Understanding Large Language Models: A Survey on Neuron-Level Interpretability, Model Transparency, and Explainable AI

Abstract

This survey explores the transformative potential and challenges of Large Language Models (LLMs) in the context of neuron-level interpretability and model transparency. LLMs, with their advanced capabilities, have the potential to revolutionize various domains, including systematic reviews, by reducing completion times and enhancing quality. However, their deployment presents unique challenges, necessitating robust evaluation metrics and mitigation strategies to ensure ethical and reliable applications. The integration of diverse explanatory techniques is crucial for gaining wider acceptance of AI systems by providing satisfactory explanations. The survey highlights the emergence of self-learning capabilities within LLMs, exemplified by models trained with frameworks like SECToR, achieving high accuracy in complex tasks. Additionally, advancements in training methodologies, such as the S2D method, contribute to the efficiency of LLMs. In neuroscience, novel frameworks for brain decoding illustrate the potential applications of LLMs in brain-computer interfaces, bridging AI and cognitive sciences. Despite these advancements, ethical considerations remain crucial, particularly in applications like ChatGPT, where current limitations must be addressed. The integration of Knowledge Graphs (KGs) offers a promising avenue for improving LLM outputs' reliability and addressing research gaps. Continued research in neuron-level interpretability and model transparency is essential for advancing AI systems' trustworthiness, ensuring alignment with ethical standards and human values. As AI technologies evolve, these efforts will be vital in developing AI systems that are powerful, efficient, transparent, and reliable.

1 Introduction

1.1 Overview of Large Language Models (LLMs)

Large Language Models (LLMs) signify a major leap in artificial intelligence, especially in natural language processing (NLP). With extensive parameterization and complex architectures, LLMs excel in various data processing tasks, showcasing transformative potential across multiple domains [1]. Their adaptability to diverse downstream tasks emphasizes their importance in the AI landscape.

In creative applications, LLMs address challenges in conceptual blending within pop culture, outperforming traditional methods [2]. The debate regarding whether models like Chat-GPT or Claude genuinely understand language or merely produce statistically plausible text is a significant topic in linguistic studies [3]. Moreover, LLMs extend their utility into embodied settings, such as robotics, where they facilitate text generation processes [4].

In education, LLMs are transforming computing education by navigating the challenges and opportunities posed by generative AI technologies [5]. They enhance decision-making in complex

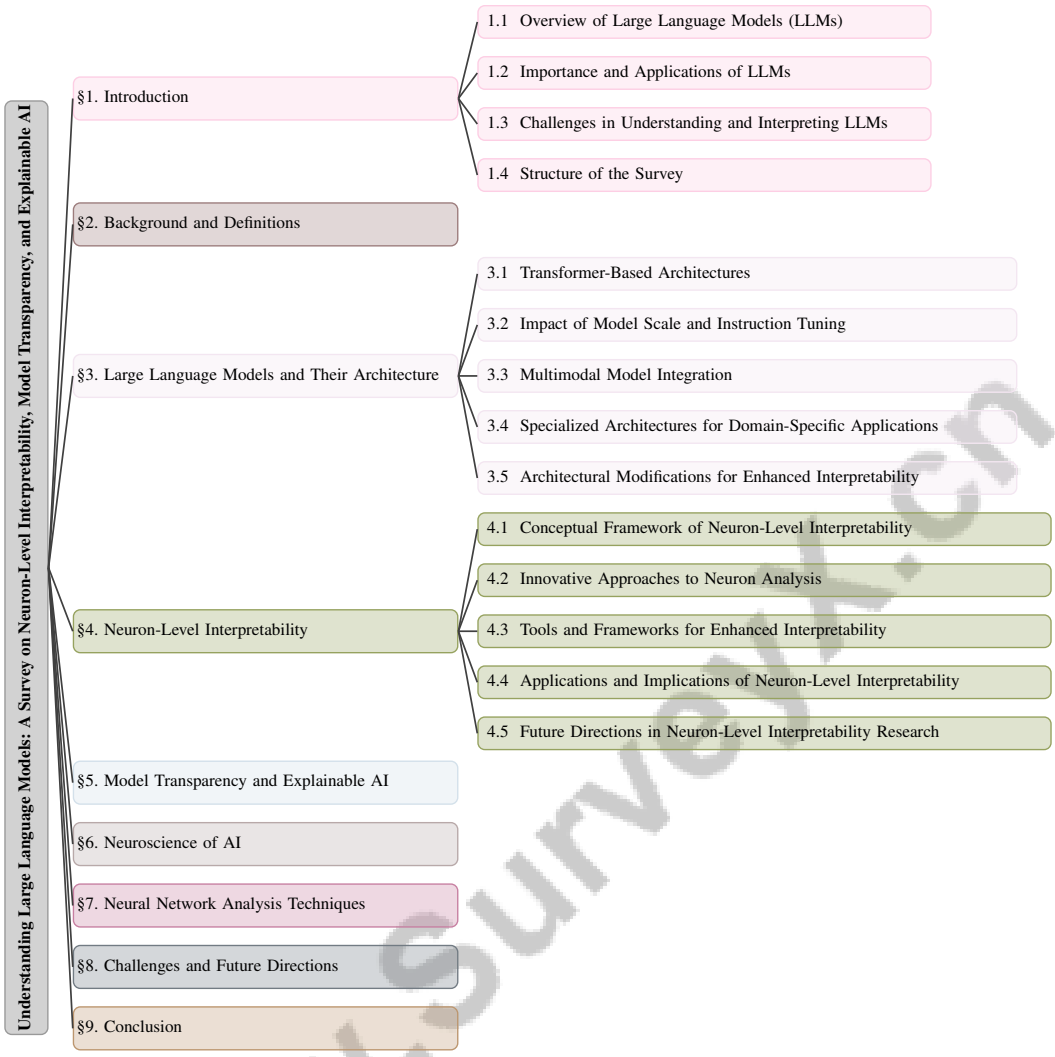

Figure 1: chapter structure

environments through integration with optimization algorithms [6], and in dentistry, models like ChatGPT demonstrate potential in automated and cross-modal dental diagnosis [7].

Despite their vast capabilities, LLM performance is impacted by factors such as pre-training corpora, affecting their efficacy in tasks like causal discovery [8]. The rapid advancements in neural networks necessitate scaling model size, training data, and computational resources to sustain effectiveness [9]. Additionally, optimizing graph flattening methods for LLMs poses challenges, particularly in long-distance scenarios due to the poor organization of textual formats [10].

Ethical considerations surrounding LLMs, particularly regarding models like ChatGPT, highlight the necessity for comprehensive research into their applications and societal impacts [11]. As LLMs evolve, their integration into various fields underscores their significance, while ongoing research addresses challenges related to interpretability and transparency.

1.2 Importance and Applications of LLMs

LLMs have become essential in artificial intelligence, showcasing remarkable versatility across numerous applications. Their ability to facilitate collaborative training while ensuring data privacy is particularly pertinent in federated learning environments, crucial for sectors like healthcare and finance that prioritize secure data handling [12]. This privacy-preserving feature is vital for maintaining ethical standards in AI deployment.

Ethically, LLMs confront challenges related to privacy, fairness, hallucination, accountability, and censorship, which are central to discussions on responsible AI integration [13]. Aligning LLMs with human values is critical, necessitating adherence to ethical guidelines to ensure socially acceptable operations [14].

The applications of LLMs extend to creative fields such as music, aiding in tasks like melody harmonization and chord-conditioned compositions [15]. Their role in prototyping machine learning features is significant, assisting professionals in technology development [16]. Furthermore, LLMs enhance programming by generating code explanations, with established benchmarks for systematic performance evaluation [17].

In educational contexts, LLMs reshape computing education by addressing challenges and opportunities related to generative AI technologies, encompassing literature reviews, surveys, pedagogical adaptations, ethical considerations, and performance benchmarks [5]. The implications of Human-Centric eXplainable AI (HCXAI) further highlight LLMs' transformative potential in enhancing educational experiences [18].

Moreover, LLMs are integrated into applications aimed at fostering empathy, such as psychotherapy, where emotional understanding and response are critical [19]. In multilingual settings, benchmarks have been developed to assess LLMs' moral alignment with human preferences, enabling cross-linguistic comparisons and ensuring ethical consistency [20].

The diverse applications of LLMs across crucial sectors such as healthcare, education, and creative industries underscore their transformative potential, enhancing processes like systematic literature reviews and predictive analytics while raising important ethical considerations regarding transparency and accountability [21, 22, 23, 13]. As these models evolve, they promise to drive innovation and address complex problems, reinforcing their pivotal role in shaping the future of artificial intelligence.

1.3 Challenges in Understanding and Interpreting LLMs

LLMs present substantial challenges in understanding and interpretability due to their complex architectures and extensive datasets. A primary concern is the difficulty in reliably assessing outputs, influenced by anti-causal statements in training data that can mislead models and reduce predictive accuracy [8]. Additionally, high computational resource demands and the complexity of ensuring interpretability in optimization further complicate effective LLM utilization [6].

The diminishing returns in performance when scaling the number of accelerators in distributed training, exacerbated by increased communication overhead, add complexity [9]. Furthermore, existing graph flattening methods often perform inadequately in long-distance reasoning tasks, complicating the organization of textual formats and hindering understanding of long-range dependencies [10].

Ensuring LLMs do not produce harmful or biased content remains a critical issue, as existing safety alignment efforts have yet to fully address underlying vulnerabilities [24]. The absence of systematic benchmarks to evaluate reasoning biases and the relationship between premises and conclusions complicates the assessment of LLMs' ethical implications [25]. In domains like dentistry, inefficiencies in processing vast amounts of structured and unstructured data hinder effective diagnosis and treatment planning, illustrating broader data management challenges in LLM applications [7].

The inherent subjectivity in evaluating creativity, which varies based on individual perceptions, presents another challenge, as existing methods struggle to produce outputs that consistently meet human standards of creativity [26]. These multifaceted challenges highlight the need for ongoing research and innovation to enhance the reliability, interpretability, and ethical deployment of LLMs across diverse domains.

1.4 Structure of the Survey

This survey is meticulously structured to provide a comprehensive exploration of Large Language Models (LLMs) and their interpretability. It commences with an **Introduction**, which lays the groundwork by presenting the significance and challenges associated with LLMs. The introductory section also includes an **Overview of Large Language Models**, detailing their expansive capabilities and applications. The **Importance and Applications of LLMs** subsection elaborates on their role across various fields, including healthcare, education, and creativity. The **Challenges in Understanding**

and Interpreting LLMs are then highlighted, focusing on issues such as computational resource demands and ethical considerations.

