A Survey on Actionable Interpretability in Large Language Models

Jie Cai[1] , **Mafizur Rahman**[2] , **James Enouen**[1] , **Lijun Qian**[2] , **Yan Liu**[1]
[1]University of Southern California, USA
[2]Prairie View A&M University, USA
{caijie,enouen,yanliu.cs}@usc.edu, {mrahman13,liqian}@pvamu.edu

Abstract

Large Language Models (LLMs) have become central to modern AI, with interpretability playing an increasingly important role in investigating the opaque and complex mechanisms encoded within billions of parameters and ensuring trustworthy deployment. However, descriptive interpretability approaches for LLMs remain largely post-hoc, explaining model behavior without actionable guidance for model improvement, thereby limiting their practical utility. Recent work has therefore shifted toward actionable LLM interpretability, emphasizing methods that connect internal mechanisms to interventions and model improvement rather than explanation alone. This survey reviews LLM interpretability through the lens of actionability, presenting a taxonomy of attributional and mechanistic approaches, along with emerging methods tailored to vision–language models (VLMs). We further examine how actionable interpretability supports downstream objectives such as hallucination mitigation, knowledge editing, fairness, safety, capability and efficiency. By positioning actionable interpretability as a pathway for better-guided LLM design and practice, this survey outlines key challenges and future directions toward trustworthy and controllable foundation models.

1 Introduction

Large Language Models (LLMs) have become a foundational paradigm in modern artificial intelligence, demonstrating remarkable performance across diverse domains and tasks. Their rapid progress has enabled increasingly capable systems that process language, images, and structured information within a unified modeling framework, leading to widespread deployment in real-world applications. However, the increasing capability and deployment of LLMs have amplified a fundamental challenge: the growing epistemic gap between observable model behavior and the underlying computational mechanisms that produce it. LLMs are large-scale, data-driven nonlinear systems whose internal representations and computations remain only partially understood. In particular, the relationship between their observable behaviors and the underlying mechanisms that encode information, perform reasoning, and generate outputs remains opaque. Such opacity is not merely a theoretical concern—it limits our ability to verify correctness, diagnose failure modes, or enforce safety constraints, thereby presents a central challenge for trustworthy deployment. These limitations are particularly consequential in high-stakes settings, where reliability, accountability, fairness, and safety are essential requirements [Doshi-Velez and Kim, 2017].

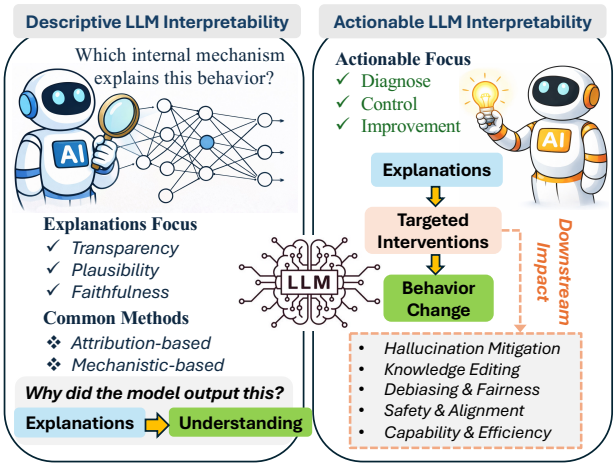

Figure 1: Descriptive interpretability focuses on explaining model behavior, whereas actionable interpretability seeks to translate insights into interventions that improve downstream model behavior.

Before the advent of LLMs, interpretability research primarily focused on traditional smaller-scale machine learning models, with the goal of understanding how input features influence model predictions. Representative approaches include post-hoc explanation methods such as LIME and Shapley-value-based methods, as well as inherently interpretable models such as linear models and decision trees [Dwivedi *et al.*, 2023]. Although interpretability continues to serve both as a scientific lens for understanding model mechanisms and as a tool for explanation-oriented applications, the rise of LLMs has substantially expanded the scope, complexity, and practical significance of interpretability.

Early LLM interpretability research primarily focused on descriptive explanations of model behavior and improve

Application Objectives	Interpretability Targets	Interpretability Methods	Actionable Interventions
① Mitigating Hallucinations	① Tokens	① Attribution-based Methods	① Steering
② Knowledge Editing	② Representations	② Attention-based Methods	② Editing
③ Fairness and Bias Mitigation	③ Layers	③ Sparse Autoencoder	③ Suppression
④ Improving Safety	④ Attention Heads	④ Probing	④ Enhancement
⑤ Enhancing Capabilities	⑤ Neurons	⑤ Logit Lens	⑤ Pruning
⑥ Improving Efficiency	⑥ Circuits	⑥ Component Analysis	

Table 1: A unified taxonomy of actionable LLM interpretability. The framework can be viewed through four dimensions: application objectives, interpretability targets, interpretability methods, and actionable interventions. The categories shown are representative rather than exhaustive, highlighting the most commonly studied directions in current research.

transparency. However, descriptive explanation alone provides limited support for modifying model behavior. Practical objectives such as improving reliability, editing knowledge, and enforcing safety constraints require the ability to identify and intervene on the mechanisms that give rise to model behavior. In this context, interpretability must move beyond describing model behavior toward enabling principled intervention. Explanations should clarify not only how models behave, but also which mechanisms can be monitored, modified, or controlled. Without this actionable role, interpretability risks remaining passive observation with limited practical utility. [Miller, 2019] emphasizes the utility of interpretability for specific stakeholders and tasks, referring to methods that provide explanations with tangible benefits-for example, supporting human decision-making, building trust, or facilitating model improvement. Building on the utility of interpretability, recent work [Zhang *et al.*, 2026] emphasizes ***"actionable"*** as moving beyond insight toward enabling direct use, informing concrete steps for further model improvement, behavioral control or downstream applications of explanations. Under this view, the value of interpretability lies not only in what it explains, but also in the actions it enables.

This survey provides a systematic review of LLM interpretability methods through the lens of actionability. We organize existing methods into attributional and mechanistic paradigms, while also discussing emerging actionable interpretability techniques for VLMs. We further examine how actionable interpretability supports downstream objectives, including hallucination mitigation, knowledge editing, fairness, safety, capability enhancement and efficiency inference. Figure 1 contrasts prior descriptive interpretability, which primarily explains model behavior, with actionable interpretability, which seeks to identify intervention targets and enable behavior change. Our objective is not merely to catalog existing techniques, but to develop a conceptual framework for actionable LLM interpretability: one that moves beyond descriptive explanations toward understanding and control.

