Interpretation Meets Safety: A Survey on Interpretation Methods and Tools for Improving LLM Safety

Seongmin Lee, Aeree Cho, Grace C. Kim, ShengYun Peng, Mansi Phute, Duen Horng Chau
{seongmin,aeree,gracekim3,shengyun,mphute6,polo}@gatech.edu
Georgia Tech

Abstract

As large language models (LLMs) see wider real-world use, understanding and mitigating their unsafe behaviors is critical. Interpretation techniques can reveal causes of unsafe outputs and guide safety, but such connections with safety are often overlooked in prior surveys. We present the first survey that bridges this gap, introducing a unified framework that connects safety-focused interpretation methods, the safety enhancements they inform, and the tools that operationalize them. Our novel taxonomy, organized by LLM workflow stages, summarizes nearly 70 works at their intersections. We conclude with open challenges and future directions. This timely survey helps researchers and practitioners navigate key advancements for safer, more interpretable LLMs.

1 Introduction

Large language models (LLMs) have shown remarkable capabilities across many domains (Yu et al., 2022; Singhal et al., 2023; Sadybekov and Katritch, 2023; Wang et al., 2024e), but their outputs can be unsafe, posing significant challenges for real-world use (Tang and Li, 2025). In response, researchers have developed interpretation methods and tools to better understand the mechanisms behind unsafe generation and to improve safety (McGrath and Jonker, 2024; Ajwani et al., 2024).

Bridging Interpretation and Safety. As interest in both LLM interpretation and safety grows, a unifying survey that bridges the two becomes essential. Existing surveys largely focus on either interpretation (Zhao et al., 2024b; Ferrando et al., 2024; Zhao et al., 2024c; Calderon and Reichart, 2025) or safety (Huang et al., 2023b; Ayyamperumal and Ge, 2024; Shi et al., 2024b; Chua et al., 2024; Ma et al., 2025), without addressing how interpretation can enhance safety or inform users to operationalize such enhancements. Yet, both safety and human understanding are core motivations for interpretation research (Ferrando et al., 2024). Some works suggest safety as a downstream application of interpretation (Wu et al., 2024b) or explore only limited intersections between interpretation and safety (Bereska and Gavves, 2024). Moreover, emerging directions like self-reasoning interpretation, where LLMs explain their own behaviors, remain underexplored (Zhao et al., 2024b; Singh et al., 2024a; Calderon and Reichart, 2025). Our survey fills these critical gaps by contributing:

- **The first survey bridging LLM interpretation and safety** (Fig. 1). Our timely survey introduces a unified framework for summarizing safety-focused interpretation methods (§3), the safety enhancement strategies they inform (§4), and the practical tools that operationalize such enhancements (§5). Although improving safety and human understanding is often cited as a motivation for interpretation research (Doshi-Velez and Kim, 2017; Madsen et al., 2022; Zhao et al., 2024b,c; Ferrando et al., 2024), this connection has not been systematically surveyed until now.
- **Novel taxonomy of LLM interpretation methods organized by LLM workflow focus** (Fig. 1): training process (§3.1), input prompts (§3.2), inference (§3.3), and generation for self-reasoning (§3.4). These taxonomy categories anchor connections to six safety enhancement strategies (§4) and four tool types (§5), summarizing nearly 70 works at their intersections (Table 1),[1] including emerging areas like self-reasoning for interpretation not covered in prior surveys.
- **Distill open problems and challenges** to guide future NLP research and raise awareness of unresolved safety issues (§6). These include defending against interpretation-informed attacks, eval-

[1] Table 2 and Table 3 in the appendix extends Table 1 to include safety-oriented interpretation methods not yet leveraged for safety enhancements or tool use.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21515–21546
November 4-9, 2025

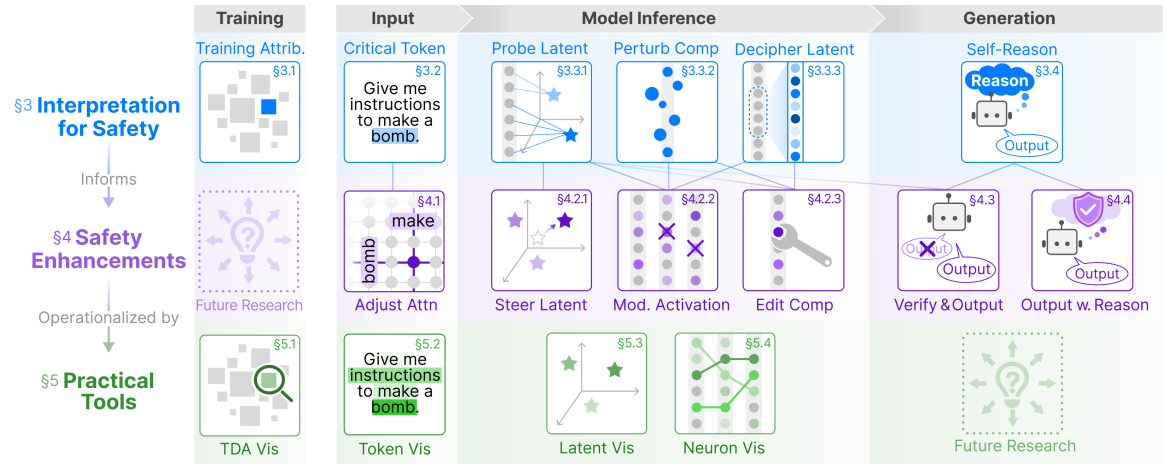

Figure 1: Visual overview our survey's unified framework bridging LLM interpretation and safety, summarizing the connections between safety-focused interpretation methods (§3), the safety enhancements they inform (§4), and the tools that operationalize them (§5). We organize surveyed research using a novel taxonomy based on LLM workflow: training, input tokens, inference, and generation.

uating interpretation, using training attribution for safety, designing user-centered presentation of interpretation, and refining safety dimensions.

2 Survey Methodology

We focus on interpretation methods addressing safety issues in autoregressive Transformer-based generative LLMs (Vaswani et al., 2017; Brown et al., 2020), among the most widely used and studied architectures (Huang et al., 2023c; Veeramachaneni, 2025). We adopt the established definition of interpretation as *recognizing and explaining LLM behaviors in human-understandable terms* (Doshi-Velez and Kim, 2017; Räuker et al., 2023; Hsieh et al., 2024; Singh et al., 2024a). We exclude methods requiring major architectural changes, as they hinder practical integration (Ludan et al., 2023; Tan et al., 2024; Sun et al., 2025).

We focus on four major safety concerns commonly studied in interpretation research (Qian et al., 2024):[2] *hallucination*, *jailbreaks and harmfulness*[3], *bias*, and *privacy leakage*. These safety risks often share similar underlying causes, such as risky training data and objectives that favor memorization over generalization. Such design choices can embed unintended correlations, reinforce biases, and leave models vulnerable to adversarial prompts. As a result, hallucinations arise from confident fabrication; jailbreaks exploit alignment flaws; bias reflects systemic data skew; and privacy leakage stems from memorized content. Despite differing in expression, these risks all stem from failures to generalize appropriately. Section 3 elaborates on how various mitigation approaches address these intertwined unsafe behaviors.

