Neuron-Level Interpretability in Large Language Models: A Survey

Abstract

This survey paper provides a comprehensive exploration of neuron-level interpretability in large language models (LLMs), emphasizing its critical role in enhancing transparency and trust in AI systems. The rapid advancement of LLMs has necessitated the development of Explainable AI (XAI) techniques focusing on neuron behavior and activation patterns, especially in high-stakes domains like healthcare and education. By dissecting neuron behavior, researchers aim to align AI decision-making with human cognitive frameworks, thereby fostering trust and promoting effective collaboration. The survey delves into methodologies for neuron behavior analysis, including relevance propagation and dynamic learning processes, highlighting their contributions to model transparency. Visualization and quantitative analysis techniques are discussed as pivotal tools for elucidating neuron activation patterns, enhancing model interpretability. The integration of cognitive neuroscience principles into AI methodologies is explored, underscoring their significance in aligning AI systems with human cognitive processes. The paper also addresses challenges in achieving effective interpretability, such as the generation of noisy explanations and the lack of standardized evaluation metrics. Future research directions are proposed, including the development of robust frameworks for interpretable representations and interdisciplinary collaboration to address ethical considerations. The survey concludes by reiterating the importance of neuron-level interpretability for developing AI systems that are transparent, accountable, and aligned with human values, ultimately fostering greater trust and collaboration between AI technologies and human users.

1 Introduction

1.1 Contextualizing Neuron-Level Interpretability

The rapid advancement of large language models (LLMs) underscores the critical need for Explainable Artificial Intelligence (XAI), particularly in neuron-level interpretability, which is essential for enhancing transparency and trust in AI systems [1]. In high-stakes fields such as healthcare and geosciences, understanding individual neuron activation patterns is vital for deciphering complex decision-making processes [2]. The opaque nature of deep learning models presents significant challenges, obstructing the derivation of comprehensible explanations from algorithmic outputs [3].

Neuron-level interpretability serves as a bridge between algorithmic outputs and human understanding, thereby enhancing human-AI interaction and fostering trust [4]. This approach is increasingly relevant in areas where distinguishing between human and AI-generated content is complex, impacting sectors like education and cybersecurity [5]. Instruction tuning has emerged as a pivotal mechanism for aligning LLMs with user intentions, highlighting the necessity of understanding neuron-level behavioral shifts following tuning [6].

Integrating neuron-level interpretability into AI research aims to elucidate AI decision-making processes for a diverse range of stakeholders, enhancing trust and accountability [2]. This interdis-

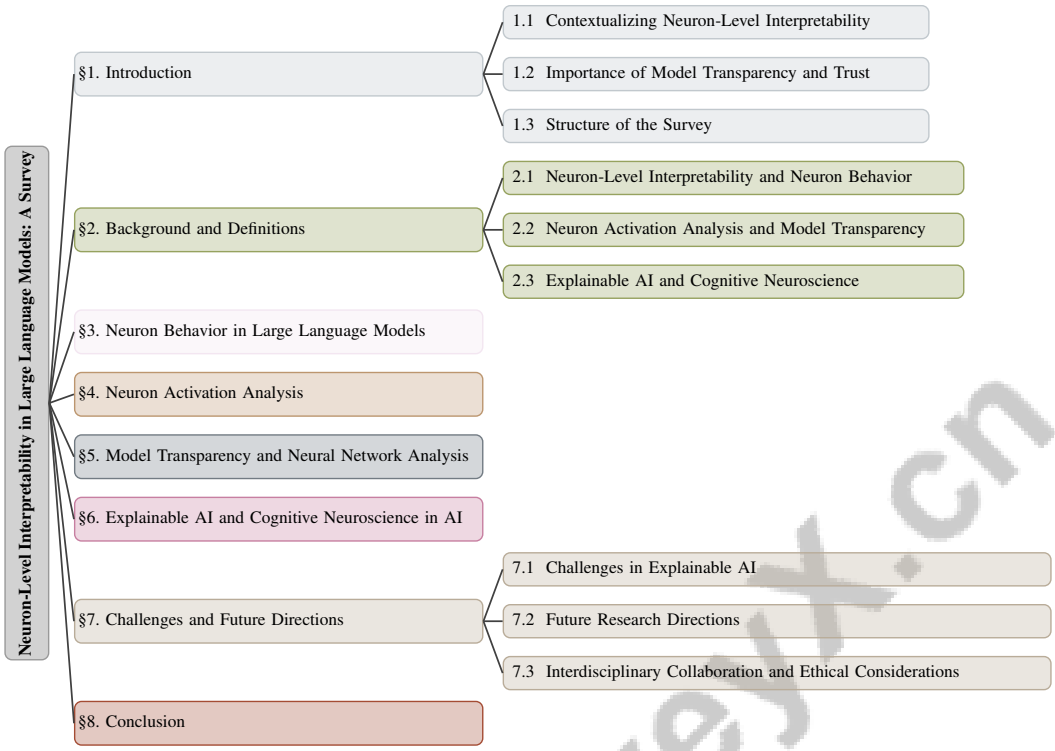

Figure 1: chapter structure

ciplinary approach is crucial for overcoming challenges in achieving explainability in AI models and advancing the field of explainable AI [5]. By clarifying the decision-making processes of AI systems, neuron-level interpretability significantly improves model transparency and accountability, addressing the limitations of existing methods that rely on closed-source models [1].

1.2 Importance of Model Transparency and Trust

The deployment of large language models (LLMs) in critical domains necessitates a steadfast focus on model transparency and trust, particularly in sectors such as healthcare, education, and finance, where erroneous outputs can have profound consequences. The inherent opacity of AI models complicates the understanding of their decision-making processes, a significant concern in multimodal explainable AI (MXAI) [7]. This opacity not only diminishes the utility of AI systems across various fields, including manufacturing and medicine, but also erodes user trust and acceptance [8].

The demand for usable explainability in LLMs, often criticized for their lack of transparency, exacerbates this issue [9]. A lack of transparency can lead to the rejection of AI systems, especially in high-stakes scenarios [10]. The necessity for explainable AI arises from the imperative for users to understand and trust AI models, ensuring fairness, accountability, and transparency [11]. Understanding the impact of explainable AI (XAI) on user compliance with AI recommendations is crucial for fostering model transparency and trust [4].

Privacy concerns also significantly influence this landscape, as the tension between explainability and privacy in machine learning models, particularly involving sensitive data, poses substantial barriers to deployment [12]. Regulatory frameworks increasingly mandate intelligible explanations from AI systems, reflecting societal demands for accountability and reliability [13]. The Contextual Importance and Utility (CIU) framework exemplifies efforts to bridge the gap between AI developers and end-users by facilitating human-like explanations of black-box outcomes [14].