2 Definitions and Taxonomy

The rise of LLMs has shifted interpretability research from explaining small, task-specific models to analyzing open-ended generative models trained on broad and heterogeneous corpora. LLMs encode implicit knowledge, implement context-dependent computations and exhibit complex ability, posing challenges beyond the assumptions of traditional interpretability. Prior work has shown that interpretability is inherently context-dependent, with the validity and utility of explanations varying across stakeholders, tasks, and operational settings [Miller, 2019]. In the LLM era, this dependency becomes even more pronounced, necessitating principled frameworks that situate interpretability claims within clearly defined user roles and objectives. Accordingly, this survey adopts an operational view of LLM interpretability as the study of how internal representations and computations give rise to model behavior. Under this view, interpretability aims to make the mechanisms underlying model generations understandable, analyzable, and ultimately actionable for diagnosis, intervention, control and improvement.

Actionable LLM Interpretability. In this survey, we define ***actionable*** LLM interpretability as interpretability that enables targeted interventions with verifiable effects on model behavior. Such interpretability goes beyond providing descriptive or post hoc accounts of model behavior; it must identifie functionally relevant representations, features, neurons, heads, layers, or circuits that can be manipulated to achieve a specified downstream objective. Although no universally accepted quantitative metric for actionability exists, a reasonable qualitative criterion is to ask: Who is the intended user of this explainability method and what concrete notion of utility is increased by using it? In practice, this entails both (i) accurately characterizing the aspects of the interpretability most relevant to the user's goals, and (ii) enabling targeted interventions that address identified issues, ranging from data- and input-level modulation to representation-level interventions and parameter-level editing. Importantly, the operational meaning of "actionable" is inherently role- and context-dependent: end-users may value clarity and transparency in the reasoning process, while researchers may prioritize mechanistic insights into internal computation that guide architectural design or training process.

Taxonomic Dimensions of Actionable LLM Interpretability. LLM interpretability methods can be systematically categorized along several orthogonal dimensions that together structure the landscape of LLM interpretability.

1. **Scope:** This dimension distinguishes between local and global interpretability. Local interpretability explains model behavior for a particular generation, while global interpretability provides an understanding of general model behavior across diverse inputs.
2. **Stage:** Intrinsic interpretability incorporates transparency into the model design, allowing the model's in-

ternal generation process to be examined directly. Post-hoc interpretability, in contrast, explains the behavior of already trained black-box models using auxiliary explanatory methods.

3. **Approach Paradigm:** Interpretability methods can be broadly categorized into two major paradigms. Attribution-based methods estimate the contribution of input features to model outputs. Mechanistic interpretability seeks to reverse-engineer the computations underlying model behavior by identifying functionally relevant representations and components.
4. **Task-Specific Functionality.** Selecting suitable interpretability approaches requires considering the specific task context because different LLM capabilities engage different computational processes. For instance, reasoning and memory retrieval involve different computational processes and thus require tailored interpretability techniques [Jin *et al.*, 2025a].
5. **Application Objectives.** Interpretability becomes practically valuable when it supports concrete interventions in model behaviorr. Its value lies not merely in explanatory plausibility, but in enabling downstream actions such as hallucination mitigation, knowledge editing, debiasing, and safety alignment.

Based on the taxonomy discussed above, we summarize the landscape of actionable LLM interpretability in Table 1. The table presents a unified taxonomy that organizes existing research along four complementary dimensions: application objectives, interpretability targets, interpretability methods, and actionable interventions. We include key related-works for Transformer interpretability because many remain applicable to LLMs when appropriately adapted. Nonetheless, extending these methods to LLMs requires careful consideration of the fundamental differences between large-scale language models and earlier, smaller Transformer models.

3 Explainability Paradigms for LLMs

3.1 Feature Attribution Interpretability for LLMs

Feature attribution methods seek to quantify the contribution of individual input tokens to an LLM's output. Given an input token sequence $x = \{x_1, x_2, \ldots, x_n\}$, and an LLM, an attribution method assigns each token x_i an importance score $S(x_i)$ that estimates its contribution to a target output. The attribution map provides a local explanation of which parts of the input strongly influence the model's output for this instance x. Representative families include perturbation-based and local-surrogate methods, attention-based methods, and gradient- and backpropagation-based methods.

Perturbation and Local Surrogate. Perturbation-based methods estimate feature importance by measuring how model outputs change when parts of the input are masked, removed, or otherwise modified. SHAP [Lundberg and Lee, 2017] and LIME [Ribeiro *et al.*, 2016] are representative examples: the former estimates Shapley-value contributions, while the latter fits an interpretable local surrogate to approximate the model's behavior using simpler, interpretable functions. Recent work extends these ideas to LLMs by perturbing internal components like attention weights [Deiseroth *et al.*, 2023], scaling attribution to long-context generation through efficient Shapley-value approximations [Enouen *et al.*, 2024; Liu *et al.*, 2025], and modeling higher-order feature interactions using sparse representations [Kang *et al.*, 2025a]. [Sudhi *et al.*, 2024] extends perturbation-based explanation to retrieval-augmented generation (RAG) by modifying retrieved contexts and measuring the resulting changes in model outputs, enabling attribution of model responses to specific retrieved evidence. A key advantage of perturbation-based approaches is their model-agnostic nature and direct causal interpretation: a feature is deemed important if modifying it changes the prediction. However, perturbation and surrogate-based methods often incur substantial computational overhead, rely on simplifying assumptions about feature interactions, and provide only indirect approximations of model behavior. Moreover, attribution estimates can be sensitive to the choice of perturbation scheme, while modern models may remain highly confident under heavily perturbed or semantically degraded inputs, raising concerns about the reliability of perturbation-based explanations [Feng *et al.*, 2018].

Attention-based. Attention-based methods leverage Transformers attention patterns as a source of token-level attribution. The most direct approach treats raw attention weights as importance scores. However, high attention weights do not necessarily correspond to high predictive importance, making raw attention an unreliable explanation signal. A variety of improved methods have been proposed - for example, Attention Rollout aggregates attention across layers, while Attention Flow models information propagation from output tokens back to input tokens using a maximum-flow formulation [Abnar and Zuidema, 2020]. [Chefer *et al.*, 2021a; Chefer *et al.*, 2021b] augment attention weights with gradient signals to derive class-specific relevance estimates and leverage positive gradient-attention interactions to improve faithfulness. Nevertheless, attention weights do not necessarily identify the causal factors driving model outputs [Wiegreffe and Pinter, 2019; Jain and Wallace, 2019]. Recent studies on attention sinks provide additional evidence for this limitation, showing that LLMs may assign disproportionately high attention to semantically uninformative tokens [Kang *et al.*, 2025b; Gu *et al.*, 2025].