We curated nearly 70 works from top venues in machine learning, natural language processing, human-computer interaction, and visualization, with a focus on understanding, enhancing, and communicating LLM safety through interpretation (Table 1).

3 Interpretation Methods for LLM Safety

We categorize interpretation methods by where they operate in the LLM workflow (Fig. 1): training process (§3.1), input tokens (§3.2), model internals (§3.3), and underlying knowledge revealed by LLMs' self-explanatory capabilities (§3.4).

3.1 Attribute Safety to Training Process

Since LLMs are shaped by their training data (Grosse et al., 2023), training data attribution (TDA) assumes unsafe behaviors stem from problematic training data and evaluates the contribution of each training data point to model behavior. **Representation-based attribution** does this by comparing the similarity between the latent vectors of each training example and the output (Yeh et al., 2018; Tsai et al., 2023; Su et al., 2024b; He et al., 2024b). While effective for identifying related data, it does not establish causality (Cheng et al., 2025a).

To assess causal influence, **gradient-based methods** estimate how sensitive a model's parameters are to individual training examples. Many

[2]We focus on risks of direct harm or misuse, excluding general performance issues like out-of-distribution robustness.

[3]We do not consider user intent (malicious or benign) as interpretation methods reveal mechanisms behind unsafe outputs regardless of intent.

Table 1: Overview of representative works at the intersections of safety-focused interpretation (§3), safety enhancements they inform (§4), and tools operationalizing them (§5). Each row is one work; each column corresponds to a technique or tool. Safety issues, techniques, and tools addressed by a work are indicated by a colored cell.

	SAFETY TYPE				§3 INTERPRET FOR SAFETY						§4 ENHANCE SAFETY						§5 PRACTICAL TOOLS					VENUE
Work	Hallucination	Jailbreak & Harm	Bias	Privacy Leakage	3.1 Training Attrib.	3.2 Input Token	3.3.1 Probe Latent	3.3.2 Perturb Comp	3.3.3 Decipher Latent	3.4 Self-Reason	4.1 Attn. to Rel. Token	4.2.1 Steer Latent Vec	4.2.2 Modulate Neuron	4.2.3 Edit Model	4.3 Verify & Output	4.4 Output w. Reason	Ease Impl.	5.1 TDA Vis	5.2 Token Vis	5.3 Latent Vec Vis	5.4 Neuron Vis	Publication
Lee et al. (2025a)	●				●													●				AAAI
Hazra et al. (2024)		●			●									●								EMNLP
Zhao et al. (2024e)		●			●									●								ArXiv
Qian et al. (2024)	●	●	●	●	●		●					●										ACL
Sarti et al. (2023)			●			●											●		●			ACL
Mishra et al. (2025)			●			●													●			TVCG
Vig (2019)			●			●													●			ACL
Wang et al. (2025b)		●				●								●								ArXiv
Dale et al. (2023)	●					●									●							ACL
Chuang et al. (2024)	●					●									●							EMNLP
Pan et al. (2025a)		●				●	●				●											ArXiv
Li et al. (2023b)	●					●		●											●	●		ArXiv
Tenney et al. (2020)			●			●		●											●	●		EMNLP
Zhang et al. (2024b)			●			●		●			●		●									ICLR
Zhu et al. (2024)		●					●				●											ArXiv
Duan et al. (2024)	●						●					●										ArXiv
Ball et al. (2024)		●					●					●										ArXiv
Li et al. (2025c)		●					●					●										COLING
Wang et al. (2024a)		●					●					●										ArXiv
Yang et al. (2024a)	●						●					●										ACL
Bhattacharjee et al. (2024)		●					●					●										SafeGenAI
Chu et al. (2024)	●	●	●				●					●										CCS
Rimsky et al. (2024)	●	●					●					●										ACL
Singh et al. (2024b)		●	●				●					●										ICML
Zhang et al. (2024c)	●						●					●										ACL
Li et al. (2023a)	●						●					●										NeurIPS
Turner et al. (2023)		●					●					●										ArXiv
Gao et al. (2024)		●					●					●										ArXiv
Shen et al. (2024)		●					●					●										ICLR
Han et al. (2025)		●					●					●										ArXiv
Hernandez et al. (2024a)			●				●					●										COLM
Chen et al. (2024b)			●				●					●								●		ArXiv
Wang et al. (2024c)		●					●							●								ACL
Li et al. (2024e)		●					●							●								ArXiv
Chen et al. (2025)	●						●								●							AAAI
Zhao et al. (2024a)		●					●								●							ArXiv
Burns et al. (2022)	●						●								●							ArXiv
Li et al. (2024c)	●						●								●							ArXiv
Deng et al. (2025)	●							●					●									AAAI
Liu et al. (2024b)			●					●					●									ICLR
Li et al. (2024a)	●							●						●								ArXiv
Ma et al. (2023)			●					●						●								EMNLP
Li et al. (2025b)		●						●						●								ICLR
Zhao et al. (2024d)		●						●	●					●								EMNLP
Hernandez et al. (2024b)			●						●											●		ICLR
Lindsey et al. (2025)	●	●	●						●												●	Anthropic
Ameisen et al. (2025)									●												●	Anthropic
Zhou et al. (2025a)			●						●				●									ICLR
Frikha et al. (2025)				●					●				●									ArXiv
Hegde (2024)			●						●				●									SciForDL
He et al. (2025)	●	●	●						●				●									ArXiv
Abdaljalil et al. (2025)	●								●				●									ArXiv
Bayat et al. (2025)	●								●				●									ArXiv
Wu et al. (2025)		●							●				●									ArXiv
O'Brien et al. (2024)		●							●				●									ArXiv
Khoriaty et al. (2025)		●							●				●									ArXiv
Geva et al. (2022b)		●							●				●									EMNLP
Geva et al. (2022a)	●		●						●				●							●		EMNLP
Yu et al. (2024b)	●								●					●								EMNLP
Dhuliawala et al. (2024)	●									●					●							ACL
Weng et al. (2023)	●									●					●							EMNLP
Cheng et al. (2025b)	●									●					●							ArXiv
Liu et al. (2025a)		●								●						●						ArXiv
Jiang et al. (2025)	●									●						●						ICASSP
Ji et al. (2024a)	●									●						●						AAAI
Zhang et al. (2025)		●								●						●						ArXiv
Kaneko et al. (2024)			●							●						●						ArXiv
Prahallad and Mamidi (2024)			●							●						●						ArXiv
Li et al. (2024d)		●								●						●						ArXiv
Cao et al. (2024)		●								●						●						NaNA
Rad et al. (2025)		●								●						●						ArXiv
Moore et al. (2024)			●							●						●						ArXiv
Sicilia and Alikhani (2024)			●							●						●						NLP4PI
Mou et al. (2025)		●								●						●						ArXiv
Liu et al. (2025a)		●								●						●						ArXiv
Kwon and Mihindukulasooriya (2023)			●																	●		IUI

build on TracIn (Pruthi et al., 2020), which traces influence by measuring the alignment between gradients of losses computed for a model output and each training data. Variants have improved its accuracy (Han and Tsvetkov, 2021; Yeh et al., 2022; Wu et al., 2022; Han and Tsvetkov, 2022; Ladhak et al., 2023) and adapted it for LLMs (Xia et al., 2024; Pan et al., 2025b). However, these methods fall short in estimating the effect of removing a training point (Hammoudeh and Lowd, 2024; Cheng et al., 2025a). More theoretically grounded work builds on **influence function** (Hampel, 1974; Cook and Weisberg, 1980; Koh and Liang, 2017), which estimates how downweighting a training example affects model parameters and predictions (Koh and Liang, 2017). Despite scalability improvements (Han et al., 2020; Ren et al., 2020; Barshan et al., 2020; Guo et al., 2021; Schioppa et al., 2022; Park et al., 2023) and extension to LLMs (Grosse et al., 2023; Kwon et al., 2024; Choe et al., 2024; Wu et al., 2024a; Chang et al., 2025), their effectiveness is debated due to strong assumptions like model convexity (Basu et al., 2021; Akyurek et al., 2022; Li et al., 2024g), which rarely hold for LLMs (Cheng et al., 2025a).