Thus, the pursuit of transparency and trust in LLMs transcends technical challenges, addressing societal needs and promoting effective human-AI collaboration [15]. The ongoing challenge of ensuring user understanding and trust in AI systems through effective explanations remains central

to the field, further complicated by the absence of formal definitions and evaluation metrics for explainability [1].

1.3 Structure of the Survey

This survey is meticulously structured to explore neuron-level interpretability in large language models (LLMs), aiming to enhance transparency and trust in AI systems. It begins with an introduction to the imperative need for neuron-level interpretability, contextualizing its significance in Explainable Artificial Intelligence (XAI) across various critical domains. The subsequent analysis emphasizes the importance of model transparency and trust, particularly in sectors where AI decisions can have substantial societal consequences, underscoring the necessity for explainable AI (XAI) techniques to enhance understanding and mitigate risks associated with the black-box nature of machine learning models. This is especially pertinent in safety-critical applications, where ethical and regulatory considerations demand a balance between transparency and privacy, as well as the integration of domain knowledge to foster trust in AI systems [16, 17, 18, 12].

The second section provides background and definitions, offering an overview of key concepts related to neuron-level interpretability, including neuron behavior and activation patterns within LLMs, setting the foundation for subsequent analyses.

Section three delves into the intricate behavior of neurons within LLMs, emphasizing methodologies employed to analyze individual neuron contributions to the model's decision-making process. It discusses innovative approaches such as relevance propagation and natural language explanations to assess neuron activations, while evaluating the faithfulness and causal efficacy of these explanations. The impact of prompt tuning on the quality of neuron explanations is also highlighted, illustrating how reformulating prompts can enhance interpretability and reduce computational costs. This exploration aims to demystify the black-box nature of LLMs and improve understanding of their internal mechanisms, crucial for applications in safety-critical domains like healthcare, education, and law [19, 20, 21, 22, 23]. Additionally, it examines the dynamic properties and learning processes of neurons, providing insights into their functional roles.

In section four, neuron activation analysis is detailed, highlighting techniques for visualizing and quantitatively analyzing neuron activation patterns, which are essential for model transparency.

The fifth section investigates the role of neuron-level interpretability in enhancing model transparency, discussing mechanistic interpretability and the importance of neural network analysis in high-stakes domains, emphasizing the necessity of transparency for the acceptance and reliability of AI models.

Section six explores the intersection of explainable AI and cognitive neuroscience principles, examining how these principles inform AI methodologies and enhance interpretability, alongside theoretical frameworks that contribute to improved interpretability in AI systems.

The penultimate section identifies current challenges in achieving neuron-level interpretability in LLMs and discusses future research directions, emphasizing the need for interdisciplinary collaboration and ethical considerations in advancing explainable AI.

The survey concludes by synthesizing primary findings, emphasizing that achieving neuron-level interpretability is crucial for enhancing model transparency and fostering trust in AI systems. This interpretability allows users to gain insights into neural network decision-making processes, essential for applications in sensitive domains like healthcare and finance, where understanding model behavior can mitigate risks and biases associated with automated decisions [24, 20, 25, 21, 26]. This structured approach ensures a thorough understanding of the topic, providing a roadmap for future research and development in the field.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Neuron-Level Interpretability and Neuron Behavior

Neuron-level interpretability is pivotal in Explainable AI (XAI), focusing on demystifying large language models (LLMs) by examining individual neuron behavior [27]. This is crucial in addressing the opacity of deep learning models, often perceived as 'black boxes', particularly in sensitive areas like medical diagnostics [28]. A primary challenge is the lack of transparency in understanding the

features that neurons detect, which can lead to negative outcomes in fields requiring accountability [13].

In high-stakes domains such as healthcare and education, enhancing transparency and trust through neuron-level interpretability is essential [10]. Analyzing neuron behavior aligns AI decision-making with human cognitive frameworks, fostering trust and collaboration [4]. This interpretability provides actionable insights into AI systems' workings, addressing suboptimal decision-making in human-AI interactions [17].

Incorporating neuron-level interpretability into AI research enhances transparency and empowers stakeholders to comprehend and trust model decisions [29]. By elucidating neuron activation patterns and their contributions to outputs, this approach tackles the core challenge of explainability in black-box AI models, paving the way for a more reliable AI ecosystem [30].

2.2 Neuron Activation Analysis and Model Transparency

Neuron activation analysis is critical in elucidating the decision-making processes of LLMs, often considered opaque due to their complex architectures [31]. This analysis deciphers neuron activation patterns, providing insights into how models process inputs and generate outputs [6]. Traditional XAI methods like LIME and SHAP have limitations, highlighting the need for robust neuron activation analysis techniques to enhance model transparency [3].

The complexity of deep learning models often results in explanations misaligned with human cognition and domain knowledge, challenging interpretability [10]. Innovative methodologies, such as the NEON framework using Layer-wise Relevance Propagation (LRP), enhance transparency by linking decisions to input features [30]. Similarly, the TELL method integrates interpretability into model design, promoting transparency [29].

Despite advancements, significant barriers to transparency remain due to the variability of explanations from different XAI methods, creating uncertainty about decision-making processes [2]. The Decision Stack Framework, mimicking biological decision-making, exemplifies efforts to enhance AI explainability [13]. Neuron activation analysis is essential for enhancing model transparency through interpretable representations that align AI explanations with human cognitive processes [31].

Akula et al. highlight generating context-sensitive explanations to improve transparency by considering context [27]. However, as Dhore et al. note, current interpretability methods like GradCAM and LRP often yield noisy visualizations, obscuring insights into model predictions and affecting transparency [28]. Danilevsky et al. categorize explanations in NLP into local vs. global and self-explaining vs. post-hoc methods, enriching the understanding of neuron activation analysis in the context of transparency [1].

2.3 Explainable AI and Cognitive Neuroscience

Integrating Explainable AI (XAI) with cognitive neuroscience principles enhances LLM interpretability by leveraging insights from human cognition to inform AI design and explanation strategies [32]. Cognitive neuroscience provides a foundational understanding of human cognitive processes, enabling AI systems to produce explanations aligned with human reasoning and perception, fostering trust and collaboration [33].

Efforts to develop relatable explanations through frameworks utilizing contrastive saliency, counterfactuals, and cues address diverse stakeholder needs, making AI predictions comprehensible and actionable [14]. Applying XAI methods in cognitive neuroscience, such as functional brain imaging analysis, illustrates the potential of interdisciplinary approaches to enhance interpretability [12].

Moreover, integrating machine learning models to decode emotional states from complex data emphasizes XAI's applicability in understanding human-like cognitive processes. This approach underscores the necessity for AI systems providing transparent and comprehensible explanations while enhancing our understanding of human cognition. By incorporating user-centric design and empirical methodologies, these systems bridge the gap between AI technologies and human users, fostering trust and collaboration. Recent literature emphasizes integrating insights from psychological theories and user feedback to improve AI usability and acceptance across domains [34, 35, 36].