Gradient-Based and backpropagation-based. Gradient-Based and backpropagation-based methods estimate feature importance by tracing how prediction-relevant signals propagate through the network. Early approaches such as Input × Gradient (I×G) [Simonyan *et al.*, 2013] estimate token importance by multiplying input representations with their gradients, but can be noisy and unstable due to gradient saturation or shattering effects in deep architectures. SmoothGrad [Smilkov *et al.*, 2017] stabilizes attributions by averaging gradients across perturbations, while Integrated Gradients (IG) [Sundararajan *et al.*, 2017] mitigates gradient saturation by integrating gradients along an interpolation path. To better accommodate language models, [Enguehard, 2023] propose Sequential Integrated Gradients (SIG), which attributes importance by interpolating one token at a time while hold-

ing all others fixed, avoiding unrealistic mixed-token interpolation paths. However, gradient integration methods remain computationally expensive for large Transformer models and provide limited insight into how intermediate representations contribute to the final prediction. A related line of work uses Layer-wise Relevance Propagation (LRP) [Bach *et al.*, 2015; Ding *et al.*, 2017] backpropagates relevance scores through layers. More recent variants such as AttnLRP [Achtibat *et al.*, 2024] enable relevance propagation across both input tokens and internal representations. Despite their strong theoretical grounding and fine-grained attributions, backpropagation-based methods remain constrained by computational cost and the difficulty of relating relevance scores to underlying computational mechanisms.

3.2 Mechanistic Interpretability for LLMs

Mechanistic interpretability aims to identify the internal computational mechanisms that explain model behavior [Rai *et al.*, 2024]. Its objects of study range from residual-stream directions and sparse features to neurons, attention heads, and multi-component circuits. In contrast to attribution-based methods, which quantify input-output relationships, mechanistic approaches aim to uncover how information is encoded, transformed, and routed through the network. Recent advances span several directions: (1) SAEs that extract interpretable feature bases, (2) probing methods that localize task-relevant information across internal representations, (3) logit-lens-based methods that track the emergence of output-relevant signals across layers, and (4) mechanism-discovery methods that identify causal computational structures at different levels of granularity.

Sparse Autoencoder. Sparse Autoencoders (SAEs) are widely used to decompose internal features into sparse representations, which can often be interpreted as human-understandable concepts [Dong *et al.*, 2025]. Given a vector x, for example, neuron activation, a SAE encodes it into a sparse latent representation and reconstructs it through a decoder, typically written as $\mathrm{SAE}(x) = \mathrm{Dec}(\mathrm{Act}(\mathrm{Enc}(x)))$, where Enc and Dec are learned linear maps and Act enforces sparsity. The model is trained to minimize reconstruction error while encouraging only a small number of latent features to be active for each input. After training, individual neurons are interpreted via inputs that strongly activate them. The corresponding decoder vectors serve as candidate directions in representation space, yielding a sparse feature (or concept) dictionary. SAEs have been used to identify features associated with semantic concepts, factual knowledge, and alignment-relevant behaviors, and have also been extended to multimodal representations [Huben *et al.*, 2024; Lou *et al.*, 2025; He *et al.*, 2025].

Probing. Probing is a representation analysis technique that treats an LLM as a frozen feature extractor and trains an auxiliary lightweight classifier, typically a linear probe, to determine whether a target property can be decoded from internal representations. If the probe achieves high predictive performance, the corresponding property is considered to be encoded in the representation. Probes can be applied to residual-stream states, attention-head outputs, feed-forward network (FFN) activations, and other intermediate representations. By evaluating decodability across layers and components, probing enables the localization of properties such as factual knowledge, linguistic information, truthfulness signals, and sentiment [Morris *et al.*, 2023; Yu *et al.*, 2025]. For example, [Jiang *et al.*, 2025c] show that first-token logit distributions contain signals predictive of confidence, truthfulness and safety, enabling probes to detect unanswerable questions, deceptive prompts, and jailbreak attempts before generation. While probing is simple, efficient, and widely applicable, it requires supervised datasets with property-specific labels, making probe construction a major limitation.

Logit Lens. The Logit Lens is a post hoc vocabulary-projection method that interprets intermediate representations by projecting them into the vocabulary space using the unembedding matrix [Belrose *et al.*, 2023]. Applied layer by layer, it tracks how output-relevant token probabilities change during forward computation, providing a direct view of the semantic content accessible from hidden states. Beyond standard layer-wise projections, recent extensions apply Logit Lens to neuron-level activations in FFN layers, revealing specialized neurons that encode domain-specific knowledge and contribute to model generalization [Huo *et al.*, 2024]. The Attention Lens [Jiang *et al.*, 2025c] visualizes how visual evidence is reflected in generation trajectories and attention patterns, finding that hallucinated tokens often show weaker visual support in critical layers. Overall, the Logit Lens provides a training-free and intuitive tool for tracing how output-relevant information becomes accessible across layers. Its diagnostic value lies in revealing the vocabulary-level content linearly readable from intermediate states. However, such projections remain inherently approximate, as vocabulary projections reflect what is linearly decodable through the final unembedding layer rather than what is causally utilized by the model during computation.

Component-Level Analysis at Different Granularities. Mechanistic analysis investigates model behavior at multiple levels of granularity, ranging from layers and attention heads to neurons and circuits. At the layer level, prior studies characterize how computational roles are distributed across model depth. For example, [Jin *et al.*, 2025b] introduce Concept Depth to characterize where different types of information become linearly decodable across model layers. Through layer-wise probing on factual, emotional, and reasoning tasks, they show that simpler concepts are often accessible from shallow representations, whereas more complex concepts typically require deeper layers. At the head level, analyses identify functionally specialized attention heads, such as truthfulness-related heads whose activations encode truthfulness-related signals [Li *et al.*, 2023], retrieval heads that access long-range contextual information and streaming heads that primarily attend to recent tokens and attention sinks [Xiao *et al.*, 2025]. Beyond layer-level and head-level explanations, interpretability research also investigate model behavior at the granularity of individual neurons and sparse circuits [Gurnee *et al.*, 2023]. A line of work seeks to identify small subsets of neurons, or causal computational pathways, that support specific capabilities, including fac-

tual knowledge storage [Dai *et al.*, 2022], language processing [Tang *et al.*, 2024] and multimodal reasoning [Yang *et al.*, 2026]. Beyond functional localization, recent studies suggest that even neurons activated by seemingly specialized tasks can participate in broader computational processes, revealing coupling between distinct capabilities and their influence on more general model behavior [Gurnee *et al.*, 2024; Chen *et al.*, 2025b].

4 VLMs Interpretability: Unique Challenges and Solutions

VLMs pose additional interpretability challenges because model behavior depends on how visual information is integrated with language representations throughout the model. This complicates the attribution of outputs to visual and textual evidence, as well as the analysis of multimodal information flow within the model. Moreover, VLMs are deployed in modality-rich tasks such as text-rich image perception, video understanding, and 3D spatial reasoning that require fundamentally different abilities, motivating task-specific analysis that can capture the unique computational processes.