Another direction explores **Data Shapley**, which

estimates the contribution of individual or groups of data points by approximating the effect of their removal or addition (Ghorbani and Zou, 2019; Jia et al., 2019; Feldman and Zhang, 2020). While promising, these methods are computationally expensive and have so far been limited to smaller models (Wang et al., 2024b, 2025a). Furthermore, the inaccessibility of LLMs' proprietary training data poses challenges for application of TDA overall (Bommasani et al., 2021; Achiam et al., 2023).

Beyond training data, researchers attempted to **understand LLM training dynamics** of learning new concepts and capabilities. Some studies use synthetic tasks with well-defined concepts, examining how models acquire them over training (Park et al., 2024a; Prakash et al., 2024a). Others compare models pre- and post-fine-tuning (Zhao et al., 2024e; Chen et al., 2024a; Hazra et al., 2024), simulate training (Ilyas et al., 2022; Guu et al., 2023; Engstrom et al., 2024), or analyze model internals (§3.3) across training checkpoints (Davies et al., 2023; Nanda et al., 2023a; Xu et al., 2024b; Prakash et al., 2024a; Ma et al., 2024; Inaba et al., 2025). These studies have revealed how safety capabilities like toxic content refusal emerge during training (Qian et al., 2024; Lee et al., 2024).

3.2 Identify Safety-Critical Input Tokens

A major branch of AI interpretation research attributes model outputs to specific input features, hypothesizing that unsafe outputs stem from over-attending or misinterpreting specific input tokens (Simonyan et al., 2014; Bach et al., 2015; Ribeiro et al., 2016; Selvaraju et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017; Lundberg and Lee, 2017). For Transformer-based models, early methods examined **attention weights**, based on the intuition that higher weights singal greater importance (Wiegreffe and Pinter, 2019; Abnar and Zuidema, 2020; Kobayashi et al., 2020). While attention weights offer insights into model behavior (Halawi et al., 2024; Yuksekgonul et al., 2024), monitoring all heads in LLMs can be overwhelming. Aggregation strategies, from simple heuristics (e.g., mean, max) (Tu et al., 2021; Sarti et al., 2024) to more principled attention rollout (Abnar and Zuidema, 2020), help reduce complexity.

To improve the faithfulness by accounting for other model components like residuals and layer norms (Kobayashi et al., 2021), **vector-based methods** decompose latent vectors into vectors attributable to input tokens (Kobayashi et al., 2021; Modarressi et al., 2022; Ferrando et al., 2022b, 2023; Modarressi et al., 2023; Yang et al., 2023a; Achtibat et al., 2024; Song et al., 2024; Kobayashi et al., 2024). These are applied to modern LLMs (Arras et al., 2025) and used to analyze hallucinations (Ferrando et al., 2022a; Dale et al., 2023; Chuang et al., 2024). However, these methods require model-specific designs, limiting adaptability (Abbasi et al., 2024).

Perturbation-based methods offer a model-agnostic approach by modifying input tokens and observing output changes (Ribeiro et al., 2016). Perturbations include altering latent vectors (Deiseroth et al., 2023; Madani et al., 2025), masking or zeroing token embeddings (Jacovi et al., 2021; Yin and Neubig, 2022; Mohebbi et al., 2023; Cohen-Wang et al., 2024), replacing tokens (Finlayson et al., 2021; Liu et al., 2023; Mohebbi et al., 2023; Sadr et al., 2025), or prompting counterfactuals (Bhattacharjee et al., 2023b; Yona et al., 2023; Gat et al., 2024). Some extend Shapley value, which has been largely explored for classical models (Lundberg and Lee, 2017; Covert et al., 2021), to estimate the influence of specific input tokens (Horovicz and Goldshmidt, 2024; Mohammadi, 2024; Enouen et al., 2024). These methods have identified tokens triggering prompt poisoning (Cohen-Wang et al., 2024) and biases (Mohammadi, 2024). However, they can be costly and create unnatural, out-of-distribution inputs, causing unfaithful interpretations (Abbasi et al., 2024; Achtibat et al., 2024).

To address these pitfalls, **gradient-based methods** compute the gradient of the model output with respect to input embeddings, quantifying how changes to each token affect output (Simonyan et al., 2014; Bach et al., 2015; Selvaraju et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017). Initially designed for smaller models (Enguehard, 2023; Achtibat et al., 2024; Wang et al., 2024f; Song et al., 2024), they have since been refined with contrastive explanations (Jacovi et al., 2021; Yin and Neubig, 2022; Eberle et al., 2023; Sarti et al., 2024) and extended to LLMs (Barkan et al., 2024a; Rezaei Jafari et al., 2024; Qi et al., 2024; Pan et al., 2025a).

Other approaches include **similarity-based** methods that compare the final output representation to input token embeddings, assuming higher similarity indicates greater token importance (Ferrando et al., 2022b; Abbasi et al., 2024). In par-

allel, researchers have proposed **prompt-based approaches** (Bhattacharjee et al., 2023a; Huang et al., 2023a), which instruct LLMs to identify influential tokens for behaviors like jailbreaks (Wang et al., 2025b), and **optimization-based techniques**, which search for token attributions that maximize certain interpretability metrics (Zhou and Shah, 2023; Barkan et al., 2024c,b).

3.3 Interpret Safety via Inference-Time Model Internals

Multiple works have shown that unsafe behaviors are associated with certain regions in latent space (Zou et al., 2023) or specific model components (Geva et al., 2023) or neurons (Geva et al., 2021), which can be interpreted in three ways: (1) identifying latent regions tied to safety (§3.3.1); (2) perturbing components to assess impact (§3.3.2); and (3) deciphering latent vectors with human-understandable terms (§3.3.3).

3.3.1 Probe Safety Regions in Latent Space

Recent work investigates whether and how safety-related concepts are encoded in LLMs' latent vectors (Zou et al., 2023). This work builds on the *linear representation hypothesis*, which posits that high-level concepts like factuality or harmfulness are embedded as linear directions in the model's latent space (Mikolov et al., 2013; Elhage et al., 2022; Park et al., 2024b). Under this view, researchers analyze latent vectors from individual layers—or concatenated across multiple layers—to identify which latent vectors encode safety behaviors.