A multidisciplinary approach to AI explainability is underscored by frameworks categorizing XAI research, advocating for the integration of social and technical perspectives to address AI system complexities. By adopting a human-centered perspective, XAI can effectively meet users' diverse needs, ensuring explanations are meaningful, contextually relevant, and tailored to individual preferences. This approach emphasizes involving end users in the development process, facilitating non-technical, personalized explanations that clarify AI competencies and enhance trust. As XAI evolves, it holds potential to play significant social roles, such as facilitating knowledge coordination and fostering cross-disciplinary insights, enriching user experiences and understanding of AI technologies [16, 37, 34, 27, 38]. This convergence of XAI and cognitive neuroscience principles promises the development of AI systems that are transparent and aligned with human cognitive processes, enhancing the overall trustworthiness and effectiveness of AI technologies.

3 Neuron Behavior in Large Language Models

Exploring neuron behavior in large language models (LLMs) is crucial for understanding their decision-making processes. The following subsection delves into interpretability techniques that analyze individual neuron contributions, enhancing transparency and comprehension of LLMs' operational mechanisms.

3.1 Neuron Behavior in LLMs

Analyzing neuron behavior in LLMs is essential for improving interpretability and transparency, thereby fostering trust in AI systems. Hybrid interpretability techniques, such as GradCAM combined with Layer-wise Relevance Propagation (LRP), provide reliable visual explanations for convolutional neural network (CNN) predictions, clarifying individual neuron contributions [28]. Classifying Explainable AI (XAI) techniques into model-specific, model-agnostic, global, and local approaches aids in selecting methodologies for specific interpretability goals [10]. The NEON method, which transforms clustering models into neural networks, exemplifies innovative approaches to elucidate neuron contributions to cluster assignments [30]. Additionally, CNN methodologies highlight critical regions for understanding model behavior, as seen in MRI scan classification [6].

Integrating neuroscience principles into AI design leverages cognitive insights to mimic human decision-making, enhancing interpretability [13]. Saliency maps, categorized by contextual factors such as human tasks, provide nuanced insights into neuron activations [39]. Despite advancements, aligning explanations with human cognition and domain knowledge remains challenging. Developing interpretable representations is crucial for valuable insights into model behavior [31]. Balancing transparency with data privacy is a core issue in XAI [12]. These methodologies collectively enhance our understanding of LLMs' complex decision-making processes, facilitating natural language explanations and addressing challenges like hallucinated outputs and high computational demands. Ultimately, these efforts promote trust and accountability in AI systems, enabling informed decision-making across fields like healthcare and finance [40, 41].

3.2 Neuron Contribution through Relevance Propagation

Relevance propagation is a pivotal XAI technique that elucidates individual neurons' contributions to LLM predictions by attributing outputs back to input features, offering insights into decision-making processes [42]. The absLRP method exemplifies this by calculating relevance scores based on neuron activation magnitudes, improving prediction clarity and highlighting neuron significance. This technique aligns AI explanations with human cognition and domain knowledge, fostering trust and understanding, especially in high-stakes domains [34]. Involving end users in XAI system development ensures explanations are accurate, contextually relevant, and actionable.

Relevance propagation techniques, particularly LRP, enhance neural network interpretability in fields like medical diagnostics and financial modeling. They enable researchers to assess neuron contributions, validating model predictions and ensuring reliability. Recent LRP advancements address limitations by incorporating evaluation metrics for faithfulness, robustness, and localization, improving accuracy in neuron relevance assessments. Hybrid methods combining LRP with GradCAM offer clearer visual explanations, aiding neural network behavior understanding [42, 28, 19, 21]. This

method provides a granular view of neuron contributions, enabling identification of key features driving decisions, refining models for performance and interpretability.

Advancements in relevance propagation underscore the importance of robust XAI systems prioritizing user-centric explanations. Enhancing LLM transparency through relevance propagation fosters user acceptance and trust, facilitating effective human-AI collaboration and supporting decision-making processes, as highlighted in various applications, including education, healthcare, and finance [11, 43, 44, 9, 41].

3.3 Dynamic Properties and Learning Processes

Understanding the dynamic properties and learning processes of neurons in LLMs is crucial for grasping their adaptability and functional capabilities. These properties capture complex, multiscale interactions within neural networks, akin to biophysical neural networks (BNNs) [45]. Modeling these interactions provides insights into neuron learning and adaptation, contributing to LLM effectiveness and reliability.

As illustrated in Figure 2, which depicts the hierarchical structure of dynamic properties and learning processes in large language models (LLMs), key areas such as hierarchical neural interactions, neuron weight adjustments, and the alignment of AI systems with biological learning processes are emphasized. This visualization enhances our understanding of these dynamic properties, capturing hierarchical neural interactions that are essential for modeling complex data patterns and enhancing AI interpretability. This understanding aids in developing robust machine learning models and identifying features impacting learning processes, optimizing performance and transparency, crucial for adopting machine learning in critical sectors like healthcare. Visualization techniques and evaluation metrics help assess trade-offs between predictive performance and explanation quality, ensuring models are effective and interpretable [46, 47, 48, 49].

Dynamic learning processes in LLMs involve continuous adjustments of neuron weights and activation patterns as models assimilate new data, enhancing performance in tasks like text generation and language translation. This adaptability is crucial for improving interpretability and safety in high-reliability applications, allowing better understanding of decision-making and biases [40, 20]. It maintains prediction relevance and accuracy, especially in evolving data environments. Advanced representation learning techniques enhance understanding of neuron adjustments, improving generalization across contexts.

Exploring dynamic properties emphasizes developing AI systems mimicking biological neural network learning processes. Aligning AI learning mechanisms with natural ones allows creation of models that perform well and offer insights into learning and cognition principles. This alignment is crucial for advancing XAI, facilitating transparent, accountable AI systems tailored to human cognitive processes. Addressing explainability challenges in complex models promotes responsible AI practices, enhancing user trust and engagement. It emphasizes contextualizing explanations based on user needs and cognitive frameworks, fostering effective human-AI interactions across applications [50, 51, 35, 52].

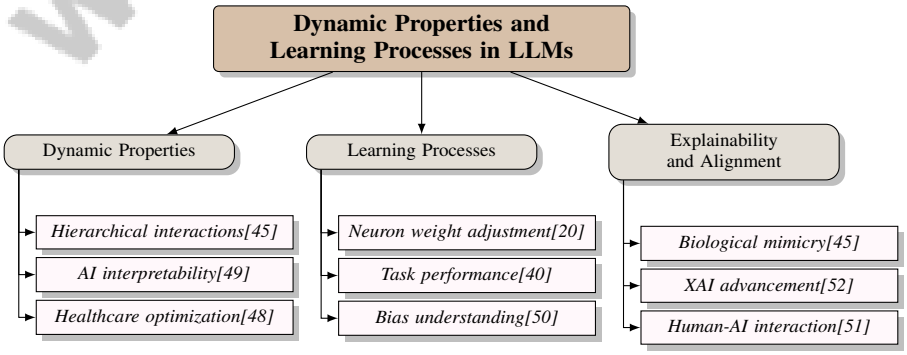

Figure 2: This figure illustrates the hierarchical structure of dynamic properties and learning processes in large language models (LLMs), emphasizing key areas such as hierarchical neural interactions, neuron weight adjustments, and the alignment of AI systems with biological learning processes.