4.1 Token-level Analysis

Token-level analysis examines how discrete visual and textual tokens encode information, interact across modalities, and influence model outputs. In particular, visual tokens are a central object of analysis because they provide the interface through which image content enters the language model and becomes available for grounding, reasoning, and generation.

Visual Token Attribution. Visual token attribution aims to characterize the contribution of individual visual tokens to model outputs and how their influence changes across layers. [Zhang *et al.*, 2025a] investigate the importance of visual tokens through an information-flow perspective. By combining attention scores with Grad-CAM, they find that image-text interactions are concentrated in shallow decoder layers, where visual information becomes progressively concentrated among a small subset of tokens, revealing substantial redundancy among visual tokens in these layers. [Li *et al.*, 2025b] identify contextual interference as a limitation of token-level explanations in VLMs, where visual activations induced by preceding tokens can obscure the evidence associated with later tokens. They propose an estimated-causal-inference-based method that separates token-specific visual cues from contextual effects.

Information Inside Visual Tokens. Another line of work examines what information visual tokens encode. [Fan *et al.*, 2026] show that visual tokens in VLMs exhibit pronounced semantic sparsity: visual tokens naturally separate into sink, dead, and alive groups, with only alive tokens carrying image-specific semantics. They further find that grounded object semantics are concentrated in a very small subset of visual tokens. [Papadimitriou *et al.*, 2025] use SAEs to decompose VLM embedding spaces into sparse concepts and study which concepts are predominantly visual, which are predominantly textual, and which support cross-modal semantic alignment. While many concepts are activated predominantly by either images or text, the authors show that numerous concept directions are nearly orthogonal to modality-separating subspaces and instead encode shared semantic structure.

4.2 Inner Workings of VLMs

Recent works investigate how visual information is transformed inside VLMs. Using logit-lens analysis and attention knockout, [Neo *et al.*, 2025; Zhang *et al.*, 2025b] suggest a stage-wise processing hierarchy: lower layers capture scene-level information, middle and late layers integrate question-relevant visual evidence, and later layers support answer formulation and syntactic refinement. Similarly, [Huang *et al.*, 2025b; Kaduri *et al.*, 2025] examine how textual instructions modulate visual information processing in VLMs. Their analyses suggest a staged process in which visual information is first pre-processed and filtered according to the instruction, then transferred into language representations and retrieved during generation. [Li *et al.*, 2026] uses logit-lens to trace the layer-wise evolution of visual-encoder representations and reveal that early layers primarily encode local attributes such as color and texture, while deeper layers progressively compose these low-level cues into object-level semantic representations. [Wu *et al.*, 2025] provide evidence for the semantic hub hypothesis, showing that semantically equivalent inputs across languages, modalities, and data types are mapped to a shared representational space. This shared space appears to be scaffolded by the model's dominant pretraining language, helping explain both cross-domain generalization and why logit-lens analyses can recover interpretable semantic structure from intermediate representations.

5 From Explanation to Intervention: Actionable Explainability for LLMs and VLMs

Actionable explainability treats explanations not as descriptive outputs, but as tools for guiding downstream interventions on model behavior. In this view, an explanation is actionable when it identifies a concrete targe whose manipulation can produce a measurable change in model behavior. Recent work [Orgad *et al.*, 2026] argues that actionability should be judged by two coupled properties: concreteness, meaning that the explanation identifies a specific action, and validation, meaning that the action is empirically shown to improve an objective. In practice, actionable LLM interpretability typically follows a three-stage workflow: localizing a relevant mechanism, verifying its causal or functional role, and leveraging the verified signal for model improvement. The following sections review representative applications of this paradigm in LLMs and VLMs.

5.1 Mitigating Hallucinations

Hallucinations refer to generated content that is unsupported by contextual evidence or inconsistent with facts [Huang *et al.*, 2025a]. Such evidence may comes from the prompt, retrieved context, or factual knowledge expected from the model. VLM hallucinations also involve failures of visual grounding, where generated content is insufficiently supported by image evidence. Actionable interpretability ad-

dresses hallucinations by identifying internal mechanisms associated with hallucinations and using these mechanisms to guide hallucination detection or mitigation.

Interpretability studies suggest that hallucinations are associated with internal mechanisms governing how models use evidence and factual knowledge. [Sun *et al.*, 2025] disentangle the roles of retrieved context and parametric knowledge, showing that hallucinations arise when copying heads fail to preserve retrieved evidence while knowledge FFNs overinject parametric knowledge into the residual stream. This imbalance provides both diagnostic signals and intervention targets for hallucination mitigation. A related line of work studies whether models internally recognize the entities they are asked about. Using SAEs, [Ferrando *et al.*, 2025] identify directions that distinguish known from unknown entities. These directions are causally linked to hallucination and refusal behavior. This suggests that hallucinations are partly associated with internal representations of entity familiarity. In VLMs, hallucinations arise when visual evidence is weakly incorporated into the model's internal generation process. Via logit lens, [Jiang *et al.*, 2025a] find that hallucinated tokens exhibit lower logit confidence, reflecting weaker internal association with visual evidence. At the attention level, [Jiang *et al.*, 2025c] show that hallucinated objects receive weaker or less consistent cross-head attention in intermediate layers than grounded objects At the representation level, [Phukan *et al.*, 2025] find lower similarity between hallucinated answers and image-region embeddings, indicating weaker alignment between hallucinated answers and visual evidence. These grounding signals also motivate targeted interventions. [Jiang *et al.*, 2025a] mitigate hallucinations by projecting the hallucinated image embeddings onto the subspace orthogonal to the corresponding hallucinated object's textual embeddings, thereby reducing unsupported object semantics in the visual representation. [Jiang *et al.*, 2025c] improve grounding by injecting an additional attention term computed from the average attention across all heads in intermediate layers, encouraging the model to prioritize image regions that consistently attract attention across heads.

5.2 Knowledge Editing

Knowledge editing encompasses several objectives, including knowledge updating and knowledge unlearning, which aim to modify targeted knowledge while preserving unrelated model behavior. Mechanistic interpretability provides a natural foundation for this objective by localizing the internal representations that encode and retrieve knowledge. Most knowledge editing methods therefore follow a locate-then-edit paradigm, where knowledge-bearing components are first localized and then selectively modified.