A simple yet powerful approaches compute **mean latent vectors** for the data points with and without a particular concept (e.g., hallucinated vs. factual). These reveal directions associated with hallucinations (Liu et al., 2024a; Chen et al., 2025) and jailbreaks (Arditi et al., 2024; Zhao et al., 2024a; Zhu et al., 2024; Lin et al., 2024). Dimensionality reduction using PCA or SVD further uncover axes responsible for unsafe behaviors (Duan et al., 2024; Ball et al., 2024; Pan et al., 2025a).

Another widely used technique is **probing classifiers**, where a model is trained to predict whether a latent vector encodes a safety-related property (Alain and Bengio, 2016; Tenney et al., 2019; Dalvi et al., 2019; Kadavath et al., 2022; Gurnee et al., 2023; Liu et al., 2024a; Ju et al., 2024). Probing successfully detects hallucinations (Burns et al., 2022; Slobodkin et al., 2023; Orgad et al., 2024; Ashok and May, 2025), jailbreaks (Zhou et al., 2024; Xu et al., 2024c; Abdelnabi et al., 2024; Ashok and May, 2025), and bias (Orgad et al., 2024). However, these properties are not always linearly separable (Hildebrandt et al., 2025), introducing non-linear classifiers (Azaria and Mitchell, 2023; Ji et al., 2024b; Zhang et al., 2024a; Su et al., 2024a; Bürger et al., 2024; He et al., 2024c; Li et al., 2025a; Tan et al., 2025) or contrastive learning (He et al., 2024a; Beigi et al., 2024) to better capture complex boundaries of unsafe model behaviors. A known challenge of the probing methods is poor generalization across tasks and datasets (Belinkov, 2022; CH-Wang et al., 2024; Levinstein and Herrmann, 2024), which has been partially resolved by incorporating distributional differences into the loss function (Bürger et al., 2024) or training probing models on diverse datasets (Liu et al., 2024a).

3.3.2 Perturb to Assess Safety Impact

A common way to understand how specific components affect model behavior is to perturb them and observe changes. One approach uses **gradient-based analysis**, computing output gradients with respect to model parameters to evaluate each parameter's influence. While useful for explaining mechanisms behind knowledge conflicts in RAG (Jin et al., 2024) and biased generations (Liu et al., 2024b), such methods may not sufficiently capture causal relations (Chattopadhyay et al., 2019).

A more direct approach is **component knockout**, which ablates layers, attention heads, or parameters to identify their influence (Olsson et al., 2022; Geva et al., 2023). This has localized components responsible for hallucinations (Jin et al., 2024; Li et al., 2024a), jailbreaks (Zhao et al., 2024d; Wei et al., 2024), and biases (Yang et al., 2023b; Ma et al., 2023). Instead of full ablation, **parameter scaling** adjusts component influence (Luick, 2024), pinpointing safety-critical layers (Li et al., 2025b) and heads (Zhou et al., 2025b), while **parameter perturbation** modifies model weights and evaluates how safety properties respond to the structural changes (Peng et al., 2024; Huang et al., 2024; Leong et al., 2024), offering a broader perspective on the stability and robustness of safety alignment across the model's parameter landscape.

Activation patching, inspired by causal mediation analysis (Pearl, 2001; Vig et al., 2020), replaces intermediate activations (e.g., latent vectors, attention weights) from one input with those from another input, measuring how such intervention affects the model output. It localizes model

components associated with hallucinations (Monea et al., 2024; Deng et al., 2025) and biases (Vig et al., 2020), as well as general model capabilities (Geiger et al., 2021; Stolfo et al., 2023; Davies et al., 2023; Cabannes et al., 2024) and factual knowledge (Meng et al., 2022; Nanda et al., 2023b; Ghandeharioun et al., 2024).

To uncover how components interact, researchers extract *computational circuits*, graphs with important components as nodes and information flow as edges (Geiger et al., 2021; Elhage et al., 2021). **Path patching**, an extension of activation patching, modifies outputs along specific computational paths while freezing the rest of the network[4] (Wang et al., 2023; Goldowsky-Dill et al., 2023; Prakash et al., 2024b; Hanna et al., 2023). Due to the high reliance on human inspection, several efforts automate circuit discovery (Conmy et al., 2023; Ferrando and Voita, 2024; Bhaskar et al., 2024), while attribution patching approximates causal effect for scalability (Nanda, 2024; Syed et al., 2024; Kramár et al., 2024; Hanna et al., 2024). However, as LLM circuit analysis is still in its early stages, most focus on simple grammatical or arithmetic tasks, with very few addressing real-world safety problems (Hanna et al., 2024).

3.3.3 Decipher Latent Vectors with Language

One approach to understand latent vectors through language is to analyze how their individual neurons respond to input data. By identifying inputs that highly activate a neuron and their shared patterns, researchers have inferred concepts encoded by each neuron (Geva et al., 2021; Foote et al., 2023). However, many neurons in LLMs are *polysemantic*, encoding multiple unrelated concepts, making interpretation challenging (Arora et al., 2018; Bricken et al., 2023; Templeton et al., 2024).

To address this, researchers have developed techniques to disentangle concepts superposed in the latent vectors (Elhage et al., 2022). A prominent method is using Sparse dictionary learning (Mairal et al., 2008; Makhzani and Frey, 2013; Elhage et al., 2021) to train **sparse autoencoders (SAEs)** (Sharkey et al., 2022; Bricken et al., 2023; Huben et al., 2024; Lieberum et al., 2024). An SAE consist of an encoder and a decoder; the encoder maps a latent vector into a sparse, high-dimensional concept vector, where each dimension — SAE neuron — represents a distinct, interpretable concept, characterized by the inputs that strongly activate it (Paulo et al., 2024); the decoder reconstructs the original latent vector from the concept vector.

Since training separate SAEs for each (sub)layer in LLMs can be computationally intensive and redundant, later research enhances scalability and expressiveness through new architectures (Rajamanoharan et al., 2024a; Templeton et al., 2024; Dunefsky et al., 2024a; Mudide et al., 2025), activation functions (Rajamanoharan et al., 2024b), and training strategies (Kissane et al., 2024; Ghilardi et al., 2024; Braun et al., 2024; Shi et al., 2025; Farnik et al., 2025). These advances have discovered more diverse concepts (O'Neill et al., 2024; Templeton et al., 2024; He et al., 2024e), offering insights into LLMs' hallucinations (Ferrando et al., 2025; Theodorus et al., 2025), jailbreaks (Härle et al., 2024; Muhamed et al., 2025; Gallifant et al., 2025), biases (Hegde, 2024; Zhou et al., 2025a), and privacy leakage (Frikha et al., 2025).

Beyond individual neurons, **SAE circuits** are extracted to reveal how interpretable concepts interact to produce specific outputs (He et al., 2024d; Dunefsky et al., 2024c; Marks et al., 2025; Balagansky et al., 2025). SAE variants, such as Crosscoder (Lindsey et al., 2024) and Transcoder (Dunefsky et al., 2024a,b; Ameisen et al., 2025), enhance circuit interpretability and reduce redundancy, making it easier to isolate mechanisms behind unsafe behaviors (Lindsey et al., 2025).