4 Neuron Activation Analysis

4.1 Visualization Techniques for Neuron Activation

Method Name	Visualization Techniques	Interpretability Enhancement	Model Trust and Transparency
CBP[53]	Direct Visualizations	Concept Backpropagation	User Confidence Increase
SAE[54]	Semantic Attention	Semantic Attention-based	Reliable Explanations
XCNN[55]	Interpretable Heatmaps	Interpretable Heatmaps	Transparent Decision-making
G+L[28]	Gaussian Blur	Clearer Visual Explanations	Greater Trust
GXAIN[56]	-	Natural Language Narratives	Enhance User Understanding
IP[5]	Convolutional Kernels	Interpretable Pipelines	Better Understanding
MFI[3]	Mask-filling Technique	Global View	Model Interpretability

Table 1: Comparison of various visualization techniques and their impact on interpretability and trust in large language models. The table summarizes different methods, detailing their visualization techniques, how they enhance interpretability, and their contribution to model trust and transparency.

Visualization techniques are essential for interpreting neuron activation patterns in large language models (LLMs), providing insights into their decision-making processes. Table 1 provides a comprehensive overview of several visualization techniques used for interpreting neuron activation in large language models, emphasizing their role in enhancing interpretability and fostering model trust. Concept backpropagation perturbs inputs using a trained concept probe to visualize concept representation, enhancing interpretability [53]. The Semantic Attention Explanation (SAE) method uses semantically-informed perturbations to clarify how attention weights influence model outputs, effectively visualizing attention mechanisms' impact on neuron activation [54].

The XCNN method produces interpretable heatmaps during classification tasks, improving convolutional neural networks' (CNNs) explainability by offering clear visual insights into model behavior [55]. A hybrid technique combining GradCAM with Layer-wise Relevance Propagation (LRP) processes GradCAM outputs to reduce noise, multiplies them with LRP outputs, and applies Gaussian blur for clearer visual explanations, enhancing CNN interpretability [28].

The Usable XAI framework introduces ten strategies to diagnose LLMs and improve user-friendliness through generative capabilities, highlighting visualization's role in fostering AI system understanding and trust [9]. The GraphXAIN approach uses narrative techniques to explain graph neural network (GNN) predictions, improving user comprehension and trust in model outputs [56].

Custode et al. propose a framework integrating a vision module to extract high-level features from raw images with a decision module optimized through evolutionary algorithms for enhanced visualization [5]. Deelaka et al. describe a method predicting classes for masked images and visualizing updates, offering a dynamic perspective on model behavior [3].

These visualization techniques, including DeepView for dimensionality reduction and neural-specific local interpretation approaches, form a robust toolkit for analyzing neuron activation patterns in LLMs. They aid in bias detection, enhance explainability, and support model safety and reliability evaluation [24, 20, 21, 57]. Employing these methods enables deeper insights into neural network interactions, promoting transparency, trust, and accountability in AI systems.

4.2 Quantitative Analysis of Neuron Activation

Method Name	Methodological Frameworks	Evaluation Metrics	User-Centric Explanations
EO[2]	Explanation Optimizer	Faithfulness And Complexity	Comprehensible Explanations
CINA[19]	Concept Induction	Mann-Whitney U	Meaningful Explanations
DXAI[58]	Decomposition-based Explainable	Area Under Curve	Informative Visualizations
GIG[59]	Generalized Integrated Gradients	C-Deletion And C-Insertion	Shared Concept Interpretations
S-XAI[60]	Row-centered Pca	Semantic Probabilities	Semantic Spaces

Table 2: Overview of methodological frameworks, evaluation metrics, and user-centric explanations employed in quantitative analysis of neuron activation within large language models. This table summarizes various approaches and their respective evaluation criteria, highlighting the importance of user-centered explanations in enhancing the interpretability and transparency of AI systems.

Quantitative analysis of neuron activation in LLMs is crucial for elucidating the processes underlying model predictions and enhancing interpretability. Various methodologies assess individual neuron

contributions and activation patterns, bolstering AI systems' transparency and reliability. The Explanation Optimizer framework integrates outputs from multiple XAI methods to derive optimal explanations for deep classifiers, balancing faithfulness and complexity [2]. This framework provides a comprehensive quantitative understanding of neuron importance.

As illustrated in Figure 3, which depicts the hierarchical structure of quantitative analysis methodologies for neuron activation in LLMs, key methods and evaluation metrics are prominently highlighted. This figure underscores the significance of user-centric explanations, which are essential for fostering trust and understanding in AI systems. Additionally, Table 2 provides a comprehensive overview of the methodological frameworks and evaluation metrics utilized in the quantitative analysis of neuron activation in large language models, emphasizing the role of user-centric explanations in fostering trust and understanding.

Statistical analyses, such as the Mann-Whitney U test, compare activations between target and non-target images for each label, offering insights into neuron behavior [19]. Techniques like DXAI provide high-resolution, multi-channel explanations differentiating class-relevant from irrelevant features, enhancing quantitative analysis with detailed neuron activation insights [58].

Performance metrics such as C-Deletion and C-Insertion evaluate concept extraction and attribution fidelity, underscoring robust evaluation frameworks' importance in quantitative analysis [59]. The semantic interpretation of CNNs is assessed by comparing semantic probabilities and explanations generated by S-XAI against CNN predictions, focusing on trustworthiness and semantic sample searching [60].

Exploring various quantitative methods emphasizes the need for robust Explainable Artificial Intelligence (XAI) systems prioritizing user-centric explanations. User engagement is crucial for enhancing AI predictions' transparency and interpretability, empowering users to understand AI systems' competencies and limitations. Emphasizing user-centric explanations fosters trust and facilitates effective AI-user collaboration, enabling significant contributions to knowledge generation and application across diverse fields [44, 38, 16, 27]. Through detailed quantitative analyses, AI systems can achieve higher acceptance and trust, enhancing human-AI collaboration and decision-making. Collectively, these methodologies provide a comprehensive toolkit for analyzing neuron activation patterns in LLMs, facilitating improved interpretability and transparency of AI systems and fostering trustworthy human-AI interactions.