At the neuron level, [Dai *et al.*, 2022] introduce the concept of knowledge neurons, identifying a small set of FFN neurons whose activations are causally associated with the expression of specific factual knowledge. By manipulating these neurons and their corresponding FFN parameters, they demonstrate preliminary capabilities for knowledge editing without full model fine-tuning. [Li *et al.*, 2024] further refine mechanistic localization for knowledge editing by disentangling the roles of self-attention and feed-forward networks. Their analysis suggests that FFNs primarily store factual associations, whereas attention layers encode more general knowledge-retrieval patterns. Guided by this insight, they restrict edits to FFN weights, leading to more precise and reliable knowledge updates. [Guo *et al.*, 2025] show that effective knowledge editing depends on localizing the mechanisms responsible for factual lookup. By restricting edits to fact-lookup MLPs, they achieve more robust unlearning across prompt variations. Similar mechanistic editing paradigms have been extended to VLMs. [Basu *et al.*, 2024] introduce a multimodal causal tracing framework to identify the layers responsible for storing and transferring factual information in VLMs. Their analyses reveal that factual associations are concentrated in early causal MLP layers, and editing these layers enables effective correction of erroneous knowledge and insertion of new factual associations.

5.3 Fairness and Bias Mitigation

Interpretability-informed debiasing seeks to identify internal mechanisms responsible for biased behavior and intervene on them directly. Existing studies suggest that bias can be localized to identifiable neurons, latent representations, and sparse semantic features, enabling targeted interventions that mitigate biased behavior while preserving general capabilities. [Liu *et al.*, 2024] introduce social bias neurons, whose activations are causally associated with demographic disparities in model outputs. Using gradient-based attribution, they localize these neurons and mitigate biased behavior by suppressing their activations during inference. In contrast to prior neuron-suppression and neuron-editing approaches, [Pan *et al.*, 2026] identify know-bias neurons associated with bias awareness and selectively amplify their activations during inference, mitigating social bias through knowledge activation rather than behavior suppression. [Neplenbroek *et al.*, 2025] show that demographic biases can also emerge in latent user representations. Using linear probes, they reveal that stereotypical conversational cues induce demographic representations that may even override explicitly stated user identities. They further steer these representations toward the user-specified identity, mitigating stereotype-driven implicit personalization. [Yalavarthi *et al.*, 2026] leverage SAEs to decompose VLM embeddings into interpretable feature directions and identify those associated with spurious attributes. By removing the corresponding directions through orthogonal projection, they mitigate spurious correlations, demonstrating that SAE features can serve as actionable targets for fairness and robustness interventions.

5.4 Improving Safety

As LLMs are deployed in increasingly consequential settings, unsafe outputs raises substantial concerns, while the opacity internal mechanisms further complicates safety auditing and risk mitigation. Actionable interpretability addresses this challenge by identifying safety-relevant components and representations that can be monitored or intervened [Lee *et al.*, 2025]. A recurring finding across both LLMs and VLMs is that unsafe behavior can be traced to a small subset of internal mechanisms. At the component level, safety-critical

heads, neurons, and layers have been shown to play disproportionate roles in refusal, toxicity mitigation, and harmful-query detection, enabling selective editing, targeted tuning, or activation control [Zhou *et al.*, 2025; Zhao *et al.*, 2025; Li *et al.*, 2025a]. At the representation level, safety-relevant behavior such as refusal behavior can be modulated through low-dimensional latent representations [Arditi *et al.*, 2024], while toxicity-related SAE features [Goyal *et al.*, 2025] provide sparse targets for steering harmful generations. In VLMs, visual inputs introduce additional failure modes by allowing harmful signals to propagate through multimodal representations. Multimodal jailbreaks have been associated with distinctive hidden-state trajectories [Jiang *et al.*, 2025b], harmful activation directions that can be suppressed during inference [Wang *et al.*, 2025a], and sparse visual-textual tokens whose pruning and restoration mitigate unsafe generations [Chen *et al.*, 2025a]. Furthermore, [Xu *et al.*, 2025] show that text-side safety mechanisms do not reliably transfer to visual inputs, motivating cross-modal alignment as a mechanism-aware safety intervention.

5.5 Enhancing Capability

Interpretability can improve performance by turning explanatory findings into actionable intervention targets. Recent studies leverage mechanistic insights into attention, representations, neurons and heads to guide inference-time interventions and targeted fine-tuning. A representative example is post-hoc attention steering. [Zhang *et al.*, 2024] propose PASTA, which identifies influential a small set of functionally influential attention heads and reweights them toward user-emphasized spans at inference time, enabling fine-grained control without additional training. This interpretability-driven intervention translates head-level mechanistic insights into substantial improvements in instruction following and knowledge integration. [Chen *et al.*, 2025c] analyze attention trajectories over image tokens and show that successful spatial reasoning hinges on attention aligning with true object locations. Building on this diagnosis, they propose to reshape attention based on confidence score: sharpening attention when confidence is high and broadening it when low. Beyond attention, residual-stream representations can serve as actionable targets for intervention. [Højer *et al.*, 2025] identify latent directions associated with successful reasoning by contrasting internal activations from correct and incorrect reasoning trajectories. The resulting control vectors are injected into the residual stream at inference time, steering the model toward reasoning-favorable representational states and improving performance across multiple reasoning benchmarks. Actionable interpretability can inform fine-tuning. By identifying a sparse set of keystone neurons that are consistently important across multiple capability dimensions, [Ji *et al.*, 2026] demonstrate that updating only parameters associated with these neurons can achieve comparable or even better performance to full-parameter fine-tuning while better preserving all capabilities.

5.6 Improving Efficiency

Interpretability reveals how computational roles are distributed across model components, exposing functional specialization and computational redundancy at different levels such as tokens, heads and layers that can be leveraged for efficient inference. At the head level, [Xiao *et al.*, 2025] identify a functional division between retrieval heads, which access long-range contextual information, and streaming heads, which primarily attend to recent tokens and attention sinks. Exploiting this decomposition, they retain full KV caches for retrieval heads and use compressed caches for streaming heads, reducing memory and latency while preserving long-context performance. At the layer level, [Chen *et al.*, 2025d] show that consecutive layers contribute unequally to representation transformation, with some layer blocks inducing only minor changes in hidden representations. Thus, they replace low-impact layer blocks with lightweight modules, achieving compression with limited performance degradation. At the channel level, [Le *et al.*, 2025] use a small subset of informative hidden states to probe future representations and infer batch-specific channel importance. These activation-derived signals are then used to guide adaptive structured pruning during inference. Similar principles extend to VLMs. [Wang *et al.*, 2025b] reveal that fewer than 5% of attention heads are consistently responsible for visual concept understanding. Exploiting this head-level sparsity, they allocate KV-cache budgets asymmetrically across attention heads. Through visual information-flow analysis, [Yin *et al.*, 2025] show that shallow layers inject visual cues into instruction tokens, whereas deeper layers consolidate unimodal visual semantics through image–image interactions. Guided by this insight, they prune image tokens at selected layers to reduce computation while preserving performance.

6 Evaluation: Datasets and Benchmarks

Developing reliable benchmarks for actionable LLM interpretability remains challenging due to the lack of consensus on evaluation standards, especially as model capabilities continue to evolve and expand beyond purely text-based tasks. Nonetheless, recent efforts have begun to address this gap by moving toward evaluations grounded in intervention effectiveness and downstream utility.