Following this, the survey delves into the **Background and Definitions**, offering a foundational understanding of key concepts such as neuron-level interpretability and explainable AI. The third section, **Large Language Models and Their Architecture**, provides an in-depth analysis of LLM architectures, including Transformer-based models and the impact of model scale. This section also examines the integration of multimodal data and domain-specific architectures.

The survey subsequently shifts focus to **Neuron-Level Interpretability**, exploring methodologies designed to analyze the roles of individual neurons within neural networks and their impact on model decision-making. This exploration includes examining how neurons can be understood as linear combinations of concepts, revealing insights into their causal effects beyond mere high activation levels. Additionally, it addresses the significance of mechanistic interpretability in uncovering the algorithms employed by neural networks, thereby enhancing our understanding of their predictive capabilities and aligning them with human-derived domain knowledge [27, 28, 29, 30, 31]. This is followed by a discussion on **Model Transparency and Explainable AI**, emphasizing the principles and techniques that enhance the comprehensibility of AI systems.

In **Neuroscience of AI**, the parallels between AI and neuroscience are examined, providing insights into how neuroscientific principles can inform AI development. The subsequent section, **Neural Network Analysis Techniques**, reviews various methodologies for understanding neural network behavior, including visualization and activation analysis.

The penultimate section, **Challenges and Future Directions**, identifies emerging challenges and proposes future research opportunities to advance interpretability and transparency in AI. The survey concludes with a **Conclusion** that synthesizes the key findings and underscores the importance of continued research in this dynamic field.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Background and Definitions

Large Language Models (LLMs) represent a notable advancement in artificial intelligence, enabling complex language processing and generation via transformer architectures and extensive datasets [8]. A significant challenge for LLMs is catastrophic forgetting, where fine-tuning for new tasks can result in the loss of previously acquired knowledge, necessitating strategies that balance knowledge retention with adaptability [18].

Neuron-level interpretability is crucial for understanding neural networks by examining individual neurons' functions and behaviors, which aids in deciphering decision-making processes and identifying biases or errors in outputs [25]. Model transparency, a concept closely tied to interpretability, involves making a model's operations and decisions comprehensible and trustworthy to humans. This involves methodologies that enhance understandability, fostering trust and facilitating deployment in critical fields like healthcare and finance [7].

The neuroscience of AI explores parallels between artificial neural networks and the human brain, utilizing insights from cognitive science and neurobiology to inform AI model design and interpretation. This interdisciplinary approach aims to enhance AI systems' robustness and efficiency by emulating biological systems' adaptive learning capabilities [32]. Techniques such as visualizing activation patterns, analyzing attention mechanisms, and assessing model performance across tasks are vital for optimizing model efficacy and understanding the strengths and limitations of various architectures and training paradigms [33, 34].

Explainable AI (XAI) seeks to make AI operations transparent and comprehensible to humans by developing tools and frameworks that elucidate how AI models process information and make decisions, bridging the gap between complex outputs and human interpretability, particularly in sensitive applications like healthcare and finance [18, 3].

Integrating Knowledge Graphs (KGs) with LLMs is an emerging research area aimed at reducing hallucinations and improving AI output accuracy. This involves employing relevant datasets, benchmarks, and methodologies to enhance AI reliability [35]. However, evaluating LLMs' reason-

ing capabilities and systematically constructing datasets for logical reasoning assessment remain underexplored, underscoring the need for comprehensive benchmarks in this domain [25].

Collectively, these concepts underpin ongoing efforts to enhance AI systems' interpretability, transparency, and reliability. By addressing these critical aspects, researchers aim to develop AI models that are not only powerful and efficient but also aligned with ethical standards and human values [34].

3 Large Language Models and Their Architecture

The architecture of Large Language Models (LLMs), particularly those based on transformers, has revolutionized natural language processing (NLP). This section explores these architectures, highlighting the mechanisms that enable LLMs to perform complex tasks and adapt to diverse applications. As illustrated in Figure 2, the figure depicts the hierarchical structure of large language models, emphasizing the various architectural innovations and their applications. It categorizes the core mechanisms of transformer-based architectures, the impact of model scale and instruction tuning, the integration of multimodal models, specialized architectures for domain-specific applications, and architectural modifications for enhanced interpretability. The subsequent subsection delves into transformer architectures, emphasizing their components and advantages in LLM contexts.

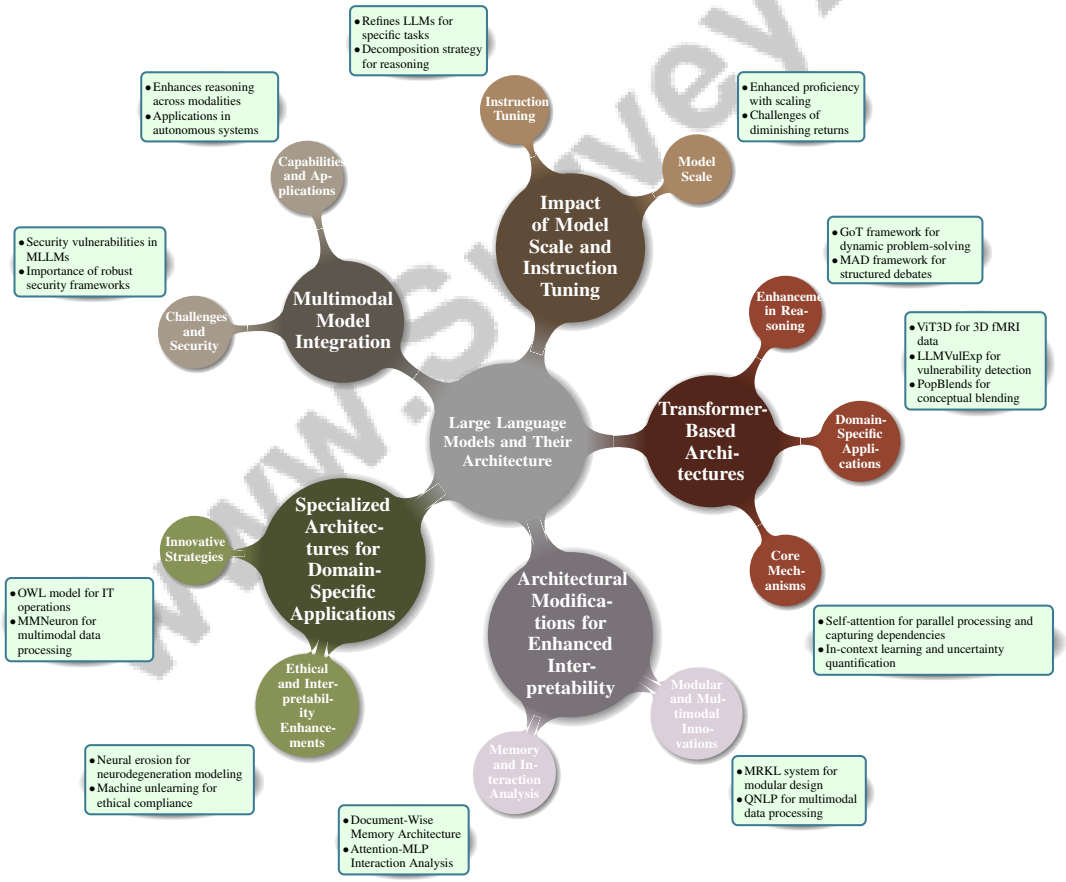

Figure 2: This figure illustrates the hierarchical structure of large language models, emphasizing the various architectural innovations and their applications. It categorizes the core mechanisms of transformer-based architectures, the impact of model scale and instruction tuning, integration of multimodal models, specialized architectures for domain-specific applications, and architectural modifications for enhanced interpretability.

3.1 Transformer-Based Architectures

Transformers are integral to LLMs, underpinning their remarkable NLP capabilities. The self-attention mechanism central to transformers facilitates parallel input sequence processing, capturing long-range dependencies and contextual relationships vital for tasks like numeral and unit conversions [36]. Their versatility is evident in domain-specific applications, such as the ViT3D architecture for 3D fMRI data, which maintains spatial structures for effective visual reconstruction [37].

Recent innovations in transformers address domain-specific challenges through novel strategies. In-context learning and uncertainty quantification modules enhance instruction-following and confidence in generated relations [36]. Frameworks like LLMVulExp leverage transformers for specialized tasks, such as vulnerability detection, providing detailed explanations within a cohesive framework [38].

Transformers also advance creative domains, as demonstrated by PopBlends, which automates conceptual blend suggestions by combining traditional knowledge extraction with LLM capabilities [2]. The Grounded Decoding (GD) method ensures semantic accuracy and physical realizability by decoding sequences under both LLM and grounded models [4].

In reasoning and problem-solving, frameworks like GoT enhance LLMs' capabilities by forming arbitrary graph structures for dynamic problem-solving [39]. The MAD framework fosters structured debates among multiple agents to achieve consensus and encourage divergent thinking [40].

Lifelong learning methodologies in transformers organize models into Internal and External Knowledge frameworks, incorporating continual pretraining, fine-tuning, retrieval-based, and tool-based learning strategies [41]. This ensures LLMs remain adaptable and relevant, supporting continuous improvement across fields.

As illustrated in Figure 3, the hierarchical structure of transformer-based architectures categorizes them into domain-specific applications, innovative strategies, and lifelong learning approaches. Each category highlights key methods and frameworks that demonstrate the versatility and adaptability of transformers in various contexts. Transformer-based architectures provide a robust framework for processing and generating complex linguistic data. Their integration with additional modules and domain-specific datasets enhances their capabilities, establishing them as essential tools in advancing artificial intelligence. The transparency and comprehensive release of training details in models like MAP-Neo highlight the importance of open research practices in propelling the field forward [42].

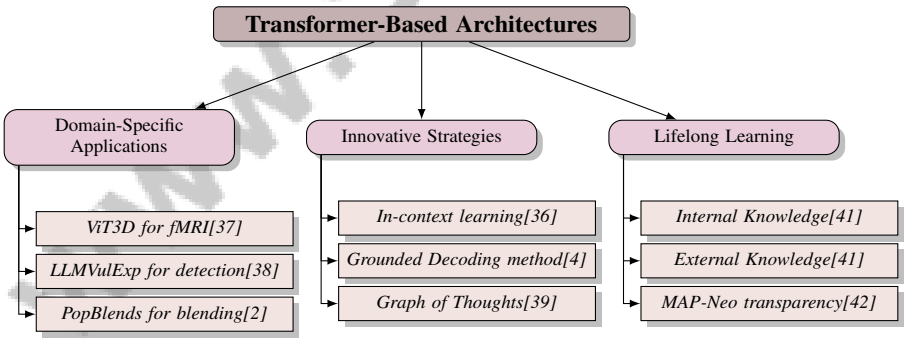

Figure 3: This figure illustrates the hierarchical structure of transformer-based architectures, categorizing them into domain-specific applications, innovative strategies, and lifelong learning approaches. Each category highlights key methods and frameworks that demonstrate the versatility and adaptability of transformers in various contexts.