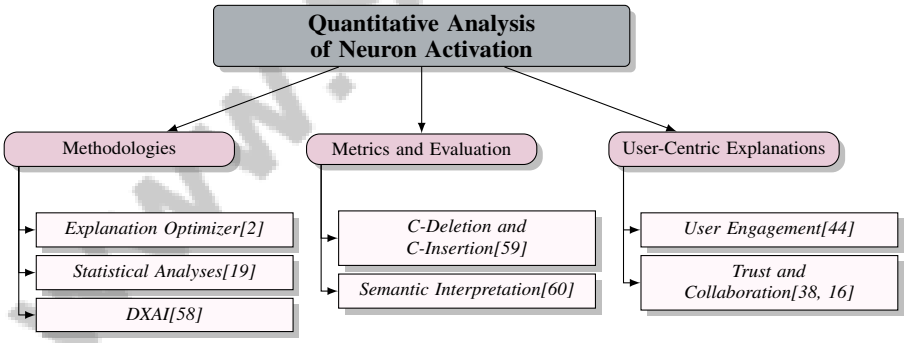

Figure 3: This figure illustrates the hierarchical structure of quantitative analysis methodologies for neuron activation in LLMs, highlighting key methods, evaluation metrics, and the importance of user-centric explanations.

5 Model Transparency and Neural Network Analysis

Understanding neural network dynamics is essential for fostering trust and accountability in AI systems. This section explores mechanistic interpretability, a key element in enhancing model transparency by elucidating internal mechanisms and illustrating how specific components influence overall behavior. This approach is vital for developing reliable and ethical AI applications, emphasizing the relationship between mechanistic interpretability and model transparency in promoting trust and reliability.

5.1 Mechanistic Interpretability and Model Transparency

Mechanistic interpretability provides crucial insights into AI systems' inner workings, essential for trust and reliability, especially in safety-critical applications. By clarifying how model components contribute to behavior, it promotes accountability [27]. Recent advancements in Explainable AI (XAI) have reinforced this field, with an emphasis on ethical considerations necessary for societal acceptance [1].

The hybrid method by Dhore et al. exemplifies the benefits of mechanistic interpretability by combining GradCAM and Layer-wise Relevance Propagation (LRP), offering enhanced interpretability compared to standalone methods and fostering greater trust in AI models [28]. This method provides clear visual explanations, crucial for understanding individual components' contributions within neural networks.

Frameworks like Akula et al.'s aim to generate meaningful explanations, aligning mechanistic interpretability with model transparency [27]. By elucidating AI decision-making processes, mechanistic interpretability fosters acceptance and trust, facilitating effective human-AI collaboration.

Despite advancements, traditional XAI methods often fall short for large language models (LLMs), highlighting the need for new strategies [1]. Ongoing research focuses on developing robust XAI systems that prioritize user-centric explanations, ensuring AI systems are interpretable, trustworthy, and aligned with stakeholder needs. These advancements provide a solid framework for enhancing model transparency.

5.2 Neural Network Analysis in High-Stakes Domains

Neural network analysis is critical in high-stakes domains where AI system reliability and accuracy have significant consequences. In healthcare, finance, and autonomous systems, understanding decision-making processes is essential for safety, fairness, and accountability [10]. The complexity of deep learning models poses challenges, as erroneous predictions can lead to harmful outcomes.

As illustrated in Figure 4, the hierarchical structure of neural network analysis encompasses key areas such as healthcare, finance, and autonomous systems, alongside relevant techniques and applications. This visual representation enhances our understanding of how these domains interconnect and the significance of employing effective analysis methods.

In healthcare, neural network analysis is vital for validating AI-driven diagnostic tools, impacting patient outcomes directly [8]. Integrating XAI techniques such as LRP and GradCAM provides insights into models' decision pathways, enabling healthcare professionals to trust AI recommendations [28]. These techniques clarify prediction rationales, ensuring alignment with clinical expertise and ethical standards.

In finance, neural network analysis aids in understanding risk factors and decision criteria in credit scoring and fraud detection [27]. Transparency through XAI methods allows stakeholders to assess fairness and bias in predictions, ensuring regulatory compliance and fostering user trust [1].

Autonomous systems, like self-driving cars, benefit from neural network analysis, where AI interpretability is crucial for safety and regulatory compliance. Understanding decision-making processes in real-time is essential for preventing accidents and ensuring ethical operations [15].

Advancements in neural network analysis underscore the necessity of robust XAI frameworks that prioritize transparency and accountability in high-stakes domains. By improving AI systems' interpretability, diverse stakeholder communities can collaboratively ensure responsible AI deployment, emphasizing safety, fairness, and ethical considerations while addressing varied concerns regarding explainability [61, 51].

6 Explainable AI and Cognitive Neuroscience in AI

6.1 Integration of Cognitive Neuroscience Principles

The incorporation of cognitive neuroscience principles into AI methodologies provides a comprehensive framework for improving the interpretability and functionality of large language models (LLMs). This interdisciplinary strategy leverages human cognition insights to guide AI design,

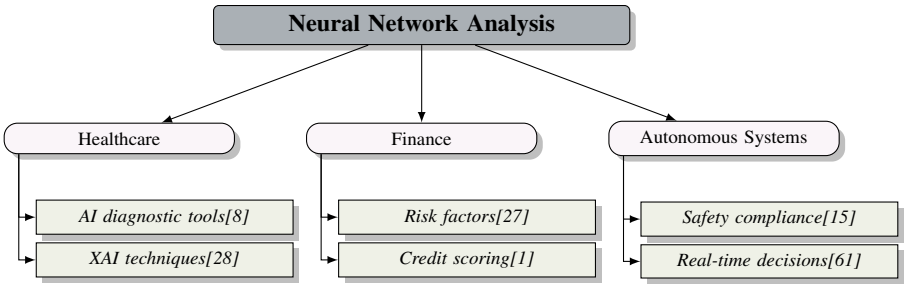

Figure 4: This figure illustrates the hierarchical structure of neural network analysis in high-stakes domains, highlighting key areas such as healthcare, finance, and autonomous systems, along with relevant techniques and applications.

ensuring alignment with human reasoning and perception, which is vital for trust and usability [62]. By focusing on conceptual understanding rather than raw data, as illustrated by the FunnyBirds benchmark, AI models can better align with human cognitive tasks, enhancing interpretability [63].

Layer-wise Relevance Propagation (LRP) exemplifies this integration by producing pixel-wise segmentation masks from image-level labeled data, clarifying AI decision-making processes [64]. This technique enhances model transparency by aligning AI explanations with human cognitive processes, promoting a deeper comprehension of AI behavior.

Frameworks categorizing Explainable AI (XAI) research based on explanation types further demonstrate how cognitive neuroscience principles can inform AI methodologies [4]. By aligning AI explanations with cognitive neuroscience insights, these frameworks ensure AI outputs are not only precise but also contextually relevant and meaningful.