Different from descriptive LLM interpretability benchmarks that primarily assess whether an explanation is faithful, stable, or human-understandable, actionable LLM interpretability introduces a stronger objective: after using the explanation, can the evaluator produce a verified improvement in model behavior without unacceptable collateral damage [Orgad *et al.*, 2026]? This shift necessitates new evaluation protocols that measure not only explanatory quality but also the effectiveness of the actions derived from those explanations. In this survey, we argue that actionable interpretability should be evaluated along five complementary dimensions. First, *localization validity* assesses whether a method identifies the feature, token, head, layer, or circuit that mediates the behavior. Second, *causal efficacy* measures whether intervening on the identified target changes the intended behavior. Third, *specificity* evaluates whether the intervention selectively changes the targeted behavior while preserving non-targeted model behavior. Fourth, *robustness and scalability* examine whether intervention effects persist

across prompts, distribution shifts, and model scales. Fifth, *utility* asks whether the explanation supports a better downstream action.

Existing benchmarks cover these criteria unevenly. InterpBench [Gupta *et al.*, 2024] constructs semi-synthetic Transformers with known ground-truth circuits using Strict Interchange Intervention Training (SIIT), enabling controlled evaluation of circuit discovery methods. By constraining non-circuit nodes from influencing model outputs, SIIT provides realistic testbeds while preserving circuit-level ground truth. MIB [Mueller *et al.*, 2025] benchmarks mechanistic interpretability methods along circuit localization and causal-variable localization. It evaluates whether methods recover concise causal pathways and whether internal features align with task-relevant causal variables. SAEBench [Adam *et al.*, 2025] evaluates SAEs beyond reconstruction quality, covering interpretability, feature disentanglement, concept detection, and practical applications such as unlearning, thereby providing a more comprehensive assessment of SAE utility. AxBench [Zhengxuan *et al.*, 2025] evaluates representation-based steering methods against prompting and finetuning baselines, finding that prompting and finetuning consistently outperform SAE-based interventions, while even simple steering vectors achieve stronger control than SAEs. [Guo *et al.*, 2025] use knowledge editing and unlearning as functional testbeds for evaluating localization quality. Their results show that localization methods should be assessed not only by whether the identified components influence model outputs, but also by whether interventions on those components produce robust changes with limited unintended effects.

Looking forward, progress in interpretability benchmarking will depend on the development of rigorous evaluation standards. Future benchmarks should clearly distinguish among explanatory objectives and intervention utility—and measure their relationships rather than treating interpretability as a single property. For VLMs, a key challenge is establishing whether explanations are grounded in the evidence actually used by the model.

7 Challenges and Opportunities

7.1 Challenges

Several challenges limit the effectiveness and broader adoption of actionable interpretability in modern foundation models. The first challenge is how to rigorously measure whether an LLM interpretability method is effectively actionable. There is no widely accepted standard for what constitutes a "good" explanation. Human evaluations suffer from subjectivity and low scalability, while automatic metrics—such as fidelity or stability—often fail to align with human understanding [Braun and Forell, 2025]. This lack of consensus undermines reproducibility and meaningful cross-method comparison. Developing reliable benchmarks remains challenging due to the absence of standardized evaluation criteria, particularly as model scale and modality complexity continue to increase. Existing benchmarks [Gupta *et al.*, 2024; Adam *et al.*, 2025] remain fragmented and task-specific, underscoring the need for unified evaluation frameworks that jointly assess faithfulness, robustness, and practical utility across textual and multimodal settings. The second challenge arises from the ambiguity and multifactorial nature of LLM outputs. Outputs are often shaped by entangled influences—including input phrasing, training distributions, prompt context, and internal hidden states. Attribution-based interpretability tends to conflate correlation with causation, making it hard to determine whether a model truly "used" a feature or merely "noticed" it. The localization of decisive mechanisms an open problem.

A third challenge concerns intervention reliability. Many interpretability-guided interventions show effectiveness under specific settings, yet their robustness across prompts, domains, modalities, or model scales remains poorly understood. Recent studies suggest that internal representations can shift substantially across contexts and languages through specialized latent-space transition mechanisms [Tezuka and Inoue, 2025], while activation-space interventions may transfer across model families through shared representation mappings [Oozeer *et al.*, 2025]. Nevertheless, the broader stability and generalizability of mechanistic interventions remain open questions. A related challenge is the scalability of actionable interventions. Many existing methods rely on computationally expensive procedures such as perturbation-based methods and SAEs. While effective in certain settings, these approaches may introduce significant cost overhead in LLMs and VLMs. Developing lightweight and scalable intervention mechanisms therefore remains an important direction for actionable interpretability research.

Finally, modern LLMs operate as highly non-linear complex systems. Their internal dynamics, driven by deep transformer stacks and attention mechanisms, render inference process opaque and resistant to mechanistic explanation. Unlike small-scale models, tracing causal chains within these architectures is prohibitively complex. Recent work [Sharkey *et al.*, 2025] on mechanistic interpretability further underscores the difficulty of mapping high-dimensional representations back to human-intelligible concepts. Even when technical explanations (e.g., saliency maps or token attributions) are available, end users like policymakers or domain experts often struggle to translate them into actionable insight. This gap between technical explanations and human understanding calls for more user-centric designs, such as contrastive explanations or interactive interpretability systems.

7.2 Opportunities

Despite these challenges, actionable interpretability presents significant opportunities for improving LLM deployment. It is central to building trust in high-stakes settings by enabling the detection of hallucinations, bias, and unsafe behaviors. It also serves as a powerful diagnostic tool and guides targeted interventions across the entire deployment pipeline including data curation, prompting strategies, and architecture adjustments. By generating intelligible and verifiable explanations, interpretability strengthens human–AI collaboration, facilitating oversight and reducing misuse. Moreover, actionable LLM interpretability have catalyzed methodological innovation, inspiring self-explaining models, LLM-as-explainer frameworks, and interactive explanation systems. Thus, future research may further enable scalable and real-time in-

terpretability frameworks that support continuous monitoring and adaptive intervention in LLMs.

Actionable interpretability also enables LLMs to surface mechanistic patterns in biology, chemistry, and medicine, as early efforts have shown that interpretable models can reveal meaningful structure, thereby transforming models from passive predictors into tools for hypothesis generation and controlled scientific reasoning. Moreover, emerging approaches [Kim *et al.*, 2025] treat LLMs as cooperative agents that provide multi-turn, concept-aware explanations to support human oversight. The concept of "agentic interpretability" advocates for LLMs to function not merely as providers of static explanations, but as interactive partners that adapt their explanatory behavior to support timely and effective human intervention.