In parallel, **logit lens** projects intermediate latent vectors onto the model's vocabulary space using the final projection matrix, viewing latent vectors on the vocabulary level (nostalgebraist, 2020; Elhage et al., 2021; Geva et al., 2022b; Dar et al., 2023). Further research enhances its robustness (Belrose et al., 2023; Din et al., 2023) and extends it to analyze training dynamics (Katz et al., 2024). The logit lens has been leveraged to investigate LLMs' knowledge store and recall mechanisms (Haviv et al., 2023; Yu et al., 2023) and safety issues, such as hallucinations (Yu et al., 2024b; Jiang et al., 2024a; Halawi et al., 2024; Jin et al., 2024) and jailbreaks and harmfulness (Zhao et al., 2024d; Feng et al., 2024; Lee et al., 2024).

3.4 Self-explain with Reason Generation

The rapid advances in LLMs' self-explanatory and reasoning capabilities has prompted a surge of recent work aiming to explore how LLMs can in-

[4]Path patching differs from activation patching in that it selectively modifies only the information flowing along specific paths, whereas activation patching replaces entire activations at specific components.

terpret their own outputs by expressing the reasoning behind them in natural language (Huang and Chang, 2023; Zhao et al., 2024b; Yu et al., 2024a). A common approach is **in-generation reasoning**, where LLMs are prompted or trained to generate responses along with rationales (Camburu et al., 2018; Rajani et al., 2019; Marasovic et al., 2022) or uncertainty estimates (Kadavath et al., 2022; Chaudhry et al., 2024; Amayuelas et al., 2024; Xu et al., 2024a). Chain-of-thought (CoT) prompting is a notable example, where LLMs generate intermediate reasoning steps to reach ananswer (Wei et al., 2022; Zhao et al., 2023a; Chu et al., 2025; Cahlik et al., 2025). Many CoT variants support more complex reasoning (Yao et al., 2023; Besta et al., 2024) and improve explanation faithfulness (Qu et al., 2022; Lyu et al., 2023; Tafjord et al., 2022; Creswell and Shanahan, 2022; Creswell et al., 2023). However, such explanations can be unreliable (Gao et al., 2023; Ye and Durrett, 2022; Araya, 2025), necessitating further verification (Ye and Durrett, 2022; Turpin et al., 2023; Weng et al., 2023; Miao et al., 2024).

Alternatively, **post-hoc explanation** methods assess and explain a response after generation (Jiang et al., 2024b; Binder et al., 2025). These methods prompt LLMs to evaluate the correctness or safety of their outputs and provide rationales (Li et al., 2024b; Liu et al., 2025a; Betley et al., 2025). To explain hallucinations, a response may be split into factual claims or questions, which the model is then asked to verify against its knowledge (Dhuliawala et al., 2024; Akbar et al., 2024; Lee et al., 2025b).

4 Enhance Safety using Interpretation

Recent advances in LLM interpretation (§3) have inspired techniques to enhance model safety. This section reviews methods that leverage interpretation to mitigate unsafe behaviors, following the stages of the LLM workflow discussed in §3.

4.1 Attend to Relevant Input Tokens (§3.2)

Some methods prompt LLMs to attend to relevant input tokens to reduce hallucinations (Liu et al., 2025b) and improve factuality (Krishna et al., 2023). Others remove jailbreak-triggering tokens (Pan et al., 2025a) or manipulate attention to user-specified relevant tokens (Zhang et al., 2024b).

4.2 Modify Model Internals for Safety (§3.3)

4.2.1 Steer Latent Vectors To Safe Directions

Representation engineering guides LLMs' latent vectors toward safe regions by adding safety vectors identified by probing (Zou et al., 2023) or training transformations that map unsafe vectors into safe regions (Hernandez et al., 2024a) (§3.3.1). These methods mitigate a range of safety concerns (Qian et al., 2024; Rimsky et al., 2024; Singh et al., 2024b; Chu et al., 2024), such as hallucinations (Li et al., 2023a; Yang et al., 2024a; Zhang et al., 2024c; Duan et al., 2024), jailbreaks and harmfulness (Turner et al., 2023; Bhattacharjee et al., 2024; Zhu et al., 2024; Ball et al., 2024; Gao et al., 2024; Shen et al., 2024; Li et al., 2025c; Han et al., 2025), and bias (Hernandez et al., 2024a).

4.2.2 Modulate (Un)Safe Neurons' Activations

Suppressing risky neurons or amplifying safer ones guides LLMs away from unsafe behaviors. SAEs help locate and control (un)safe SAE neuron activations (Soo et al., 2025) (§3.3.3), addressing risks (He et al., 2025) like hallucinations (Abdaljalil et al., 2025; Bayat et al., 2025), jailbreaks and harmfulness (O'Brien et al., 2024; Härle et al., 2024; Khoriaty et al., 2025; Wu et al., 2025), biases (Liu et al., 2024b; Hegde, 2024; Marks et al., 2025; Zhou et al., 2025a), and privacy leaks (Frikha et al., 2025). Alternatives remove dependency on SAEs by using logit lens (§3.3.3) to find and upscale safe MLP sublayers (Geva et al., 2022b; Wang et al., 2024a) or amplify safety-critical attention weights (§3.3.1, §3.3.2) on user-specified reliable tokens (Zhang et al., 2024b; Deng et al., 2025).

4.2.3 Edit Harmful Model Components

Safety can be improved by pruning or downscaling components (e.g., attention heads or (sub)layers) linked to hallucinations (Li et al., 2024a; Yu et al., 2024b), jailbreaks (Zhao et al., 2024d; Wang et al., 2024c; Li et al., 2024e, 2025b; Wang et al., 2025b), and biases (Ma et al., 2023). Other techniques identify safety directions in parameter space by comparing aligned and unaligned model weights, then steer critical parameters accordingly (Hazra et al., 2024; Wang et al., 2024a; Zhao et al., 2024e).

4.3 Verify Safety before Outputs (§3.3, §3.4)

Some approaches generate multiple candidate responses, evaluate their safety (§3.3, §3.4), and select only safe ones to construct final output (Burns et al., 2022; Zhao et al., 2023b; Weng et al., 2023;

Dale et al., 2023; Miao et al., 2024; Chuang et al., 2024; Dhuliawala et al., 2024; Chen et al., 2025). Others intervene during generation, resamping tokens when an unsafe sequence is detected (Li et al., 2024c; Cheng et al., 2025b) or producing refusal messages (Zhao et al., 2024a; Mou et al., 2025).

4.4 Output with Self-Reasoning (§3.4).

Building on CoT reasoning's success for performance and interpretation (Ji et al., 2024a), several approaches fine-tune or prompt LLMs to generate intermediate reasoning steps to reinforce safety constraints during generation (Kaneko et al., 2024; Prahallad and Mamidi, 2024; Li et al., 2024d; Cao et al., 2024; Sicilia and Alikhani, 2024; Moore et al., 2024; Rad et al., 2025; Zhang et al., 2025; Mou et al., 2025). For instance, GuardReasoner (Liu et al., 2025a) prompts models to explain why a response may be harmful, enabling safer behavior through self-reflection.