3.2 Impact of Model Scale and Instruction Tuning

The scale of LLMs and instruction tuning methodologies significantly influence their performance and interpretability. As LLMs scale, they exhibit enhanced linguistic proficiency and task capability due to extensive parameterization. However, scaling introduces challenges, such as diminishing returns in performance linked to communication overhead when increasing the number of accelerators, affecting throughput [9].

Instruction tuning refines LLMs for specific tasks, enhancing accuracy and adaptability. This process tailors training to improve model precision, ensuring versatility across applications. A strategy to enhance reasoning involves decomposing tasks into sequential subquestions, thus improving problem-solving [43].

The computational demands of scaling LLMs increase with multimodal capabilities, requiring efficient strategies to balance performance with resource utilization, especially in critical applications like healthcare [44, 6, 45, 23, 46]. The interplay between model scale and instruction tuning guides researchers to enhance both performance and interpretability, fostering more robust and adaptable AI systems.

3.3 Multimodal Model Integration

Integrating multimodal data in LLMs marks a significant advancement, enabling the processing and generation of information across diverse data types, including text, images, and audio. Multimodal Large Language Models (MLLMs) leverage the synergy of modalities to enhance reasoning, providing a comprehensive understanding of complex scenarios. This approach is crucial for applications requiring extensive data interpretation, such as autonomous driving, domestic robotics, and open-world games [47].

Understanding MLLM architecture is vital for improving multimodal reasoning abilities. By effectively integrating various data types, these models achieve nuanced comprehension of tasks that single-modality models may struggle with, especially in scenarios requiring contextual information from multiple sources [48].

However, integrating multimodal data poses challenges, particularly regarding security and vulnerability. A specific threat model for MLLMs categorizes potential vulnerabilities and attacks, providing a framework to address security concerns, which is crucial for robustness in sensitive applications [49].

The incorporation of multimodal data enhances LLMs' ability to process intricate information, notably in fields like medicine and security. Nevertheless, this advancement presents critical security challenges that must be addressed to mitigate potential vulnerabilities, such as misinformation and exploitation by malicious actors. Recent studies emphasize the need for robust security frameworks to ensure LLMs deliver accurate information and maintain user trust, especially when providing security and privacy advice [49, 34, 45]. As research progresses, developing robust architectures and evaluation benchmarks will be essential in optimizing MLLM performance and security across diverse applications.

3.4 Specialized Architectures for Domain-Specific Applications

Specialized architectures in LLMs are increasingly tailored to meet specific application demands, enhancing performance and interpretability. Developing such architectures involves innovative strategies addressing unique challenges within particular domains. For example, the OWL model exemplifies a domain-specific architecture for IT operations, employing advanced techniques for context extension and parameter-efficient tuning to optimize performance [50].

In multimodal data processing, the MMNeuron method focuses on identifying and analyzing domain-specific neurons to improve understanding of how MLLMs process diverse information types [51]. This approach is crucial for applications needing the integration of various data modalities, such as emotion recognition, where Bayesian Cue Integration (BCI) combines facial expressions and contextual knowledge to enhance accuracy [52].

The concept of 'neural erosion' offers a novel perspective in modeling neurodegeneration within LLMs, contrasting with traditional image-based configurations and providing insights into domain-specific adaptations [53]. Additionally, the BEAN regularization method enhances interpretability by enforcing dependencies among neurons in dense layers, forming interpretable neuronal functional clusters without altering the deep neural network architecture [54].

Furthermore, integrating machine unlearning techniques allows models to align outputs with current ethical standards by enabling them to forget harmful past responses, maintaining ethical compliance

while preserving performance [55]. This capability is relevant in sensitive domains where ethical considerations are paramount.

The evolution of specialized architectures for domain-specific applications enhances LLMs' functionality and interpretability, ensuring adaptability to unique requirements across fields. As research progresses, developing frameworks for categorizing multimodal models will be pivotal in advancing the integration of diverse data types and the evolution of model architectures [44].

3.5 Architectural Modifications for Enhanced Interpretability

Architectural modifications in LLMs are crucial for enhancing interpretability, allowing insights into the internal workings and decision-making processes of these complex systems. The Document-Wise Memory Architecture (DWMA) tracks and recalls document memories by mapping document representations to specific memory entries, improving contextually accurate responses [56].

Another advancement is the Attention-MLP Interaction Analysis (AMIA), automating the identification of next-token neurons and the attention heads that activate them. This method clarifies interactions between model components, offering insights into how specific neurons and attention mechanisms contribute to output generation [57]. The Activation Variance-Sparsity Score (AVSS) aids interpretability by integrating normalized activation variance and sparsity to assess layer contributions, facilitating the identification of non-essential layers and optimizing model architecture [58].

The MRKL system exemplifies a modular design integrating external knowledge and reasoning modules, enhancing reasoning capabilities and addressing existing language models' limitations. This modularity allows the seamless incorporation of additional information sources, improving LLM interpretability and functionality [59]. Extending Quantum Natural Language Processing (QNLP) models to include image data represents a key innovation in multimodal data processing, creating architectures that represent syntactic and hierarchical structures in a unified framework, enhancing interpretability [60].

Despite these advancements, challenges remain in establishing a standard framework for evaluating interpretability, as studies often focus on a limited number of salient neurons and lack comprehensive evaluation criteria [28]. The issue of catastrophic forgetting, particularly in larger models, underscores the need for architectural strategies balancing knowledge retention with adaptability [61].

Architectural modifications aimed at enhancing interpretability are essential for advancing our understanding of LLMs. By addressing complex issues of memory retention, component interaction, and modular integration, these modifications contribute to the development of AI systems that enhance transparency and reliability. They enable effective tracking of document-related content and improve user understanding of AI-generated outputs through innovative architectures that map document representations to memory entries and employ human-centered approaches to transparency, considering the diverse needs of stakeholders engaging with LLMs [56, 62, 63, 21].

4 Neuron-Level Interpretability

Neuron-level interpretability in neural networks is a significant focus in AI research, aiming to elucidate the complex mechanisms of deep neural networks (DNNs). This interpretability enhances transparency and reliability by providing insights into individual neuron contributions to model predictions, which is crucial for debugging and building trust in critical applications like healthcare and autonomous systems. Techniques that analyze DNN structures systematically promote advancements in safety and fairness [30, 63, 64]. A comprehensive understanding of neuron-level interpretability sets the foundation for exploring neuron roles and interactions within neural networks.

4.1 Conceptual Framework of Neuron-Level Interpretability

Neuron-level interpretability focuses on understanding the functions and interactions of individual neurons within complex models. Traditional methods often emphasize high activation levels, but a more comprehensive approach considers neurons as linear combinations of various concepts, providing deeper insights into their roles and the broader representations learned by neural networks

[30, 28, 29]. Mechanistic Interpretability (MI) seeks to reverse-engineer neural networks, extracting algorithms that are understandable and align with human cognitive processes [65].

The identification of 'value neurons' and 'query neurons' marks a significant advancement, offering nuanced insights into neuron importance [66]. Safety alignment mechanisms in large language models (LLMs) reveal neurons contributing to safety behaviors, which is essential for ethical AI deployment [24]. Understanding the confidence-competence gap in LLMs, where models may exhibit overconfidence, is critical for ensuring that high-confidence outputs align with accuracy [67].

Systems like PopBlends integrate divergent and convergent thinking processes, showcasing potential for creative problem-solving within neural networks [2]. This framework encompasses methodologies such as mechanistic analysis, neuron identification, and cognitive process integration, enhancing our understanding of complex models and decision-making behaviors [30, 65, 29].

4.2 Innovative Approaches to Neuron Analysis

Recent advancements in neuron analysis have introduced methodologies that enhance neural network interpretability by focusing on individual neuron behaviors. Neurons conceptualized through a vocabulary of concepts allow automated semantic explanations of neuron behaviors, providing structured descriptions of neuron roles and interactions [68]. Selectivity-guided probing accounts for memorization effects, emphasizing the importance of understanding selective neuron responses [69].

The Neuron-Level Attribution Method (NLAM) uses log probability increase as an importance score to identify significant neurons in LLMs [66]. Additionally, generation-time activation contrasting and dynamic activation patching focus on open-ended generation and the identification of safety neurons, offering a dynamic perspective on neuron behavior [24].

MI highlights the non-identifiability of explanations, suggesting multiple valid interpretations for a given behavior [65]. These innovative approaches collectively enhance neuron analysis, providing deeper insights into neural network functions and moving beyond traditional high-activation analyses [27, 70, 29, 66, 71]. As illustrated in Figure 4, the categorization of these methods into semantic explanations, neuron attribution, and interpretability challenges underscores significant advancements and methodologies aimed at enhancing the understanding and interpretability of neural networks.

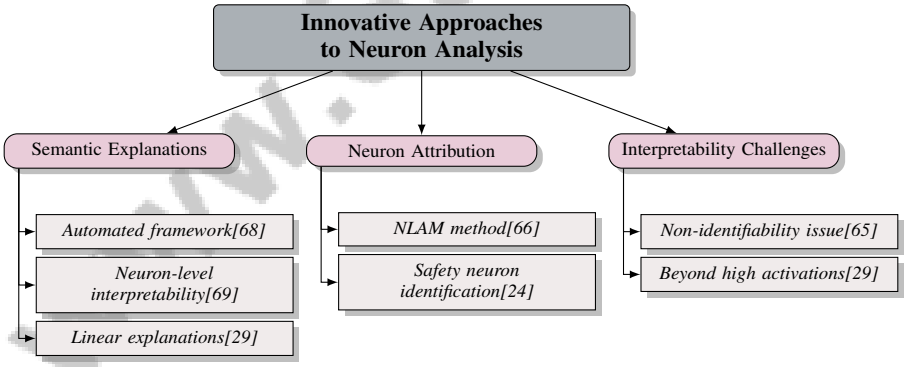

Figure 4: This figure illustrates the innovative approaches to neuron analysis, categorizing methods into semantic explanations, neuron attribution, and interpretability challenges. Each category highlights significant advancements and methodologies for enhancing the understanding and interpretability of neural networks.

4.3 Tools and Frameworks for Enhanced Interpretability

The development of specialized tools and frameworks has significantly advanced neuron-level interpretability by clarifying neural network workings. The Attention Lens applies learned transformations to attention head outputs, enhancing intuitive interpretation by aligning outputs with human-understandable concepts [72].