The integration of cognitive neuroscience principles is also crucial for regulatory compliance, especially with frameworks like the European GDPR, which demands transparency and accountability in AI decision-making [65]. Tools such as DoX, providing an objective, model-agnostic measure of explainability, are vital for developers and regulators to ensure AI systems adhere to regulatory standards while promoting transparency and user trust [66].

6.2 Frameworks for Enhanced Interpretability

The development of theoretical frameworks for enhanced interpretability in AI is essential for advancing Explainable AI (XAI) by integrating theoretical insights with practical applications. These frameworks aim to connect complex AI models with human understanding, thereby fostering trust and usability among diverse stakeholders [67]. A robust framework that combines theoretical perspectives with practical methodologies can significantly improve AI interpretability, ensuring AI-generated explanations are accurate, meaningful, and contextually relevant.

Ethical considerations are crucial in designing and deploying interpretability frameworks. Albarracin et al. emphasize the need for ethical guidelines to ensure the responsible development and implementation of AI systems [68]. By embedding ethical principles into interpretability frameworks, developers can align AI systems with societal values and regulatory standards, enhancing acceptance and trustworthiness.

Innovative methods like LRP-Seg highlight the potential of theoretical frameworks to simplify complex AI processes, such as image segmentation, by reducing dependence on extensive pixel-wise labeled data [64]. This approach not only improves AI model interpretability but also enhances their accessibility and applicability across various domains, particularly in scenarios where labeled data is scarce or costly to obtain.

7 Challenges and Future Directions

Exploring challenges and future directions in Explainable AI (XAI) involves addressing critical obstacles to effective interpretability in AI models. Understanding these challenges reveals the complexities of achieving explainability and identifies viable research pathways. A critical examination

of the multifaceted challenges in generating clear and meaningful explanations for users highlights the need for standardized methods of explainability across opaque, interpretable, comprehensible, and truly explainable systems. As illustrated in Figure 5, the hierarchical structure of challenges and future directions in XAI categorizes key issues into interpretability, user-centered, and trust-related challenges, while outlining future research directions focused on advancements in interpretability and standardization. This figure also emphasizes the importance of interdisciplinary collaboration and ethical considerations in developing socially responsible AI systems. High-performing machine learning models often lack interpretability, complicating users' understanding of AI-driven decisions. Aligning AI explanations with users' cognitive needs, particularly in high-stakes environments, is essential. A hybrid approach combining knowledge-based frameworks with machine learning can enhance AI systems' explanatory capabilities [69, 70, 1, 37, 71].

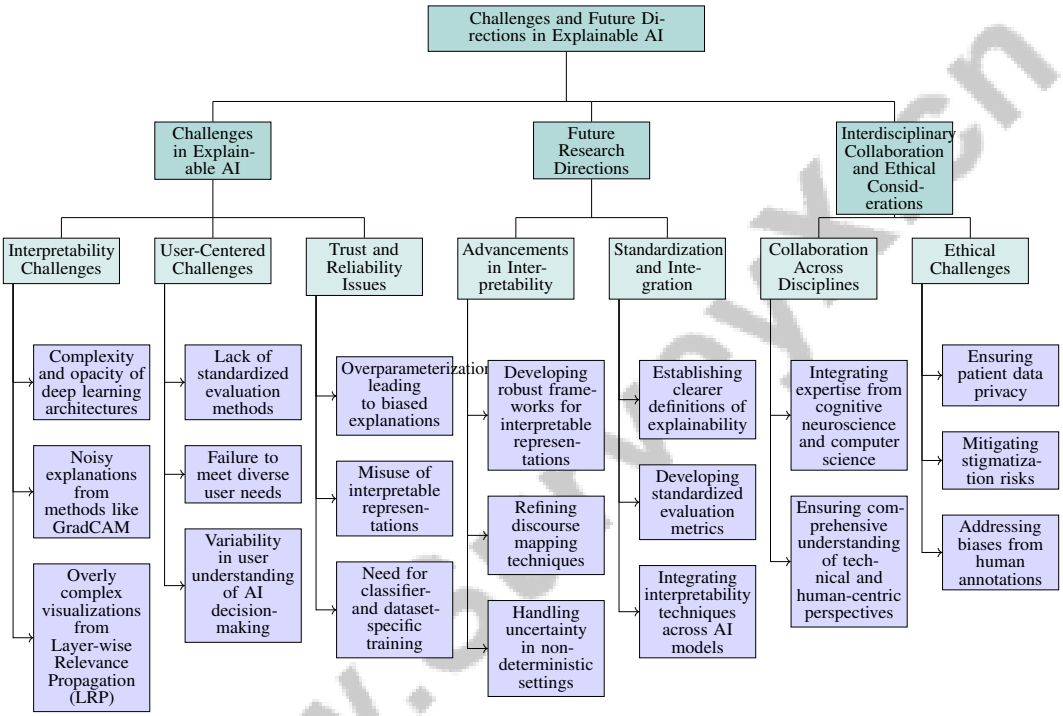

Figure 5: This figure illustrates the hierarchical structure of challenges and future directions in Explainable AI (XAI), categorizing key challenges into interpretability, user-centered, and trust-related issues, while outlining future research directions focused on advancements in interpretability and standardization. It also highlights the importance of interdisciplinary collaboration and ethical considerations in developing socially responsible AI systems.

7.1 Challenges in Explainable AI

Achieving explainability in AI models faces several challenges due to the inherent complexity and opacity of deep learning architectures. Methods like GradCAM can produce noisy explanations that obscure interpretability, while overly complex visualizations from Layer-wise Relevance Propagation (LRP) hinder clear insights from model predictions [28]. The lack of standardized evaluation methods impedes the assessment and comparison of different XAI techniques [1]. This issue is compounded by the failure to meet diverse user needs, resulting in explanations that may not be meaningful to all stakeholders [1]. The variability in user understanding of AI decision-making processes complicates explainability pursuits, as current approaches often overlook user needs and contextual factors [4]. This gap underscores the necessity for more user-centered XAI systems.

Moreover, overparameterization and misuse of interpretable representations can lead to biased and unreliable explanations, undermining trust in AI systems [31]. The requirement for classifier- and dataset-specific training adds complexity, making the process resource-intensive [58]. Addressing these challenges necessitates developing robust, interpretable, and user-centered AI systems that

facilitate comprehensible outputs for user-driven explanations. Standardization in explainability methods is crucial, as varying interpretations of 'explainable' and 'interpretable' AI exist among different stakeholder communities. Clarifying these concepts and their implications can align with diverse user needs, enhancing trust and usability in AI applications [37, 61, 70]. Such efforts can advance XAI towards greater transparency, fairness, and accountability, fostering trust and acceptance among users.