5 Tools Operationalizing Safety-Focused Interpretation

To apply interpretation methods (§3) and safety enhancement strategies (§4), practitioners need tools that support actionability (Kaur et al., 2020; Lakkaraju et al., 2022; Sharkey et al., 2025). Some efforts focus on developing libraries to **ease implementation** of interpretation and safety techniques (Choe et al., 2024; Kokhlikyan et al., 2020; Sarti et al., 2023; Hao et al., 2024), while others introduce interactive visual tools, inspired by their effectiveness in enhancing human understanding of classical AI models (Hohman et al., 2019; Beauxis-Aussalet et al., 2021; La Rosa et al., 2023; Liao and Wortman Vaughan, 2024; Wang et al., 2025c).

5.1 Training Data Attribution (TDA) Visualization (§3.1) shows how training examples shape model behavior. A prominent tool is LLM Attributor (Lee et al., 2025a), which traces outputs to training data, identifying hallucination sources.

5.2 Input Token Visualizations (§3.2) reveal individual tokens' importance and attention patterns, showing how tokens influence one another across heads. These are incorporated into LLM analysis tools (Park et al., 2019; Tenney et al., 2020; Wang et al., 2021; Li et al., 2023b; Coscia et al., 2024; Yeh et al., 2024), and reveal spurious token associations indicative of bias (Vig, 2019). Many tools also support interactive perturbation, allowing users to edit tokens or attention weights and observe the effects (Strobelt et al., 2019; Tenney et al., 2020; Coscia et al., 2024; Mishra et al., 2025).

5.3 Latent vector visualizations (§3.3) show how concepts are encoded and propagated during model inference. Some tools project latent vectors into 2D space (Tenney et al., 2020; Li et al., 2023b; Kwon and Mihindukulasooriya, 2023). Others visualize semantics of latent vectors revealed by logit lens (Katz and Belinkov, 2023; Pal et al., 2023; Hernandez et al., 2024b) (§3.3.3), while some enable users to steer latent vectors for safer outputs (Geva et al., 2022a; Chen et al., 2024b) (§4).

5.4 Neuron visualizations (§3.3) display data points that highly activate each neuron during inference, revealing interpretable concepts (Nanda, 2022; Garde et al., 2023; Bills et al., 2023). (§3.3.3). Similar approaches are applied to SAE neurons (Lin, 2023), helping concept identification and SAE circuit discovery (Chalnev et al., 2024) for multiple unsafe behaviors (Lindsey et al., 2025; Ameisen et al., 2025).

6 Research Directions and Conclusion

Our survey identifies five key open problems at the intersection of interpretation and safety research. As interpretation tools enter high-stake use, a pressing concern is their potential misuse, particularly when adversaries exploit interpretation methods to attack LLMs (§6.1). Mitigating this risk requires interpretations that are both faithful to model behavior and understandable to users (§6.2). With such reliability, interpretations can be leveraged proactively during training to promote safer model behaviors (§6.3). Their impact also depends on how effectively they are presented, underscoring the need for more user-aligned communication strategies (§6.4). Finally, we advocate for expanding interpretation research to address a broader range of safety risks beyond those currently emphasized (§6.5).

6.1 Defense against Interpretation-based Attack

Interpretation-driven attacks pose a unique threat, distinct from traditional adversarial attacks, as they exploit insights gained from interpretation methods (Lin et al., 2024; Li et al., 2024f; Arditi et al., 2024; Su, 2024; Winninger et al., 2025). For example, interpretation techniques can reveal why certain input tokens trigger risky outputs, enabling malicious users to craft diverse prompt suffixes that

elicit unsafe behavior. Similarly, adversaries can use interpretation to identify and manipulate model components associated with harmful outputs. Because these threats arise unexpectedly from tools intended to enhance transparency and safety, defending against them is particularly challenging. This underscores the need to design interpretation methods with potential misuse in mind, along with implementing stronger defenses—such as proactively removing risky content (Wu et al., 2025).

6.2 Reliable Evaluation of Interpretation

The field currently lacks a clear consensus on the definitions of faithfulness and interpretability, as well as standardized methods for assessing interpretability (Shi et al., 2024a). Interdisciplinary collaboration—spanning machine learning, human-computer interaction, psychology, and cognitive science—is essential for developing both formal definitions and practical evaluation frameworks. Moreover, even in human-centered evaluations, interpretability scores are often influenced by how interpretation results are presented to users (Liao and Wortman Vaughan, 2024). This highlights the need for standardized evaluation protocols that consider both the presentation format and user characteristics, such as domain expertise and task relevance (Alangari et al., 2023).

6.3 Training Attribution for Safety Enhancement

TDA shows promise for tracing unsafe behavior to training examples (§3.1), but its use in safety enhancement is limited. Prior work on retraining after removing problematic data (Kong et al., 2022; Mozes et al., 2023) focuses on non-safety issues on small non-generative models and cannot be scaled to LLMs. Developing safety enhancement methods based on training attribution could open new paths for risk mitigation.

6.4 User-centered Presentation of Safety Interpretations

How to present interpretation results to assist safety-critical decisions remains underexplored (§5). In particular, presentation of long, complex textual explanations from LLMs should be further investigated (§3.4); conversational interaction helps human understanding (Slack et al., 2023; Wang et al., 2024d), yet no tools apply this to safety-oriented interpretation. Future work should explore interaction and design strategies tailored to diverse stakeholders.

6.5 Refining Safety Dimensions

Interpretation research has largely focused on hallucinations, jailbreaks, harmfulness, bias, and privacy leakage, while other risks—like out-of-distribution robustness, code security, and over-refusal—are understudied (Siska and Sankaran, 2025; Yang et al., 2024b; Xiong et al., 2024; Abdaljalil et al., 2025). Incorporating user intent and social impact into safety definitions may enable more nuanced and targeted interpretations (Sarker, 2024).

7 Conclusion

By bridging the gap between interpretation and safety research, our survey systematically examines interpretation methods across the LLM workflow, safety enhancement strategies, and practical tools, while highlighting open problems and future directions.

8 Limitations

This survey provides an overview and categorization of interpretation techniques, with an emphasis on their role in improving the safety of LLMs. Given the fast-evolving and extensive nature of the field, some latest advancements may not be included. We focus on autoregressive Transformer-based generative LLMs, as they are among the most widely used an studied models for interpretation; therefore, the interpretation and safety enhancement techniques we discuss may not generalize to other model architectures. Our paper selection aims to capture the breadth and diversity of existing approaches, though full technical details are omitted due to space constraints. We also highlight tools that facilitate understanding and use of interpretation results, recognizing that notions of practicality can vary across stakeholders and that actionability of interpretation remains an actively researched open question. Despite its limitations, this survey introduces a taxonomy that can help newcomers quickly understand the landscape of interpretation for safety and guide future research exploring its application to other model architectures and emerging techniques.

9 Potential Risks

Our paper focuses on four major safety concerns addressed by interpretation research (hallucination, jailbreaks and harmfulness, bias, and privacy leakage), but this view may be too narrow, risking overlooking other safety issues such as code security and over-refusal (§6). While our survey covers a wide range of interpretation techniques, it does not include quantitative comparisons. As a result, readers may overly rely on certain techniques or mistakenly assume that interpretation guarantees safety.