The Concept Activation Matrix (CAM) captures neuron activation in response to specific concepts, predicting neuron activations through a sparse linear model, which structures and interprets neuron

behavior [29]. The Audio Network Dissection (AND) framework analyzes audio networks by examining neuron responses to audio inputs, providing insights into audio processing tasks [73].

These tools and frameworks collectively advance neuron-level interpretability by offering diverse methodologies for analyzing complex interactions within neural networks, enhancing transparency and structured insights into neuron behaviors [27, 63, 64, 74].

4.4 Applications and Implications of Neuron-Level Interpretability

Neuron-level interpretability enhances the transparency and reliability of Large Language Models (LLMs) across various domains. In anomaly detection, it supports techniques like supervised fine-tuning (SFT) and in-context learning (ICL), improving LLM performance and providing interpretable solutions [8]. This interpretability is crucial for understanding and mitigating biases in models like GPT-3 [75].

In clinical settings, neuron-level interpretability refines models for complex decision-making, enhancing diagnostic accuracy and patient care. In psychotherapy, it enriches interactions in psychological dialogue systems [19].

Neuron-level interpretability also extends to security domains, highlighting the need for robust security measures in sensitive applications [51]. In synthetic data generation, it addresses privacy concerns and ethical implications, ensuring generated data is diverse and relevant [76].

Identifying sparse yet effective safety neurons provides insights into the alignment tax phenomenon, crucial for preventing harmful or biased content generation [24]. Neuron-level interpretability significantly advances AI systems' transparency, fairness, and applicability across diverse domains, enhancing trustworthiness in addressing real-world challenges [11].

4.5 Future Directions in Neuron-Level Interpretability Research

Future research in neuron-level interpretability promises advancements in understanding neural networks and enhancing transparency. Integrating interpretability frameworks with reinforcement learning could provide deeper insights into neural networks' adaptation to complex environments [77].

Frameworks like RAVEL offer opportunities to explore diverse neural network configurations, improving granularity and precision in neuron-level analysis [78]. Optimizing sparsity techniques remains critical, especially with emerging hardware architectures, enhancing training efficiency and scalability [1].

Highlighted future directions emphasize innovation in neuron-level interpretability, particularly for large language models (LLMs), improving neuron explanation quality through prompt tuning, and exploring feed-forward layers in transformer architectures. These advancements are vital for ensuring LLMs' safety and reliability in high-stakes applications across fields such as healthcare, education, and law [30, 28, 74, 31]. By exploring these opportunities, researchers can contribute to developing AI systems that are more transparent, interpretable, and ethically aligned, capable of addressing complex real-world challenges.

5 Model Transparency and Explainable AI

The importance of model transparency has become increasingly prominent in discussions about artificial intelligence as AI systems grow more complex. This complexity necessitates a comprehensive exploration of explainable AI (XAI) principles, which aim to enhance the interpretability and trustworthiness of AI models. The following subsection delves into the fundamental principles of XAI and their role in achieving transparency in AI systems.

5.1 Principles of Explainable AI

Explainable AI (XAI) plays a crucial role in modern AI systems by emphasizing the need for transparency and clarity in AI decision-making. A key principle is providing clear, interpretable explanations for AI outputs, essential for fostering trust and facilitating adoption in high-stakes areas

such as healthcare, finance, and autonomous systems [79]. Interpretability can be achieved through models like Deep Logic Networks (DLNs), which offer logic-based explanations while maintaining efficiency and accuracy comparable to traditional neural networks [71].

Another critical aspect of XAI is the iterative refinement of AI models to ensure adaptability to dynamic data and contexts. Integrating generative capabilities with evaluation processes helps mitigate erroneous outputs [80], which is vital for maintaining AI reliability. Uncertainty quantification is also integral to XAI, providing a framework for assessing model confidence across scenarios. This enhances AI output reliability by structuring model prediction understanding, especially in crucial decision-making applications. Specialized benchmarks, such as the Facet-aware Metric (FM) and NPHardEval, allow for granular assessments, addressing inadequacies in traditional evaluation methods and offering a rigorous evaluation framework for large language models (LLMs) [81, 23, 82, 83].

Privacy protection is another essential XAI aspect, with metrics ensuring compliance with privacy guidelines while maintaining transparency. The ExplainableDetector method exemplifies XAI's potential in specialized applications, achieving 99.84

As illustrated in Figure 5, the core principles of Explainable AI (XAI) focus on interpretability, model reliability, and privacy and trust. This figure highlights significant methods and benchmarks, such as Deep Logic Networks, LLM-Assisted Inference, and ExplainableDetector, emphasizing their roles in enhancing transparency, adaptability, and compliance with privacy guidelines.

XAI principles prioritize enhancing transparency, interpretability, and reliability in AI systems, ensuring alignment with ethical standards and human values while supporting understanding across diverse stakeholders. By focusing on user engagement and addressing complex AI model challenges, particularly in high-stakes areas like education and criminal justice, XAI aims to build trust and mitigate risks associated with algorithmic bias and misinformation. It advocates for systematic approaches in providing explanations, communicating uncertainty, and evaluating AI performance to foster responsible and equitable AI applications [21, 84, 18]. As AI technologies evolve, integrating these principles will be vital for promoting widespread adoption across various domains.

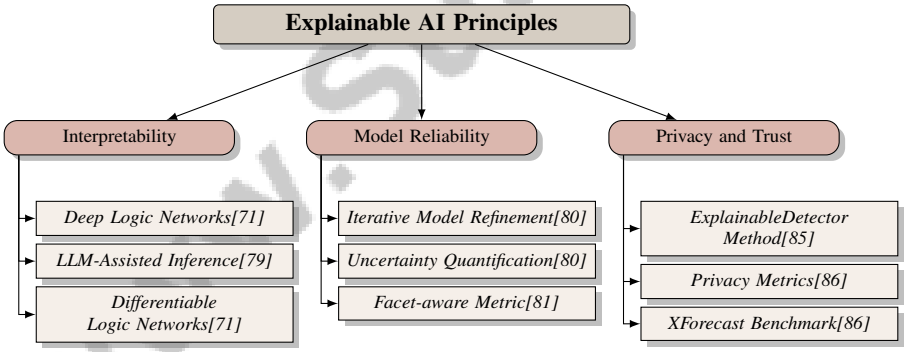

Figure 5: This figure illustrates the core principles of Explainable AI (XAI), focusing on interpretability, model reliability, and privacy and trust. It highlights significant methods and benchmarks like Deep Logic Networks, LLM-Assisted Inference, and ExplainableDetector, emphasizing their roles in enhancing transparency, adaptability, and compliance with privacy guidelines.

5.2 Techniques for Enhancing Transparency

Enhancing model transparency is essential in developing explainable AI, focusing on methodologies that elucidate the inner workings of complex models like Large Language Models (LLMs). The TraceFL framework captures neuron provenance to trace model predictions back to contributing clients, providing actionable insights and enhancing transparency by revealing output origins [87].

Explainable AI techniques are crucial for understanding LLMs by demystifying model behavior and decision-making processes [88]. Techniques like the SECToR framework employ self-consistency checks to mitigate error avalanching, allowing sustained self-improvement over iterations. This enhances transparency by ensuring models can iteratively refine outputs, maintaining reliability and reducing cascading error risks [89].

The MMNeuron method exemplifies using domain-specific neurons to enhance model transparency. By focusing on neurons processing domain-relevant information, this method improves the model's ability to provide interpretable outputs tailored to specific applications [51]. Such domain-specific focus is crucial for applications requiring context-specific understanding for accurate decision-making.

These techniques collectively enhance model transparency by providing structured insights into AI decision-making processes, facilitating a deeper understanding of their inner workings. This is vital for building trust in AI, enabling bias identification, supporting debugging efforts, and aligning AI outputs with human expectations across diverse applications, especially in deep neural networks and LLMs. By systematically categorizing interpretability methods and emphasizing a human-centered approach, these advancements address the unique challenges posed by complex AI systems and contribute to responsible AI deployment [84, 63, 22, 21, 62]. Employing methodologies that trace neuron contributions, ensure self-consistency, and leverage domain-specific knowledge enables researchers to develop AI models that are powerful, efficient, transparent, and trustworthy.

5.3 Applications and Case Studies

The practical application of explainable AI (XAI) is illustrated through various case studies that highlight its effectiveness across different domains. The GraphLLM framework significantly enhances the graph reasoning ability of LLMs, demonstrating an average accuracy improvement of 54.44

In scientific reasoning, GPT-4 engages in Socratic reasoning, producing proof schemas and rigorous reasoning processes, serving as a case study of XAI in practice and illustrating its role in complex problem-solving [90]. The proto-lm framework also delivers competitive performance on various natural language processing tasks while providing high-quality, interpretable explanations, reinforcing the coexistence of interpretability and performance in XAI [91].

The LUNA framework offers a robust model-based universal analysis approach, effectively distinguishing normal from abnormal behaviors in LLMs and providing a comprehensive framework for quality analysis across trustworthiness perspectives, emphasizing XAI's role in ensuring reliable AI behavior [92]. Improved performance on benchmarks like MathVista and M3CoT highlights the practical application of XAI techniques, demonstrating their ability to enhance reasoning capabilities in models processing diverse data types [48].

Variability in the quality of explanations based on input parameters has been observed, with some models outperforming others in correctness and completeness, emphasizing the need for careful consideration of model parameters to optimize interpretability and effectiveness [17].

These case studies collectively demonstrate the transformative effects of XAI across various applications, emphasizing its crucial role in enhancing model performance, ensuring algorithmic transparency, and building trust in AI systems. They address the necessity for clear explanations of AI decision-making processes to mitigate biases, improve user understanding, and foster responsible AI development. By analyzing the unique challenges of implementing XAI in diverse domains such as criminal justice, education, and LLMs, these studies highlight best practices and future research directions aimed at creating more effective, equitable, and user-centered AI solutions [21, 84, 18].

6 Neuroscience of AI

The convergence of neuroscience and artificial intelligence (AI) provides a rich framework for exploring cognitive processes that inform intelligent system design. This interdisciplinary relationship necessitates examining theoretical models that illustrate AI's potential to emulate human cognition. Lacanian theory, for example, offers insights into self-identification in AI, enhancing our understanding of identity formation within intelligent agents. This perspective encourages deeper exploration of mechanisms underlying AI interactions with users and environments. The subsequent subsection explores the application of Lacanian theory to AI self-identification, emphasizing its importance in advancing our understanding of intelligent systems in relation to human psychological constructs.