8 Conclusion

This survey reviewed recent advances in LLM interpretability through the lens of actionability, organizing existing methods across attributional and mechanistic paradigms, extending this taxonomy to emerging approaches for VLMs, and examining how interpretable signals enable targeted interventions for downstream applications. Despite notable progress, the opacity and growing complexity of modern LLMs continue to limit trustworthy understanding and control. Future work should develop reliable evaluation frameworks and practical interventions that remain scalable across prompts, domains, modalities, and model scales, while enabling human–AI collaboration that is more transparent, controllable, and aligned with human needs. Advancing actionable interpretability will be essential for building models that are not only more effectively improved and controlled, but also more closely aligned with human needs.

Acknowledgments

This work was supported by the Army Research Office (ARO) under cooperative agreement W911NF-24-2-0133.

References

[Abnar and Zuidema, 2020] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020.

[Achtibat *et al.*, 2024] Reduan Achtibat *et al.* Attnlrp: Attention-aware layer-wise relevance propagation for transformers. In *ICML*, 2024.

[Adam *et al.*, 2025] Karvonen Adam *et al.* Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *ICML*, 2025.

[Arditi *et al.*, 2024] Andy Arditi *et al.* Refusal in language models is mediated by a single direction. In *NeurIPS*, 2024.

[Bach *et al.*, 2015] Sebastian Bach *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10:e0130140, 2015.

[Basu *et al.*, 2024] Samyadeep Basu *et al.* Understanding information storage and transfer in multi-modal large language models. In *NeurIPS*, 2024.

[Belrose *et al.*, 2023] Nora Belrose *et al.* Eliciting latent predictions from transformers with the tuned lens. *arXiv*, 2023.

[Braun and Forell, 2025] Bertil Braun and Martin Forell. (towards) scalable reliable automated evaluation with large language models. In *ACL GEM Workshop*, 2025.

[Chefer *et al.*, 2021a] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *ICCV*, 2021.

[Chefer *et al.*, 2021b] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021.

[Chen *et al.*, 2025a] Beitao Chen *et al.* Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism. In *NeurIPS*, 2025.

[Chen *et al.*, 2025b] Qian Chen *et al.* The tug of war within: Mitigating the fairness-privacy conflicts in large language models. In *ACL*, 2025.

[Chen *et al.*, 2025c] Shiqi Chen *et al.* Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. In *ICML*, 2025.

[Chen *et al.*, 2025d] Xiaodong Chen *et al.* Streamlining redundant layers to compress large language models. In *ICLR*, 2025.

[Dai *et al.*, 2022] Damai Dai *et al.* Knowledge neurons in pretrained transformers. In *ACL*, 2022.

[Deiseroth *et al.*, 2023] Björn Deiseroth *et al.* ATMAN: Understanding transformer predictions through memory efficient attention manipulation. In *NeurIPS*, 2023.

[Ding *et al.*, 2017] Yanzhuo Ding *et al.* Visualizing and understanding neural machine translation. In *ACL*, 2017.

[Dong *et al.*, 2025] Shu Dong *et al.* A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *EMNLP Findings*, 2025.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017.

[Dwivedi *et al.*, 2023] Rudresh Dwivedi *et al.* Explainable ai (xai): Core ideas, techniques, and solutions. In *ACM Comput. Surv.*, 2023.

[Enguehard, 2023] Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. In *Findings of ACL*, 2023.

[Enouen *et al.*, 2024] James Enouen *et al.* TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In *Findings of ACL*, 2024.

[Fan *et al.*, 2026] Yingqi Fan *et al.* What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. In *CVPR*, 2026.

[Feng *et al.*, 2018] Shi Feng *et al.* Pathologies of neural models make interpretations difficult. In *EMNLP*, 2018.

[Ferrando *et al.*, 2025] Javier Ferrando *et al.* Do i know this entity? knowledge awareness and hallucinations in language models. In *ICLR*, 2025.

[Goyal *et al.*, 2025] Agam Goyal *et al.* Breaking bad tokens: Detoxification of llms using sparse autoencoders. In *EMNLP*, 2025.

[Gu *et al.*, 2025] Xiangming Gu *et al.* When attention sink emerges in language models: An empirical view. In *ICLR*, 2025.

[Guo *et al.*, 2025] Phillip Guo *et al.* Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *ICML*, 2025.

[Gupta *et al.*, 2024] Rohan Gupta *et al.* Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. In *NeurIPS*, 2024.

[Gurnee *et al.*, 2023] Wes Gurnee *et al.* Finding neurons in a haystack: Case studies with sparse probing. *TMLR*, 2023.

[Gurnee *et al.*, 2024] Wes Gurnee *et al.* Universal neurons in gpt2 language models. *TMLR*, 2024.

[He *et al.*, 2025] Zirui He *et al.* Sae-ssv: Supervised steering in sparse representation spaces for reliable control of language models. In *EMNLP*, 2025.

[Huang *et al.*, 2025a] Lei Huang *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. In *ACM Transactions on Information Systems*, 2025.

[Huang *et al.*, 2025b] Zihan Huang *et al.* Traceable and explainable multimodal large language models: An information-theoretic view. In *COLM*, 2025.

[Huben *et al.*, 2024] Robert Huben *et al.* Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

[Huo *et al.*, 2024] Jiahao Huo *et al.* MMNeuron: Discovering neuron-level domain-specific interpretation in multimodal large language models. In *EMNLP*, 2024.

[Højer *et al.*, 2025] Bertram Højer, Oliver Jarvis, and Stefan Heinrich. Improving reasoning performance in large language models via representation engineering. In *ICLR*, 2025.

[Jain and Wallace, 2019] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL*, 2019.

[Ji *et al.*, 2026] Xiangtian Ji *et al.* Tiny brains, giant impact: Uncovering the keystone neurons of llm with just a few prompts. In *ICML*, 2026.

[Jiang *et al.*, 2025a] Nick Jiang *et al.* Interpreting and editing vision-language representations to mitigate hallucinations. In *ICLR*, 2025.

[Jiang *et al.*, 2025b] Yilei Jiang *et al.* Hiddendetect: Detecting jailbreak attacks against large vision-language models via monitoring hidden states. In *ACL*, 2025.

[Jiang *et al.*, 2025c] Zhangqi Jiang *et al.* Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*, 2025.

[Jin *et al.*, 2025a] Mingyu Jin *et al.* Disentangling memory and reasoning ability in large language models. In *ACL*, 2025.

[Jin *et al.*, 2025b] Mingyu Jin *et al.* Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *COLING*, 2025.

[Kaduri *et al.*, 2025] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *CVPR*, 2025.

[Kang *et al.*, 2025a] Justin Singh Kang *et al.* SPEX: Scaling feature interaction explanations for LLMs. In *ICML*, 2025.