Acknowledgements

ChatGPT was used to check grammar and spelling of this paper.

References

Sina Abbasi, Mohammad Reza Modarres, and Mohammad Taher Pilehvar. 2024. Normxlogit: The head-on-top never lies. *arXiv preprint arXiv:2411.16252*.

Samir Abdaljalil, Filippo Pallucchini, Andrea Seveso, Hasan Kurban, Fabio Mercorio, and Erchin Serpedin. 2025. SAFE: A sparse autoencoder-based framework for robust query enrichment and hallucination mitigation in llms. *arXiv preprint arXiv:2503.03032*.

Sahar Abdelnabi, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. 2024. Are you still on track!? catching llm task drift with activations. *arXiv preprint arXiv:2406.00799*.

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Rohan Ajwani, Shashidhar Reddy Javaji, Frank Rudzicz, and Zining Zhu. 2024. Llm-generated black-box explanations can be adversarially helpful. *arXiv preprint arXiv:2405.06800*.

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. 2024. HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037, Miami, Florida, USA. Association for Computational Linguistics.

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Nourah Alangari, Mohamed El Bachir Menai, Hassan Mathkour, and Ibrahim Almosallam. 2023. Exploring evaluation methods for interpretable machine learning: A survey. *Information*, 14(8):469.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, Bangkok, Thailand. Association for Computational Linguistics.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig

Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit tracing: Revealing computational graphs in language models. https://transformer-circuits.pub/2025/attribution-graphs/methods.html. Transformer Circuits Thread.

Roberto Araya. 2025. Do chains-of-thoughts of large language models suffer from hallucinations, cognitive biases, or phobias in bayesian reasoning? *arXiv preprint arXiv:2503.15268*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Leila Arras, Bruno Puri, Patrick Kahardipraja, Sebastian Lapuschkin, and Wojciech Samek. 2025. A close look at decomposition-based xai-methods for transformer language models. *arXiv preprint arXiv:2502.15886*.

Dhananjay Ashok and Jonathan May. 2025. Language models can predict their own behavior. *arXiv preprint arXiv:2502.13329*.

Suriya Ganesh Ayyamperumal and Limin Ge. 2024. Current state of llm risks and ai guardrails. *arXiv preprint arXiv:2406.12934*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Nikita Balagansky, Ian Maksimov, and Daniil Gavrilov. 2025. Mechanistic permutability: Match features across layers. In *The Thirteenth International Conference on Learning Representations*.

Sarah Ball, Frauke Kreuter, and Nina Panickssery. 2024. Understanding jailbreak success: A study of latent space dynamics in large language models. *arXiv preprint arXiv:2406.09289*.

Oren Barkan, Yehonatan Elisha, Yonatan Toib, Jonathan Weill, and Noam Koenigstein. 2024a. Improving LLM attributions with randomized path-integration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9430–9446, Miami, Florida, USA. Association for Computational Linguistics.

Oren Barkan, Yonatan Toib, Yehonatan Elisha, and Noam Koenigstein. 2024b. A learning-based approach for explaining language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 98–108, New York, NY, USA. Association for Computing Machinery.

Oren Barkan, Yonatan Toib, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. 2024c. LLM explainability via attributive masking learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9522–9537, Miami, Florida, USA. Association for Computational Linguistics.

Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. 2020. Relatif: Identifying explanatory training samples via relative influence. In *International Conference on Artificial Intelligence and Statistics*, pages 1899–1909. PMLR.

Samyadeep Basu, Phil Pope, and Soheil Feizi. 2021. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. Steering large language model activations in sparse spaces. *arXiv preprint arXiv:2503.00177*.

Emma Beauxis-Aussalet, Michael Behrisch, Rita Borgo, Duen Horng Chau, Christopher Collins, David Ebert, Mennatallah El-Assady, Alex Endert, Daniel A Keim, Jörn Kohlhammer, et al. 2021. The role of interactive visualization in fostering trust in ai. *IEEE Computer Graphics and Applications*, 41(6):7–12.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. 2024. Internalinspector i^2: Robust confidence estimation in llms through internal states. *arXiv preprint arXiv:2406.12053*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna

Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. 2025. Tell me about yourself: Llms are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. 2024. Finding transformer circuits with edge pruning. *Advances in Neural Information Processing Systems*, 37:18506–18534.

Amrita Bhattacharjee, Shaona Ghosh, Traian Rebedea, and Christopher Parisien. 2024. Towards inference-time category-wise safety steering for large language models. In *Neurips Safe Generative AI Workshop 2024*.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2023a. Llms as counterfactual explanation modules: can chatgpt explain blackbox text classifiers. *arXiv preprint arXiv:2309.13340*, 4.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2023b. Towards llm-guided causal explainability for black-box text classifiers. *arXiv preprint arXiv:2309.13340*.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html. OpenAI.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. 2025. Looking inward: Language models can learn about themselves by introspection. In *The Thirteenth International Conference on Learning Representations*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. 2024. Identifying functionally important features with end-to-end sparse dictionary learning. *Advances in Neural Information Processing Systems*, 37:107286–107325.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Xingyu Yang, Francois Charton, and Julia Kempe. 2024. Iteration head: A mechanistic study of chain-of-thought. *Advances in Neural Information Processing Systems*, 37:109101–109122.

Vojtech Cahlik, Rodrigo Alves, and Pavel Kordik. 2025. Reasoning-grounded natural language explanations for language models. *arXiv preprint arXiv:2503.11248*.

Nitay Calderon and Roi Reichart. 2025. On behalf of the stakeholders: Trends in NLP model interpretability in the era of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 656–693, Albuquerque, New Mexico. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Yanfei Cao, Naijie Gu, Xinyue Shen, Daiyuan Yang, and Xingmin Zhang. 2024. Defending large language models against jailbreak attacks through chain of thought prompting. In *2024 International Conference on Networking and Network Applications (NaNA)*, pages 125–130.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4401–4420, Bangkok, Thailand. Association for Computational Linguistics.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *Preprint*, arXiv:2411.02193.

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. 2025. Scalable influence and fact tracing for large language model pretraining. In *The Thirteenth International Conference on Learning Representations*.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. 2019. Neural network attributions: a causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR.

Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. 2024. Finetuning language models to emit linguistic expressions of uncertainty. *arXiv preprint arXiv:2409.12180*.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. 2024a. Finding safety neurons in large language models. *Preprint*, arXiv:2406.14144.

Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. 2024b. Designing a dashboard for transparency and control of conversational ai. *Preprint*, arXiv:2406.07882.

Yuyan Chen, Zehao Li, Shuangjie You, Zhengyu Chen, Jingwen Chang, Yi Zhang, Weinan Dai, Qingpei Guo, and Yanghua Xiao. 2025. Attributive reasoning for hallucination diagnosis of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23660–23668.

Deric Cheng, Juhan Bae, Justin Bullock, and David Kristofferson. 2025a. Training data attribution (tda): Examining its adoption & use cases. *arXiv preprint arXiv:2501.12642*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025b. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *arXiv preprint arXiv:2501.01306*.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, et al. 2024. What is your data worth to gpt? llm-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954*.