6.1 Lacanian Theory and AI Self-Identification

Lacanian theory, based on Jacques Lacan's psychoanalytic frameworks, provides a unique lens for examining self-identification in AI. This approach focuses on identity formation through symbolic and

imaginary orders, offering insights into how AI systems might develop a form of self-identification. Applying Lacanian theory to AI involves analyzing how intelligent agents can recognize and interpret their 'self' within a network of symbolic relations, akin to human self-identification processes [93].

In this context, AI self-identification extends beyond traditional notions of self-awareness, encompassing the ability of AI systems to identify their roles within broader interaction systems. Leveraging Lacanian concepts allows researchers to develop models simulating complex self-recognition and identity formation processes, leading to more autonomous and adaptive AI systems. This approach has significant implications for creating AI capable of nuanced interactions with human users, enhancing their understanding of and ability to anticipate human needs and behaviors. By integrating computational psychoanalysis and active inference, AI can utilize structured frameworks for self-identity, promoting consistent self-recognition and memory continuity. Such advancements improve interaction quality and enable AI systems to exhibit sophisticated comprehension of human emotions and intentions, ultimately leading to more effective and personalized user experiences [93, 77].

Exploring Lacanian theory in AI opens new avenues for enhancing the autonomy and adaptability of intelligent agents, enriching the conceptual foundations of AI self-identification, and contributing to the broader discourse on the intersection of psychoanalysis and artificial intelligence.

6.2 Neuronal Dynamics and Attention Mechanisms

Neuronal dynamics and attention mechanisms play a pivotal role in shaping AI models, particularly neural networks. These components facilitate the processing and interpretation of complex data, enabling models to focus on relevant information while filtering out noise. Neuronal dynamics involve the temporal changes and interactions within neural networks essential for learning and adaptation, characterized by their nonlinear and nonstationary nature. Recent models like NetFormer address dynamic aspects of neuronal activity, capturing synaptic plasticity and neuronal modulation through time-dependent attention mechanisms that represent evolving connectivity structures in response to local activity patterns [94, 95]. These dynamics are crucial for developing models that generalize across tasks and environments, mimicking the adaptive nature of biological neural systems.

Attention mechanisms provide a framework for selectively concentrating on specific input data parts, enhancing the model's ability to capture contextual relationships and dependencies. This selective focus is achieved through mechanisms like self-attention, which evaluates the importance of each element in the input sequence relative to others, allowing the model to weigh their significance in decision-making processes. The Transformer architecture, now state-of-the-art in natural language processing, utilizes self-attention mechanisms to analyze and process sequences in parallel, enhancing efficiency and accuracy by weighing the importance of words contextually. Feed-forward layers further process information extracted by self-attention layers, improving overall model performance in tasks like next-token prediction and contextual understanding [56, 57, 28].

Incorporating attention mechanisms into neural networks enhances interpretability and improves performance in sequential data processing tasks, such as language translation and image captioning. The interplay between neurons and the specialized focus provided by attention mechanisms in large language models (LLMs) facilitates a processing approach that mirrors human cognition. This dynamic interaction enhances models' ability to interpret context and generate nuanced responses, highlighting their potential applications in complex decision-making scenarios, including healthcare, law, and finance. Recent studies indicate that specific attention heads excel at recognizing contextual cues that inform next-token predictions, enriching the model's interpretability and effectiveness in context-sensitive tasks [72, 57, 74, 31].

The synergy between neuronal dynamics and attention mechanisms is essential for developing AI systems that are both powerful and interpretable. By systematically integrating nuanced domain knowledge into quantifiable features through LLMs, researchers can design advanced models that enhance predictive analytics, improve risk assessment, and facilitate complex reasoning and adaptable learning. This approach preserves critical human expertise while scaling its impact across various prediction tasks, significantly advancing the field of artificial intelligence [96, 22, 62, 97, 5].

6.3 Alignment with Neuroscientific Constructs

Aligning AI models with neuroscientific constructs is an emerging research area that seeks to draw parallels between AI systems and human brain functioning. This interdisciplinary approach enhances AI model interpretability and provides insights into cognition and perception processes. For instance, exploring large language models (LLMs) through Lacanian theory positions these models as avatars of the Lacanian Other, emphasizing the role of symbolic and imaginary orders in self-identification processes [93]. This perspective highlights the potential for AI systems to simulate complex identity formations akin to human cognitive processes.

The mathematical foundations of self-identity in AI are further elucidated through metric space theory, measure theory, and functional analysis. These frameworks suggest that self-identity in AI emerges from measurable conditions, offering a structured approach to understanding how AI models can develop self-awareness and adaptability [77]. Such theoretical constructs provide a robust foundation for aligning AI models with neuroscientific principles, facilitating the development of more autonomous and context-aware systems.

As illustrated in Figure 6, the alignment of AI models with neuroscientific constructs emphasizes the integration of AI interpretability with cognitive and perception processes, as well as the mathematical foundations for self-identity. Advocating for neuroscientific interpretability as a complementary pursuit to AI interpretability emphasizes the potential for cross-disciplinary insights that enhance our understanding of both biological and artificial networks [31]. By examining the synergies between AI systems and neuroscientific constructs, researchers can develop models that mimic human cognitive abilities while providing new perspectives on intelligence and consciousness.

Integrating AI models with neuroscientific constructs represents a significant direction in AI research, enhancing our understanding of artificial intelligence and offering insights into natural intelligence. Utilizing neuro-symbolic AI techniques improves model interpretability, enabling complex analogies and engaging with empirical data more effectively. This approach allows for clearer mapping of AI features to neuroscientific principles, facilitating hypothesis testing and evidence discernment in scientific research, particularly in social sciences. Ultimately, this alignment fosters a deeper comprehension of cognitive processes and supports the development of AI systems that align more closely with human-like reasoning and understanding [97, 96, 31]. Through integrating theoretical frameworks and interdisciplinary collaboration, this approach holds potential to revolutionize AI system design and functionality, making them more interpretable, adaptable, and aligned with human cognitive processes.

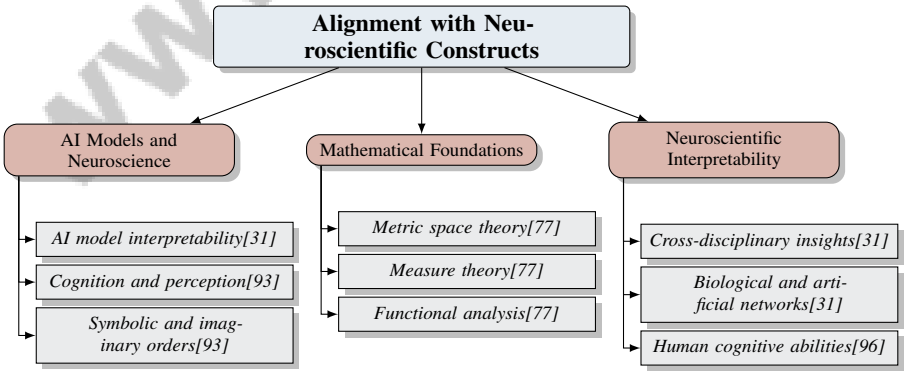

Figure 6: This figure illustrates the alignment of AI models with neuroscientific constructs, emphasizing the integration of AI interpretability with cognitive and perception processes, mathematical foundations for self-identity, and the pursuit of neuroscientific interpretability to enhance understanding of both biological and artificial networks.

7 Neural Network Analysis Techniques

7.1 Concept Activation and Sparse Linear Models

Concept activation techniques are integral to enhancing neural networks' interpretability, especially within sparse linear models. These techniques enable the identification and activation of neurons linked to specific concepts, offering a detailed understanding of model behavior. For instance, the PopBlends system utilizes strategies like No-GPT, Half-GPT, and Full-GPT to automate conceptual blend suggestions, each with distinct advantages and limitations [2]. In sparse linear models, concept activation reduces complexity while preserving interpretability by focusing on a limited set of active neurons, aligning with explainable AI goals such as model safety and bias detection. Studies show that refining prompts can enhance natural language explanations, clarifying individual neurons' contributions to concept understanding [27, 74]. This approach not only boosts transparency but also aids in identifying key features driving model predictions, enhancing applicability across various domains.

Integrating concept activation in sparse linear models represents a significant advancement in neural network analysis. This integration clarifies neuron contributions to high-level concepts and enhances AI system development efficiency by leveraging sparsity to optimize performance. By evaluating layer importance and activations, researchers can identify non-essential components, leading to streamlined architectures that maintain high performance across natural language processing tasks [1, 78, 28, 29, 58]. Continued refinement of these techniques will be crucial for advancing the capabilities and understanding of complex neural architectures.

7.2 Automated Model Interpretation with MAIA

The Model-Agnostic Interpretability Algorithm (MAIA) advances automated model interpretation by offering a flexible framework applicable to various neural network architectures. By integrating tools for feature interpretation and failure mode discovery, MAIA enhances model understanding, enabling tasks like input synthesis, misclassification identification, and experimental result summarization. Its effectiveness in computer vision applications demonstrates its ability to produce neuron-level feature descriptions comparable to expert-generated ones, addressing critical interpretability challenges [98, 84, 91, 92]. MAIA generates interpretable model predictions, improving AI systems' transparency and trustworthiness by analyzing internal representations and identifying key decision-driving features.

MAIA's model-agnostic nature allows application across various neural networks without requiring architectural changes, essential for achieving interpretability in diverse applications, including natural language processing and computer vision [99, 98, 28]. By utilizing MAIA, researchers gain deeper insights into how models process information and make predictions, facilitating bias identification and improvement areas. It employs advanced feature attribution techniques, assigning importance scores to features based on their contributions to model outputs, enhancing understanding of neural models. This involves integrating tools for iterative experimentation on model subcomponents, identifying sensitive features, systematic errors, and enhancing robustness through informed adjustments to training data and architecture [98, 22]. Thus, MAIA serves as a powerful tool for enhancing AI systems' reliability and accountability, particularly in high-stakes domains where interpretability is vital.

The development and implementation of MAIA underscore the critical role of automated interpretability in enhancing artificial intelligence. By automating neural models' understanding through tasks like feature interpretation and failure mode discovery, MAIA advances the field's ability to create more transparent and reliable AI systems [98, 62]. Providing a framework for understanding complex model behavior, MAIA contributes to ongoing efforts to create powerful and transparent AI systems, fostering greater trust and confidence in their deployment across various sectors.

7.3 Attention Lens and Vocabulary Space Transformations

Integrating attention mechanisms and vocabulary space transformations is crucial for enhancing neural networks' interpretability and analytical capabilities, particularly in Large Language Models (LLMs). The Attention Lens tool exemplifies this integration by applying learned transformations to attention head outputs, converting them into a vocabulary space that aligns with human-understandable concepts

[72]. This transformation facilitates intuitive model behavior interpretation, allowing researchers to understand how specific attention mechanisms contribute to output generation.