7.2 Future Research Directions

Advancing neuron-level interpretability in large language models (LLMs) requires a comprehensive research agenda focused on transparency, usability, and accountability. Developing robust frameworks for interpretable representations and clear evaluation metrics for their effectiveness is essential for progressing in XAI [31]. This includes refining discourse mapping techniques to capture neuron interaction intricacies, enhancing neural network interpretability [27]. Research must incorporate mechanisms to handle uncertainty in non-deterministic settings, utilizing fuzzy or probabilistic decision trees to provide nuanced insights into model behavior and improve decision-making processes [5]. Exploring the performance of hybrid interpretability methods, such as those combining GradCAM and LRP across various convolutional neural network (CNN) architectures, can yield versatile and effective interpretability solutions [28].

Establishing clearer definitions of explainability and developing standardized evaluation metrics are critical for advancing neuron-level interpretability, facilitating consistent assessments of XAI methods and contributing to the development of transparent and accountable AI systems [1]. Integrating these methods with other interpretability techniques can enhance their effectiveness across various AI models and domains. These research avenues can significantly advance neuron-level interpretability, leading to AI systems that are transparent, accountable, and closely aligned with human cognitive processes. This evolution is vital for enhancing trust in AI applications across critical domains such as healthcare and finance, where understanding the rationale behind machine decisions is crucial. Integrating interpretable deep learning models providing human-understandable justifications for their outputs will foster greater acceptance and reliability in human-machine interactions [70, 24, 25, 35, 26].

7.3 Interdisciplinary Collaboration and Ethical Considerations

Interdisciplinary collaboration is vital for advancing neuron-level interpretability in LLMs, integrating diverse expertise from cognitive neuroscience, computer science, and ethics. This collaborative approach addresses XAI's complex challenges, ensuring AI systems are developed with a comprehensive understanding of technical and human-centric perspectives. Insights from cognitive neuroscience can guide the design of AI systems that align more closely with human cognitive processes, enhancing interpretability and usability [72]. Ethical considerations are paramount in AI system development, particularly in sensitive domains like healthcare and mental health. Ensuring patient data privacy and mitigating stigmatization risks associated with AI applications are critical imperatives [72]. Additionally, reliance on human annotations in AI systems can introduce biases and inconsistencies in evaluation criteria across diverse tasks [73]. Addressing these ethical challenges requires developing standardized guidelines and frameworks prioritizing fairness, accountability, and transparency in AI systems.

Integrating ethical principles into AI research fosters trust and acceptance among users and ensures compliance with regulatory standards and societal values. By embracing interdisciplinary collaboration and ethical considerations, neuron-level interpretability can advance towards developing AI systems that are technically robust, socially responsible, and aligned with human values. This comprehensive approach is crucial for enhancing understanding and collaboration between AI technologies and human users, addressing diverse stakeholder needs and expectations regarding explainability and interpretability. By fostering greater transparency, this strategy contributes to developing trustworthy and accountable AI systems aligned with users' specific intents and requirements [61, 51, 36].

8 Conclusion

The exploration of neuron-level interpretability is pivotal for advancing the transparency and trustworthiness of large language models (LLMs). By delving into neuron behavior and activation patterns, researchers can develop AI systems that better align with human cognitive processes, thereby enhancing user trust. The integration of neural learning with symbolic reasoning emerges as a robust approach for creating explainable AI systems that are both reliable and efficient. However, the current explanation methods predominantly serve expert users, indicating a pressing need for more intuitive interfaces that cater to non-experts, thereby broadening the accessibility of AI technologies.

Ensuring reproducibility in deep learning models remains essential for building trust in AI applications. The Concept Induction method has demonstrated potential in providing meaningful insights into hidden neuron activations, outperforming existing methods in both quantitative and qualitative evaluations. While LLMs offer promising avenues for improving interpretability through more accessible and detailed explanations, challenges persist in guaranteeing the reliability of these explanations. This highlights the ongoing need for research focused on developing robust, user-friendly interpretability tools that can cater to a diverse range of users.

References

[1] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[2] Michail Mamalakis, Antonios Mamalakis, Ingrid Agartz, Lynn Egeland Mørch-Johnsen, Graham Murray, John Suckling, and Pietro Lio. Solving the enigma: Enhancing faithfulness and comprehensibility in explanations of deep networks, 2025.

[3] Pathirage N. Deelaka, Tharindu Wickremasinghe, Devin Y. De Silva, and Lisara N. Gajaweera. Fill in the blanks: Rethinking interpretability in vision, 2024.

[4] Niklas Kühl, Christian Meske, Maximilian Nitsche, and Jodie Lobana. Investigating the role of explainability and ai literacy in user compliance, 2024.

[5] Leonardo Lucio Custode and Giovanni Iacca. Interpretable pipelines with evolutionarily optimized modules for rl tasks with visual inputs, 2022.

[6] Tyler Morris, Ziming Liu, Longjian Liu, and Xiaopeng Zhao. Using a convolutional neural network and explainable ai to diagnose dementia based on mri scans, 2024.

[7] Nikolaos Rodis, Christos Sardianos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Iraklis Varlamis, and Georgios Th. Papadopoulos. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions, 2024.

[8] Julian Senoner, Simon Schallmoser, Bernhard Kratzwald, Stefan Feuerriegel, and Torbjørn Netland. Explainable ai improves task performance in human-ai collaboration, 2024.

[9] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. Usable xai: 10 strategies towards exploiting explainability in the llm era, 2024.

[10] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020.

[11] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. Does explainable artificial intelligence improve human decision-making?, 2020.

[12] Fatima Ezzeddine. Privacy implications of explainable ai in data-driven systems, 2024.

[13] J. L. Olds, M. S. Khan, M. Nayebpour, and N. Koizumi. Explainable ai: A neurally-inspired decision stack framework, 2019.

[14] Kary Främling. Explainable ai without interpretable model, 2020.

[15] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, 2023.

[16] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. Reviewing the need for explainable artificial intelligence (xai), 2021.

[17] Saifullah Saifullah, Dominique Mercier, Adriano Lucieri, Andreas Dengel, and Sheraz Ahmed. Privacy meets explainability: A comprehensive impact benchmark, 2022.

[18] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. Knowledge-intensive language understanding for explainable ai, 2021.

[19] Abhilekha Dalal, Md Kamruzzaman Sarker, Adrita Barua, Eugene Vasserman, and Pascal Hitzler. Understanding cnn hidden neuron activations using structured background knowledge and deductive reasoning, 2023.

[20] Justin Lee, Tuomas Oikarinen, Arjun Chatha, Keng-Chi Chang, Yilan Chen, and Tsui-Wei Weng. The importance of prompt tuning for automated neuron explanations. *arXiv preprint arXiv:2310.06200*, 2023.

[21] Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. *arXiv preprint arXiv:2309.10312*, 2023.

[22] Abhilekha Dalal, Rushrukh Rayan, and Pascal Hitzler. Error-margin analysis for hidden neuron activation labels, 2024.