[Kang *et al.*, 2025b] Seil Kang *et al.* See what you are told: Visual attention sink in large multimodal models. In *ICLR*, 2025.

[Kim *et al.*, 2025] Been Kim *et al.* Because we have llms, we can and should pursue agentic interpretability. *arXiv*, 2025.

[Le *et al.*, 2025] Qi Le *et al.* Probe pruning: Accelerating llms through dynamic pruning via model-probing. In *ICLR*, 2025.

[Lee *et al.*, 2025] Seongmin Lee *et al.* Interpretation meets safety: A survey on interpretation methods and tools for improving llm safety. In *EMNLP*, 2025.

[Li *et al.*, 2023] Kenneth Li *et al.* Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, 2023.

[Li *et al.*, 2024] Xiaopeng Li *et al.* Pmet: Precise model editing in a transformer. In *AAAI*, 2024.

[Li *et al.*, 2025a] Shen Li *et al.* Safety layers in aligned large language models: The key to llm security. In *ICLR*, 2025.

[Li *et al.*, 2025b] Yi Li *et al.* Token activation map to visually explain multimodal llms. *ICCV*, 2025.

[Li *et al.*, 2026] Yueyan Li *et al.* Reading images like texts: Sequential image understanding in vision-language models. In *ICLR*, 2026.

[Liu *et al.*, 2025] Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. Attribot: A bag of tricks for efficiently approximating leave-one-out context attribution. In *ICLR*, 2025.

[Liu *et al.*, 2024] Yan Liu *et al.* The devil is in the neurons: Interpreting and mitigating social biases in language models. In *ICLR*, 2024.

[Lou *et al.*, 2025] Hantao Lou *et al.* Sae-v: Interpreting multimodal models for enhanced alignment. In *ICML*, 2025.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

[Morris *et al.*, 2023] John X. Morris *et al.* Text embeddings reveal (almost) as much as text. *EMNLP*, 2023.

[Mueller *et al.*, 2025] Aaron Mueller *et al.* MIB: A mechanistic interpretability benchmark. In *ICML*, 2025.

[Neo *et al.*, 2025] Clement Neo *et al.* Towards interpreting visual information processing in vision-language models. In *ICLR*, 2025.

[Neplenbroek *et al.*, 2025] Vera Neplenbroek *et al.* Reading between the prompts: How stereotypes shape LLM's implicit personalization. In *EMNLP*, 2025.

[Oozeer *et al.*, 2025] Narmeen Oozeer *et al.* Activation space interventions can be transferred between large language models. *arXiv*, 2025.

[Orgad *et al.*, 2026] Hadas Orgad *et al.* Interpretability can be actionable. In *ICML*, 2026.

[Pan *et al.*, 2026] Jinhao Pan *et al.* Knowbias: Mitigating social bias in llms via know-bias neuron enhancement. In *ICML*, 2026.

[Papadimitriou *et al.*, 2025] Isabel Papadimitriou *et al.* Interpreting the linear structure of vision-language model embedding spaces. In *COLM*, 2025.

[Phukan *et al.*, 2025] Anirudh Phukan *et al.* Beyond logit lens: Contextual embeddings for robust hallucination detection and grounding in VLMs. In *NAACL*, 2025.

[Rai *et al.*, 2024] Daking Rai *et al.* A practical review of mechanistic interpretability for transformer-based language models. *arXiv*, 2024.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *KDD*, 2016.

[Sharkey *et al.*, 2025] Lee Sharkey *et al.* Open problems in mechanistic interpretability, 2025.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*, 2013.

[Smilkov *et al.*, 2017] Daniel Smilkov *et al.* Smoothgrad: Removing noise by adding noise. *arXiv*, 2017.

[Sudhi *et al.*, 2024] Viju Sudhi *et al.* Rag-ex: A generic framework for explaining retrieval augmented generation. In *SIGIR*, 2024.

[Sun *et al.*, 2025] ZhongXiang Sun *et al.* RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *ICLR*, 2025.

[Sundararajan *et al.*, 2017] Mukund Sundararajan *et al.* Axiomatic attribution for deep networks. In *ICML*, 2017.

[Tang *et al.*, 2024] Tianyi Tang *et al.* Language-specific neurons: The key to multilingual capabilities in large language models. In *ACL*, 2024.

[Tezuka and Inoue, 2025] Hinata Tezuka and Naoya Inoue. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms. In *EMNLP*, 2025.

[Wang *et al.*, 2025a] Han Wang *et al.* Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In *CVPR*, 2025.

[Wang *et al.*, 2025b] Jiahui Wang *et al.* Sparsemm: Head sparsity emerges from visual concept responses in mllms. In *ICCV*, 2025.

[Wiegreffe and Pinter, 2019] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP-IJCNLP*, 2019.

[Wu *et al.*, 2025] Zhaofeng Wu *et al.* The semantic hub hypothesis. In *ICLR*, 2025.

[Xiao *et al.*, 2025] Guangxuan Xiao *et al.* Duoattention: Efficient long-context llm inference with retrieval and streaming heads. In *ICLR*, 2025.

[Xu *et al.*, 2025] Shicheng Xu *et al.* Cross-modal safety mechanism transfer in large vision-language models. In *ICLR*, 2025.

[Yalavarthi *et al.*, 2026] Bharat Chandra Yalavarthi, Nalini K. Ratha, and Venu Govindaraju. Label-free mitigation of spurious correlations in VLMs using sparse autoencoders. In *ICLR*, 2026.

[Yang *et al.*, 2026] Jingcheng Yang *et al.* Circuit tracing in vision-language models: Understanding the internal mechanisms of multimodal thinking. In *CVPR Findings*, 2026.

[Yin *et al.*, 2025] Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. In *CVPR*, 2025.

[Yu *et al.*, 2025] Zhuoran Yu *et al.* How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *COLM*, 2025.

[Zhang *et al.*, 2024] Qingru Zhang *et al.* Tell your model where to attend: Post-hoc attention steering for llms. In *ICLR*, 2024.

[Zhang *et al.*, 2025a] Xiaofeng Zhang *et al.* From redundancy to relevance: Enhancing explainability in multimodal large language models. *NAACL*, 2025.

[Zhang *et al.*, 2025b] Zhi Zhang *et al.* Cross-modal information flow in multimodal large language models. In *CVPR*, 2025.

[Zhang *et al.*, 2026] Hengyuan Zhang *et al.* Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. *arXiv*, 2026.

[Zhao *et al.*, 2025] Yiran Zhao *et al.* Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *ICLR*, 2025.

[Zhengxuan *et al.*, 2025] Wu Zhengxuan *et al.* Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *ICML*, 2025.

[Zhou *et al.*, 2025] Zhenhong Zhou *et al.* On the role of attention heads in large language model safety. In *ICLR*, 2025.