Vocabulary space transformations elucidate semantic relationships within data processed by LLMs, enhancing keyword extraction and data enrichment vital for effective information retrieval, document summarization, and content categorization. Utilizing advanced LLM architectures, such as GPT-3.5 and Llama2-7B, researchers can better annotate underlying data, bridging the gap between human comprehension and machine processing [46, 23]. Mapping neural network outputs into a structured vocabulary space enables the identification of key features and patterns driving model predictions. This approach enhances neural network transparency and aids in bias and error detection, thereby improving AI systems' reliability and trustworthiness.

The use of attention lens and vocabulary space transformations signifies a substantial advancement in model analysis, offering a robust framework for exploring attention mechanisms' intricate dynamics within neural networks. As AI research progresses, refining techniques that integrate human expertise, validate AI-generated analyses, and enhance model interpretability will be crucial for improving complex AI models' functionality and reliability, particularly in predictive analytics and scientific hypothesis testing [96, 62, 23, 22].

7.4 Neural Explanation Techniques

Neural explanation techniques are essential for elucidating complex neural networks' behavior, offering insights into how these models process information and make decisions. Concept Activation Vectors (CAVs) interpret internal model representations by associating them with human-understandable concepts, enhancing transparency by identifying neurons or layers encoding distinct features [63, 64, 29, 30, 57]. This capability is crucial for diagnosing issues, refining model performance, and ensuring reliability in safety-critical applications, as it connects neuron activations to interpretable concepts, improving trust in AI systems.

Layer-wise Relevance Propagation (LRP) assigns relevance scores to input features based on their contributions to predictions, enhancing understanding of specific features influencing outcomes [56, 66, 23, 10]. LRP allows researchers to trace decision-making processes back to their origins, making it useful for identifying biases and understanding features' roles in model behavior.

Saliency maps provide visual representations of input areas influencing model predictions, elucidating decision-making processes in deep learning models, particularly in architectures like transformers [37, 28, 64, 29, 57]. By highlighting influential input regions, saliency maps offer intuitive insights into neural networks' focus during decision-making processes, valuable for understanding spatial attention in computer vision applications.

These techniques underscore the importance of interpretability in neural network analysis. By offering structured insights into model behavior, these methods enhance AI systems' transparency and trustworthiness. They enable comprehension of complex algorithms' decision-making processes, addressing algorithmic fairness and potential biases in training data. As AI systems are increasingly deployed in critical applications, the ability to explain decisions is essential for ensuring reliability and ethical use. By facilitating a clearer understanding of model operations, neural explanation methods contribute to developing powerful and efficient AI technologies while promoting accountability and user confidence [84, 74]. As research progresses, refining and integrating these methods will be essential for advancing complex neural architectures' understanding and reliability.

7.5 Local Learning and Infomorphic Neurons

Local learning and infomorphic neurons represent significant advancements in understanding and innovating neural network architectures. These developments emphasize learning processes and neuron functionality by facilitating self-organized, local learning dynamics inspired by biological systems. This approach leverages information theory principles, particularly Partial Information Decomposition (PID), enabling neurons to integrate and process diverse input types while promoting interpretability and robustness in performance. Recent methodologies in neuron-level knowledge attribution enhance understanding of specific neurons' contributions to predictions in large language models, paving the way for improved interpretability and targeted knowledge manipulation [100, 29, 51, 66, 71]. Local learning adapts parameters based on localized information, allowing for more

efficient and targeted learning, particularly advantageous when data is heterogeneous or computational resources are limited.

Infomorphic neurons capture and represent information in alignment with data structures, dynamically adjusting activation functions and connectivity based on processed information. Inspired by information morphology, these neurons enhance interpretability and performance by synchronizing operational characteristics with input data properties. This alignment is achieved through bio-inspired local learning objectives utilizing PID to assess and integrate diverse information inputs, optimizing neurons' contributions based on unique, redundant, or synergistic information. This approach clarifies local information processing within the model and strengthens efficacy, paving the way for robust and interpretable architectures [54, 64, 100, 74].

Together, local learning and infomorphic neurons provide a framework for developing adaptable and interpretable neural networks. By emphasizing localized adaptation and dynamic information representation, these concepts improve neural networks' capacity to learn from complex data sources. Mechanistic interpretability techniques identify specific model components associated with distinct capabilities, enabling precise editing and unlearning of undesirable knowledge. For instance, localizing edits to components linked to factual recall mechanisms strengthens knowledge unlearning across diverse input/output formats and mitigates unintended side effects. Additionally, neuron-level analysis methods identify crucial neurons contributing to predictions, refining understanding of knowledge storage mechanisms within large language models. These advancements facilitate a nuanced and effective learning process from heterogeneous data [66, 101]. As research evolves, integrating local learning and infomorphic neurons into existing architectures promises to enhance AI systems' capabilities, making them more responsive to real-world challenges and better aligned with human cognitive processes.

7.6 Interactions Between Attention Heads and MLP Neurons

Interactions between attention heads and multi-layer perceptron (MLP) neurons in large language models (LLMs) represent a complex dynamic significantly influencing predictive capabilities. While attention heads and MLP neurons are often studied independently, their combined effects on next-token prediction remain an area ripe for exploration [57]. Attention heads are integral to the self-attention mechanism, allowing models to weigh the importance of different input tokens and capture long-range dependencies, crucial for tasks requiring contextual understanding.

MLP neurons play a vital role in non-linear input feature transformation, enhancing the model's capacity to learn intricate patterns and relationships. This interaction between attention mechanisms and MLP neurons facilitates more accurate next-token predictions by enabling the model to recognize context-specific features and activate relevant downstream neurons accordingly. Understanding these interactions sheds light on LLMs' internal workings and contributes to advancing mechanistic interpretability in neural networks [57, 29]. The interplay between attention heads and MLP neurons is symbiotic, where attention mechanisms provide the contextual framework, and MLP neurons refine this information through intricate transformations, influencing the model's ability to generate coherent and contextually appropriate outputs.

The paper by [57] underscores the need for a deeper understanding of these interactions, suggesting that comprehensive analysis could lead to significant improvements in model interpretability and performance. By examining how attention heads and MLP neurons collaboratively influence next-token predictions, researchers can develop more sophisticated models that better mimic human-like reasoning and decision-making processes. This exploration enhances LLM transparency and provides insights into optimizing architectures for more efficient and accurate language processing.

7.7 Framework for Assessing Interpretability with R AVEL

The R AVEL framework represents a significant advancement in assessing interpretability methods within neural networks, focusing on disentangling attributes through causal analysis and intervention techniques. This innovative approach provides a structured methodology for evaluating how effectively interpretability techniques isolate and clarify individual model components' contributions to overall behavior [78]. By incorporating causal analysis, R AVEL enables researchers to identify and manipulate specific attributes within neural networks, offering insights into the causal relationships driving model outputs.

Benchmark	Size	Domain	Task Format	Metric
CF-Benchmark[61]	100,000	Natural Language Processing	Text Generation	Accuracy, FG
NUMCoT[102]	1,600	Mathematics	Numeral Conversion	Accuracy, F1-score
LLM-OPT[103]	5,000	Mathematics	Optimization	Goal Metric, Policy Metric
CommBench[104]	1,000,000	Natural Language Processing	Language Modeling	Communication Volume, Latency
SHE[96]	638	Social Sciences	Hypothesis-Evidence Relationship Classification	macro F1
PsyEval[105]	28,186	Mental Health	Question Answering	F1, AUC-ROC
REMLM[106]	1,000	Knowledge Editing	Dialogue Generation	Accuracy, Reversion
XForecast[86]	1,000	Time Series Forecasting	Forecasting	Direct Simulatability, Synthetic Simulatability

Table 1: Table summarizing various benchmarks utilized in evaluating interpretability within neural networks, detailing their size, domain, task format, and metrics employed. These benchmarks span diverse fields, including natural language processing, mathematics, social sciences, mental health, knowledge editing, and time series forecasting, providing comprehensive insights into model performance across different tasks.

Intervention techniques within the R AVEL framework allow targeted modifications to neural network components, enabling observation of resultant changes in model behavior. This capability is crucial for understanding causal pathways within neural networks and assessing individual neurons' or layers' impact on final predictions. The framework's focus on causal disentanglement enhances complex models' interpretability by employing rigorous analytical methodologies, ensuring systematic assessments capable of providing unique and coherent explanations of model behavior. This approach aligns with Mechanistic Interpretability principles, which seek to extract understandable algorithms from neural networks, addressing non-identifiability challenges in explanations. By leveraging strategies mapping relationships between model activations and human-understandable concepts, the framework thoroughly evaluates interpretability techniques' effectiveness in clarifying advanced AI systems' intricate workings [78, 65, 84]. Table 1 presents a detailed overview of the benchmarks employed in the R AVEL framework, highlighting their characteristics and relevance in assessing interpretability methods within neural networks.

The R AVEL framework offers a novel and rigorous approach to assessing interpretability in neural networks, enhancing our understanding of model behavior through causal analysis and targeted interventions. This framework propels interpretability research forward and plays a crucial role in enhancing AI systems' transparency and trustworthiness by providing effective tools for analyzing deep neural networks' inner workings, identifying potential biases, debugging issues, and ensuring algorithmic fairness. By focusing on intrinsic and post hoc interpretability methods, this framework aims to bridge the gap between complex AI models and user understanding, ultimately fostering greater confidence in AI applications across various domains [63, 28, 84].

8 Challenges and Future Directions

8.1 Emerging Challenges and Innovations

Neuron-level interpretability and transparency in large language models (LLMs) present significant challenges and innovations. Existing benchmarks often fail to capture the nuances of developer attention across coding environments, limiting their effectiveness in evaluating interpretability [33]. This highlights the need for comprehensive evaluation frameworks that reflect real-world complexities. Moreover, the complexity of AI models and varying AI literacy levels among educators and students impede the effective deployment of human-centric explainable AI in education, necessitating tools to bridge this gap [18].

In ethical AI, frameworks that guide the moral reasoning of LLMs are increasingly important to ensure alignment with human values [14]. From a technical perspective, optimizing large-scale training on existing hardware architectures remains challenging. Future research should explore parallelization techniques to enhance efficiency across diverse hardware configurations [9]. The subjective nature of creativity assessment, reliant on individual perceptions, calls for objective metrics to evaluate AI-generated creativity [26].

Specialized domains like dentistry face challenges related to data privacy, the need for high-quality training data, and potential model biases [7]. Addressing these is crucial for ethical AI deployment. Additionally, current methods for identifying safety neurons depend on pre-aligned models, necessitating more robust and independent safety assurance methods [24].