[23] Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Eugene Y. Vasserman, Md Kamruzzaman Sarker, and Pascal Hitzler. On the value of labeled data and symbolic methods for hidden neuron activation analysis, 2024.

[24] Ioannis Mollas, Nick Bassiliades, and Grigorios Tsoumakas. Lionets: A neural-specific local interpretation technique exploiting penultimate layer information, 2021.

[25] Thanos Tagaris and Andreas Stafylopatis. Hide-and-seek: A template for explainable ai, 2020.

[26] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: A survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, pages 1–6. IEEE, 2017.

[27] Arjun Akula and Song-Chun Zhu. Effective representation to capture collaboration behaviors between explainer and user, 2022.

[28] Vaibhav Dhore, Achintya Bhat, Viraj Nerlekar, Kashyap Chavhan, and Aniket Umare. Enhancing explainable ai: A hybrid approach combining gradcam and lrp for cnn interpretability, 2024.

[29] Xi Peng, Yunnan Li, Ivor W. Tsang, Hongyuan Zhu, Jiancheng Lv, and Joey Tianyi Zhou. Xai beyond classification: Interpretable neural clustering, 2022.

[30] Jacob Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. From clustering to cluster explanations via neural networks, 2021.

[31] Kacper Sokol and Peter Flach. Interpretable representations in explainable ai: From theory to practice, 2024.

[32] Stefano Teso, Öznur Alkan, Wolfang Stammer, and Elizabeth Daly. Leveraging explanations in interactive machine learning: An overview, 2022.

[33] Pradyumna Tambwekar and Matthew Gombolay. Towards reconciling usability and usefulness of explainable ai methodologies, 2023.

[34] AKM Bahalul Haque, A. K. M. Najmul Islam, and Patrick Mikalef. Notion of explainable artificial intelligence – an empirical investigation from a users perspective, 2023.

[35] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22, 2019.

[36] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai, 2019.

[37] Othman Benchekroun, Adel Rahimi, Qini Zhang, and Tetiana Kodliuk. The need for standardized explainability, 2020.

[38] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019.

[39] Romy Müller. How explainable ai affects human performance: A systematic review of the behavioural consequences of saliency maps, 2024.

[40] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

[41] Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Pohsun Feng, Yizhu Wen, Xinyuan Song, Tianyang Wang, Ming Liu, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, Junhao Song, Hong-Ming Tseng, Yichao Zhang, Lawrence K. Q. Yan, Qian Niu, Silin Chen, Yunze Wang, and Chia Xin Liang. A comprehensive guide to explainable ai: From classical models to llms, 2024.

[42] Davor Vukadin, Petar Afrić, Marin Šilić, and Goran Delač. Advancing attribution-based neural network explainability through relative absolute magnitude layer-wise relevance propagation and multi-component evaluation, 2024.

[43] Prathamesh Dinesh Joshi, Sahil Pocker, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. Hullmi: Human vs llm identification with explainability, 2024.

[44] Vinitra Swamy, Davide Romano, Bhargav Srinivasa Desikan, Oana-Maria Camburu, and Tanja Käser. illuminate: An llm-xai framework leveraging social science explanation theories towards actionable student performance feedback, 2025.

[45] Youngmok Ha, Yongjoo Kim, Hyun Jae Jang, Seungyeon Lee, and Eunji Pak. Network representation learning for biophysical neural network analysis, 2024.

[46] Tim Räz. Ml interpretability: Simple isn't easy, 2022.

[47] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning, 2018.

[48] Sreejita Ghosh, Peter Tino, and Kerstin Bunte. Visualisation and knowledge discovery from interpretable models, 2020.

[49] Mitchell Naylor, Christi French, Samantha Terker, and Uday Kamath. Quantifying explainability in nlp and analyzing algorithms for performance-explainability tradeoff, 2021.

[50] Garrick Cabour, Andrés Morales, Élise Ledoux, and Samuel Bassetto. Towards an explanation space to align humans and explainable-ai teamwork, 2021.

[51] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*, 2018.

[52] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, 2019.

[53] Patrik Hammersborg and Inga Strümke. Concept backpropagation: An explainable ai approach for visualising learned concepts in neural network models, 2023.

[54] Efimia Panagiotaki, Daniele De Martini, and Lars Kunze. Semantic interpretation and validation of graph attention-based explanations for gnn models, 2023.

[55] Amirhossein Tavanaei. Embedded encoder-decoder in convolutional networks towards explainable ai, 2020.

[56] Mateusz Cedro and David Martens. Graphxain: Narratives to explain graph neural networks, 2025.

[57] Isaac Roberts, Alexander Schulz, Luca Hermes, and Barbara Hammer. Targeted visualization of the backbone of encoder llms, 2024.

[58] Elnatan Kadar and Guy Gilboa. Dxai: Explaining classification by image decomposition, 2024.

[59] Yearim Kim, Sangyu Han, Sangbum Han, and Nojun Kwak. Decompose the model: Mechanistic interpretability in image models with generalized integrated gradients (gig), 2024.

[60] Hao Xu, Yuntian Chen, and Dongxiao Zhang. Semantic interpretation for convolutional neural networks: What makes a cat a cat?, 2022.

[61] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable ai, 2018.

[62] Ouail Kitouni, Niklas Nolte, Víctor Samuel Pérez-Díaz, Sokratis Trifinopoulos, and Mike Williams. From neurons to neutrons: A case study in interpretability, 2024.

[63] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods, 2023.

[64] Clemens Seibold, Johannes Künzel, Anna Hilsmann, and Peter Eisert. From explanations to segmentation: Using explainable ai for image segmentation, 2022.

[65] Sebastian Palacio, Adriano Lucieri, Mohsin Munir, Jörn Hees, Sheraz Ahmed, and Andreas Dengel. Xai handbook: Towards a unified framework for explainable ai, 2021.

[66] Francesco Sovrano and Fabio Vitali. An objective metric for explainable ai: How and why to estimate the degree of explainability, 2023.

[67] Marco Valentino and André Freitas. Scientific explanation and natural language: A unified epistemological-linguistic perspective for explainable ai, 2022.

[68] Mahault Albarracin, Inês Hipólito, Safae Essafi Tremblay, Jason G. Fox, Gabriel René, Karl Friston, and Maxwell J. D. Ramstead. Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making, 2023.

[69] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing, 2020.

[70] Derek Doran, Sarah Schulz, and Tarek R. Besold. What does explainable ai really mean? a new conceptualization of perspectives, 2017.

[71] Sergei Nirenburg, Marjorie McShane, Kenneth W. Goodman, and Sanjay Oruganti. Explaining explaining, 2024.

[72] Yeldar Toleubay, Don Joven Agravante, Daiki Kimura, Baihan Lin, Djallel Bouneffouf, and Michiaki Tatsubori. Utterance classification with logical neural network: Explainable ai for mental disorder diagnosis, 2023.

[73] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.