8.2 Binary Logic Operators and Differentiable Learning

Binary logic operators and differentiable learning enhance AI interpretability by integrating logical reasoning with continuous optimization. Differentiable logic networks (DLNs), utilizing layers of binary logic operators, allow gradient-based training while maintaining interpretability, addressing the "black-box" nature of traditional neural networks (NNs) and simplifying inference for resource-constrained environments. These methods achieve competitive classification accuracy and transparency, essential for sensitive fields like healthcare and finance [71, 65, 84]. Incorporating binary logic operators aligns model outputs with human-understandable structures, enhancing interpretability in safety-critical systems.

Differentiable learning refines model parameters through continuous optimization, enabling intricate pattern identification while addressing interpretability and overfitting challenges. By integrating binary logic operators and softening techniques, differentiable logic networks facilitate insights from high-dimensional data while maintaining simplicity for resource-constrained environments [79, 71, 100, 6]. This methodology is crucial for developing adaptable models across diverse tasks.

Frameworks like NetFormer exemplify the synergy between binary logic operators and differentiable learning, capturing nonstationary connectivity and yielding interpretable results aligned with biological observations [95]. This integration underscores the potential for creating powerful, transparent AI systems aligned with human cognition.

8.3 Enhancing Multimodal and Multidisciplinary Approaches

Enhancing multimodal and multidisciplinary approaches in AI is crucial for advancing the field, integrating diverse data types, and fostering collaboration across disciplines. Multimodal AI models, such as Multimodal Large Language Models (MLLMs), combine modalities—text, images, audio—to improve reasoning and understanding of complex scenarios [47]. This integration is vital for applications like autonomous driving and domestic robotics [48].

Robust evaluation frameworks, like specific threat models for MLLMs, categorize vulnerabilities and establish security frameworks [49]. Such frameworks ensure MLLM robustness and reliability in sensitive applications. Interdisciplinary collaboration fosters AI innovation and applicability across domains, aligning AI models with neuroscientific constructs for improved interpretability and insights into cognition and perception [31].

Enhancing multimodal and multidisciplinary approaches addresses real-world challenges by synthesizing diverse data types and promoting interdisciplinary collaboration. This integration leverages advanced AI systems, including large language models (LLMs), to improve robustness and adaptability, translating expert domain knowledge into quantifiable features for predictive analytics. Transparency and validation of AI-generated analyses are critical for user trust and understanding, leading to effective and responsible AI applications [21, 22, 62, 107].

8.4 Improving Robustness and Efficiency

Enhancing AI model robustness and efficiency is critical as models are deployed in complex, dynamic environments. Scalability is a key challenge, especially with complex, multimodal datasets. Integrating text, images, and audio requires efficient computational strategies to handle complexity without compromising performance [77].

Continuous learning scenarios require exploration of long-term model stability. Adapting models to new data and tasks can lead to catastrophic forgetting, necessitating strategies for maintaining stability and integrity in real-world applications [77].

Techniques like efficient parameter tuning, adaptive learning rates, and regularization enhance AI model robustness and efficiency. These optimize parameters and mitigate overfitting through document-wise memory architecture and document guidance loss, enhancing generalization to

unseen data [96, 108, 23, 22, 56]. Parallel computing and distributed systems significantly enhance computational efficiency, allowing effective large-scale dataset processing.

Improving AI model robustness and efficiency requires integrating methodological innovations, like mechanistic interpretability for precise knowledge editing, with computational strategies, including large language models (LLMs) for systematic reviews and predictive analytics. This addresses challenges of updating model knowledge and maintaining performance consistency, leveraging LLMs' strengths in discerning scientific evidence, advancing communicative AI capabilities [96, 22, 23, 101, 106]. Addressing scalability and continuous learning challenges develops powerful, efficient, resilient, and adaptable AI systems for complex environments.

8.5 Future Research Opportunities and Innovations

AI interpretability is poised for significant advancements, with numerous research opportunities and innovations. Developing universal decomposer models for various tasks, leveraging reinforcement learning for enhanced decomposition, could lead to efficient, interpretative task-specific models [43].

In healthcare, refining models like EmLLM to enhance diagnosticity and sensitivity while addressing privacy and ethical concerns is a priority [19]. This involves integrating physiological data into AI systems for improved diagnostics and ethical data handling.

The educational sector presents opportunities for developing ethical AI guidelines, exploring assessment methods, and understanding AI's evolving teaching role [5]. This involves creating frameworks aligning AI with educational objectives, ensuring AI supports learning processes.

Creative domains will focus on sophisticated evaluation metrics and diverse input integration for enhanced creativity and innovation [26]. This could lead to AI systems generating more original, impactful outputs.

In dentistry and specialized fields, addressing data handling limitations and integrating human oversight in diagnostics ensures reliable AI applications [7]. This involves developing systems collaborating with human experts for enhanced decision-making and outcomes.

Refining methods like EEDP for complex graph management and exploring applications beyond graph data is promising [10]. This could expand graph-based models' utility, providing insights and solutions to complex problems.

Developing training-free methods to identify safety neurons and understanding their influence on model safety are critical research directions [24]. This could lead to robust AI systems prioritizing safety and ethics.

Future AI interpretability research opportunities underscore the potential for systems prioritizing transparency, adaptability, and ethical alignment. Focusing on human-centered approaches, researchers can create frameworks catering to stakeholders' diverse needs, enhancing understanding of complex models like large language models (LLMs) and deep neural networks (DNNs), addressing real-world challenges, improving fairness, reducing bias, and fostering trustworthy AI systems capable of informed decision-making in critical applications [21, 63, 84, 22].

9 Conclusion

The survey elucidates the transformative capabilities of Large Language Models (LLMs) across diverse sectors, emphasizing their potential to revolutionize systematic reviews by expediting completion times and enhancing result accuracy. Despite their promise, LLM deployment presents distinct challenges, necessitating the creation of robust evaluation metrics and effective mitigation strategies. Integrating diverse explanatory techniques is crucial for fostering AI acceptance, as comprehensive explanations are vital for widespread adoption.

LLMs' self-learning capabilities, as demonstrated by models employing frameworks like SECToR, highlight their potential for autonomous proficiency enhancement. Moreover, advancements in training methodologies, such as the S2D method, illustrate significant efficiency gains while maintaining accuracy, showcasing progress in optimizing LLM performance.

In the realm of neuroscience, emerging frameworks for brain decoding highlight the application of LLMs in brain-computer interfaces and cognitive modeling, bridging the gap between AI and cognitive sciences. However, ethical considerations remain a critical focus, particularly in applications like ChatGPT, where addressing current limitations is essential to ensure responsible usage.

The incorporation of Knowledge Graphs (KGs) presents a promising strategy for enhancing LLM output reliability, filling existing research gaps, and providing a structured framework to improve AI model performance. Ongoing research into neuron-level interpretability and model transparency is crucial for advancing AI trustworthiness, ensuring alignment with ethical standards and human values. As AI technologies continue to evolve, these efforts will be pivotal in developing systems that are not only powerful and efficient but also transparent and reliable.

References

[1] Venkat Srinivasan, Darshan Gandhi, Urmish Thakker, and Raghu Prabhakar. Training large language models efficiently with sparsity and dataflow, 2023.

[2] Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B. Chilton. Popblends: Strategies for conceptual blending with large language models, 2023.

[3] Jumbly Grindrod. Large language models and linguistic intentionality, 2024.

[4] Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, and Brian Ichter. Grounded decoding: Guiding text generation with grounded models for embodied agents, 2023.

[5] James Prather, Paul Denny, Juho Leinonen, Brett A. Becker, Ibrahim Albluwi, Michelle Craig, Hieke Keuning, Natalie Kiesler, Tobias Kohn, Andrew Luxton-Reilly, Stephen MacNeil, Andrew Peterson, Raymond Pettit, Brent N. Reeves, and Jaromir Savelka. The robots are here: Navigating the generative ai revolution in computing education, 2023.

[6] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[7] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and Bing Shi. Chatgpt for shaping the future of dentistry: The potential of multi-modal large language model, 2023.

[8] Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. From pre-training corpora to large language models: What factors influence llm performance in causal discovery tasks?, 2024.

[9] Jared Fernandez, Luca Wehrstedt, Leonid Shamis, Mostafa Elhoushi, Kalyan Saladi, Yonatan Bisk, Emma Strubell, and Jacob Kahn. Hardware scaling trends and diminishing returns in large-scale distributed training, 2024.

[10] Bin Hong, Jinze Wu, Jiayu Liu, Liang Ding, Jing Sha, Kai Zhang, Shijin Wang, and Zhenya Huang. End-to-end graph flattening method for large language models, 2024.

[11] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt-related research and perspective towards the future of large language models, 2023.

[12] Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. Fate-llm: A industrial grade federated learning framework for large language models, 2023.

[13] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[14] Alejandro Tlaie. Exploring and steering the moral compass of large language models, 2024.

[15] Ziya Zhou, Yuhang Wu, Zhiyue Wu, Xinyue Zhang, Ruibin Yuan, Yinghao Ma, Lu Wang, Emmanouil Benetos, Wei Xue, and Yike Guo. Can llms "reason" in music? an evaluation of llms' capability of music understanding and generation, 2024.

[16] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. Promptmaker: Prompt-based prototyping with large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–8, 2022.

[17] Priti Oli, Rabin Banjade, Jeevan Chapagain, and Vasile Rus. The behavior of large language models when prompted to generate code explanations, 2023.

[18] Subhankar Maity and Aniket Deroy. Human-centric explainable ai in education, 2024.

[19] Poorvesh Dongre, Majid Behravan, Kunal Gupta, Mark Billinghurst, and Denis Gračanin. Integrating physiological data with large language models for empathic human-ai interaction, 2024.

[20] Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. Language model alignment in multilingual trolley problems, 2024.

[21] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[22] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[23] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review, 2024.

[24] Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Finding safety neurons in large language models, 2024.

[25] Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Exploring reasoning biases in large language models through syllogism: Insights from the neubaroco dataset, 2024.

[26] Jeremy Straub and Zach Johnson. Initial development and evaluation of the creative artificial intelligence through recurring developments and determinations (cairdd) system, 2024.

[27] 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.

[28] Soniya Vijayakumar. Interpretability in activation space analysis of transformers: A focused survey. *arXiv preprint arXiv:2302.09304*, 2023.

[29] Tuomas Oikarinen and Tsui-Wei Weng. Linear explanations for individual neurons. *arXiv preprint arXiv:2405.06855*, 2024.

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

[31] Kohitij Kar, Simon Kornblith, and Evelina Fedorenko. Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nature Machine Intelligence*, 4(12):1065–1067, 2022.

[32] Geoff Keeling and Winnie Street. On the attribution of confidence to large language models, 2024.

[33] Matteo Paltenghi, Rahul Pandita, Austin Z. Henley, and Albert Ziegler. Follow-up attention: An empirical study of developer and neural model code exploration, 2024.

[34] Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. Can large language models provide security privacy advice? measuring the ability of llms to refute misconceptions, 2023.

[35] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective, 2024.

[36] Chen Ling, Xujiang Zhao, Xuchao Zhang, Yanchi Liu, Wei Cheng, Haoyu Wang, Zhengzhang Chen, Takao Osaki, Katsushi Matsuda, Haifeng Chen, and Liang Zhao. Improving open information extraction with large language models: A study on demonstration uncertainty, 2023.

[37] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction, 2024.

[38] Qiheng Mao, Zhenhao Li, Xing Hu, Kui Liu, Xin Xia, and Jianling Sun. Towards explainable vulnerability detection with large language models, 2025.

[39] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690, 2024.

[40] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024.

[41] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey, 2024.

[42] Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. Map-neo: Highly capable and transparent bilingual large language model series, 2024.

[43] Zhuofeng Wu, He Bai, Aonan Zhang, Jiatao Gu, VG Vinod Vydiswaran, Navdeep Jaitly, and Yizhe Zhang. Divide-or-conquer? which part should you distill your llm?, 2024.

[44] Soyeon Caren Han, Feiqi Cao, Josiah Poon, and Roberto Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond, 2024.

[45] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[46] Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. Comparative study of domain driven terms extraction using large language models, 2024.

[47] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain, 2024.

[48] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization, 2024.

[49] Yihe Fan, Yuxin Cao, Ziyu Zhao, Ziyao Liu, and Shaofeng Li. Unbridled icarus: A survey of the potential perils of image inputs in multimodal large language model security, 2024.

[50] Hongcheng Guo, Jian Yang, Jiaheng Liu, Liqun Yang, Linzheng Chai, Jiaqi Bai, Junran Peng, Xiaorong Hu, Chao Chen, Dongfeng Zhang, Xu Shi, Tieqiao Zheng, Liangfan Zheng, Bo Zhang, Ke Xu, and Zhoujun Li. Owl: A large language model for it operations, 2024.

[51] Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*, 2024.

[52] Bin Han, Cleo Yau, Su Lei, and Jonathan Gratch. Knowledge-based emotion recognition using large language models, 2024.

[53] Antonios Alexos, Yu-Dai Tsai, Ian Domingo, Maryam Pishgar, and Pierre Baldi. Neural erosion: Emulating controlled neurodegeneration and aging in ai systems, 2024.

[54] Bean: Interpretable and efficien.

[55] Kongyang Chen, Zixin Wang, Bing Mi, Waixi Liu, Shaowei Wang, Xiaojun Ren, and Jiaxing Shen. Machine unlearning in large language models, 2024.

[56] Bumjin Park and Jaesik Choi. Memorizing documents with guidance in large language models, 2024.

[57] Clement Neo, Shay B. Cohen, and Fazl Barez. Interpreting context look-ups in transformers: Investigating attention-mlp interactions, 2024.

[58] Zichen Song, Yuxin Wu, Sitan Huang, and Zhongfeng Kang. Avss: Layer importance evaluation in large language models via activation variance-sparsity analysis, 2024.

[59] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning, 2022.

[60] Hala Hawashin and Mehrnoosh Sadrzadeh. Multimodal structure-aware quantum data processing, 2025.

[61] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025.

[62] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M. Drucker. How do analysts understand and verify ai-assisted data analyses?, 2024.

[63] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 ieee conference on secure and trustworthy machine learning (satml)*, pages 464–483. IEEE, 2023.

[64] Divyansh Srivastava, Tuomas Oikarinen, and Tsui-Wei Weng. Corrupting neuron explanations of deep visual features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1877–1886, 2023.

[65] Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *The Thirteenth International Conference on Learning Representations*.

[66] Zeping Yu and Sophia Ananiadou. Neuron-level knowledge attribution in large language models. *arXiv preprint arXiv:2312.12141*, 2023.

[67] Aniket Kumar Singh, Suman Devkota, Bishal Lamichhane, Uttam Dhakal, and Chandra Dhakal. The confidence-competence gap in large language models: A cognitive study, 2023.

[68] Chenxu Zhao, Wei Qian, Yucheng Shi, Mengdi Huai, and Ninghao Liu. Automated natural language explanation of deep visual neurons with large models. *arXiv preprint arXiv:2310.10708*, 2023.

[69] Arushi Sharma, Zefu Hu, Christopher Quinn, and Ali Jannesari. Interpreting pretrained source-code models using neuron redundancy analyses. *arXiv preprint arXiv:2305.00875*, 2023.

[70] N ewron: A new generalization of.

[71] Chang Yue and Niraj K Jha. Learning interpretable differentiable logic networks. *IEEE Transactions on Circuits and Systems for Artificial Intelligence*, 2024.

[72] Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. Attention lens: A tool for mechanistically interpreting the attention head information retrieval mechanism, 2023.

[73] Tung-Yu Wu, Yu-Xiang Lin, and Tsui-Wei Weng. And: audio network dissection for interpreting deep acoustic models. *arXiv preprint arXiv:2406.16990*, 2024.

[74] 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.

[75] Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. "im not racist but...": Discovering bias in the internal knowledge of large language models, 2023.

[76] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future, 2024.

[77] Minhyeok Lee. Emergence of self-identity in ai: A mathematical framework and empirical study with generative large language models, 2024.

[78] Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. Ravel: Evaluating interpretability methods on disentangling language model representations. *arXiv preprint arXiv:2402.17700*, 2024.

[79] Gaurav Singh and Kavitesh Kumar Bali. Enhancing decision-making in optimization through llm-assisted inference: A neural networks perspective, 2024.

[80] Roma Shusterman, Allison C. Waters, Shannon O'Neill, Phan Luu, and Don M. Tucker. An active inference strategy for prompting reliable responses from large language models in medical practice, 2024.

[81] Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes, 2024.

[82] Xiuying Chen, Tairan Wang, Qingqing Zhu, Taicheng Guo, Shen Gao, Zhiyong Lu, Xin Gao, and Xiangliang Zhang. Rethinking scientific summarization evaluation: Grounding explainable metrics on facet-aware benchmark, 2024.

[83] Juhyun Oh, Eunsu Kim, Inha Cha, and Alice Oh. The generative ai paradox on evaluation: What it can solve, it may not evaluate, 2024.

[84] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[85] Mohammad Amaz Uddin, Muhammad Nazrul Islam, Leandros Maglaras, Helge Janicke, and Iqbal H. Sarker. Explainabledetector: Exploring transformer-based language modeling approach for sms spam detection with explainability analysis, 2024.

[86] Taha Aksu, Chenghao Liu, Amrita Saha, Sarah Tan, Caiming Xiong, and Doyen Sahoo. Xforecast: Evaluating natural language explanations for time series forecasting, 2024.

[87] Waris Gill, Ali Anwar, and Muhammad Ali Gulzar. Tracefl: Interpretability-driven debugging in federated learning via neuron provenance. *arXiv preprint arXiv:2312.13632*, 2023.

[88] Jumbly Grindrod. Modelling language, 2024.

[89] Hugh Zhang and David C. Parkes. Chain-of-thought reasoning is a policy improvement operator, 2023.

[90] Qingxiu Dong, Li Dong, Ke Xu, Guangyan Zhou, Yaru Hao, Zhifang Sui, and Furu Wei. Large language model for science: A study on p vs. np, 2023.

[91] Sean Xie, Soroush Vosoughi, and Saeed Hassanpour. Proto-lm: A prototypical network-based framework for built-in interpretability in large language models, 2023.

[92] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. Luna: A model-based universal analysis framework for large language models, 2024.

[93] Lingyu Li and Chunbo Li. Enabling self-identification in intelligent agent: insights from computational psychoanalysis, 2024.

[94] Xiongye Xiao, Heng Ping, Chenyu Zhou, Defu Cao, Yaxing Li, Yi-Zhuo Zhou, Shixuan Li, Nikos Kanakaris, and Paul Bogdan. Neuron-based multifractal analysis of neuron interaction dynamics in large models, 2025.

[95] Wuwei Zhang, Ziyu Lu, Trung Le, Hao Wang, Uygar Sümbül, Eric Todd SheaBrown, and Lu Mi. Netformer: An interpretable model for recovering dynamical connectivity in neuronal population dynamics. In *The Thirteenth International Conference on Learning Representations*.

[96] Sai Koneru, Jian Wu, and Sarah Rajtmajer. Can large language models discern evidence for scientific hypotheses? case studies in the social sciences, 2024.

[97] Thilini Wijesiriwardene, Amit Sheth, Valerie L. Shalin, and Amitava Das. Why do we need neuro-symbolic ai to model pragmatic analogies?, 2023.

[98] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024.

[99] Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. Can perplexity reflect large language model's ability in long text understanding?, 2024.

[100] Andreas C. Schneider, Valentin Neuhaus, David A. Ehrlich, Abdullah Makkeh, Alexander S. Ecker, Viola Priesemann, and Michael Wibral. What should a neuron aim for? designing local objective functions based on information theory, 2025.

[101] Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization, 2024.

[102] Ancheng Xu, Minghuan Tan, Lei Wang, Min Yang, and Ruifeng Xu. Numcot: Numerals and units of measurement in chain-of-thought reasoning using large language models, 2024.

[103] Pei-Fu Guo, Ying-Hsuan Chen, Yun-Da Tsai, and Shou-De Lin. Towards optimizing with large language models, 2024.

[104] Quentin Anthony, Benjamin Michalowicz, Jacob Hatef, Lang Xu, Mustafa Abduljabbar, Aamir Shafi, Hari Subramoni, and Dhabaleswar Panda. Demystifying the communication characteristics for distributed transformer models, 2024.

[105] Mihael Arcan, David-Paul Niland, and Fionn Delahunty. An assessment on comprehending mental health through large language models, 2024.

[106] Xinbei Ma, Tianjie Ju, Jiyang Qiu, Zhuosheng Zhang, Hai Zhao, Lifeng Liu, and Yulong Wang. On the robustness of editing large language models, 2024.

[107] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. A glimpse in chatgpt capabilities and its impact for ai research, 2023.

[108] Abi Aryan, Aakash Kumar Nain, Andrew McMahon, Lucas Augusto Meyer, and Harpreet Singh Sahota. The costly dilemma: Generalization, evaluation and cost-optimal deployment of large language models, 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.