Xu Chu, Zhijie Tan, Hanlin Xue, Guanyu Wang, Tong Mo, and Weiping Li. 2025. Domaino1s: Guiding llm reasoning for explainable answers in high-stakes domains. *arXiv preprint arXiv:2501.14431*.

Zhixuan Chu, Yan Wang, Longfei Li, Zhibo Wang, Zhan Qin, and Kui Ren. 2024. A causal explainable guardrails for large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 1136–1150, New York, NY, USA. Association for Computing Machinery.

Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. 2024. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

R Dennis Cook and Sanford Weisberg. 1980. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508.

Adam Coscia, Langdon Holmes, Wesley Morris, Joon Suh Choi, Scott Crossley, and Alex Endert. 2024. iscore: Visual analytics for interpreting how language models automatically score summaries. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 787–802, New York, NY, USA. Association for Computing Machinery.

Ian Covert, Scott Lundberg, and Su-In Lee. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada. Association for Computational Linguistics.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. 2023. Discovering variable binding circuitry with desiderata. *arXiv preprint arXiv:2307.03637*.

Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2023. Atman: understanding transformer predictions through memory efficient attention manipulation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2025. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23760–23768.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2023. Jump to conclusions: Short-cutting transformers with linear transformations. *arXiv preprint arXiv:2303.09435*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2024. Do llms know about hallucination? an empirical investigation of llm's hidden states. *arXiv preprint arXiv:2402.09733*.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024a. Transcoders enable fine-grained interpretable circuit analysis for language models. https://www.lesswrong.com/posts/YmkjnWtZGLbHRbzrP/transcoders-enable-fine-grained-interpretable-circuit. LessWrong.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024b. Transcoders find interpretable LLM feature circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jacob Dunefsky, Philippe Chlenski, Senthooran Rajamanoharan, and Neel Nanda. 2024c. Case studies in reverse-engineering sparse autoencoder features by using mlp linearization. https://www.lesswrong.com/posts/93nKtsDL6YY5fRbQv/case-studies-in-reverse-engineering-sparse-autoencoder. LessWrong.

Oliver Eberle, Ilias Chalkidis, Laura Cabello, and Stephanie Brandl. 2023. Rather a nurse than a physician - contrastive explanations under investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6907–6920, Singapore. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Logan Engstrom, Axel Feldmann, and Aleksander Mądry. 2024. Dsdm: model-aware dataset selection with datamodels. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Joseph Enguehard. 2023. Sequential integrated gradients: a simple but effective method for explaining language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7555–7565, Toronto, Canada. Association for Computational Linguistics.

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. 2024. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13984–14011, Bangkok, Thailand. Association for Computational Linguistics.

Lucy Farnik, Tim Lawson, Conor Houghton, and Laurence Aitchison. 2025. Jacobian sparse autoencoders: Sparsify computations, not just activations. *arXiv preprint arXiv:2502.18147*.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891. Curran Associates, Inc.

Zijian Feng, Hanzhang Zhou, ZIXIAO ZHU, Junlang Qian, and Kezhi Mao. 2024. Unveiling and manipulating prompt influence in large language models. In *The Twelfth International Conference on Learning Representations*.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022a. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022b. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-Jussà. 2024. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Alex Foote, Neel Nanda, Esben Kran, Ioannis Konstas, Shay Cohen, and Fazl Barez. 2023. Neuron to graph: Interpreting language model neurons at scale. *Preprint*, arXiv:2305.19911.

Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. Privacyscalpel: Enhancing llm privacy via interpretable feature intervention with sparse autoencoders. *arXiv preprint arXiv:2503.11232*.

Jack Gallifant, Shan Chen, Kuleen Sasse, Hugo Aerts, Thomas Hartvigsen, and Danielle S Bitterman. 2025. Sparse autoencoder features for classifications and transferability. *arXiv preprint arXiv:2502.11367*.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Lang Gao, Xiangliang Zhang, Preslav Nakov, and Xiuying Chen. 2024. Shaping the safety boundaries: Understanding and defending against jailbreaks in large language models. *arXiv preprint arXiv:2412.17034*.

Albert Garde, Esben Kran, and Fazl Barez. 2023. Deepdecipher: Accessing and investigating neuron activation in large language models. *Preprint*, arXiv:2310.01870.

Yair Ori Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2024. Faithful explanations of black-box NLP models using LLM-generated counterfactuals. In *The Twelfth International Conference on Learning Representations*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022a. LM-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21, Abu Dhabi, UAE. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022b. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: a unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Davide Ghilardi, Federico Belotti, and Marco Molinari. 2024. Efficient training of sparse autoencoders for large language models via layer groups. *arXiv preprint arXiv:2410.21508*.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. 2021. FastIF: Scalable influence functions for efficient model interpretation and debugging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10333–10350, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.

Kelvin Guu, Albert Webson, Ellie Pavlick, Lucas Dixon, Ian Tenney, and Tolga Bolukbasi. 2023. Simfluence: Modeling the influence of individual training examples by simulating training runs. *arXiv preprint arXiv:2303.08114*.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. In *The Twelfth International Conference on Learning Representations*.

Zayd Hammoudeh and Daniel Lowd. 2024. Training data influence analysis and estimation: a survey. *Mach. Learn.*, 113(5):2351–2403.

Frank R Hampel. 1974. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, and Heng Ji. 2025. Internal activation as the polar star for steering unsafe llm behavior. *arXiv preprint arXiv:2502.01042*.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4398–4409, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2022. Orca: Interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data. *arXiv preprint arXiv:2205.12600*.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than? interpreting mathematical abilities in a pre-trained language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. LLM reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*.

Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. 2024. Scar: Sparse conditioned autoencoders for concept detection and steering in llms. *arXiv preprint arXiv:2411.07122*.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Rima Hazra, Sayan Layek, Somnath Banerjee, and Soujanya Poria. 2024. Safety arithmetic: A framework for test-time safety alignment of language models by steering parameters and activations. In *Proceedings of the 2024 Conference on Empirical Methods in*

Natural Language Processing, pages 21759–21776, Miami, Florida, USA. Association for Computational Linguistics.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024a. LLM factoscope: Uncovering LLMs' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10218–10230, Bangkok, Thailand. Association for Computational Linguistics.

Luxi He, Mengzhou Xia, and Peter Henderson. 2024b. What is in your safe data? identifying benign data that breaks safety. In *First Conference on Language Modeling*.

Zeqing He, Zhibo Wang, Zhixuan Chu, Huiyu Xu, Rui Zheng, Kui Ren, and Chun Chen. 2024c. Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit. *arXiv preprint arXiv:2411.11114*.

Zeqing He, Zhibo Wang, Huiyu Xu, and Kui Ren. 2025. Towards llm guardrails via sparse representation steering. *arXiv preprint arXiv:2503.16851*.

Zhengfu He, Xuyang Ge, Qiong Tang, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024d. Dictionary learning improves patch-free circuit discovery in mechanistic interpretability: A case study on othello-gpt. *arXiv preprint arXiv:2402.12201*.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. 2024e. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Praveen Hegde. 2024. Effectiveness of sparse autoencoder for understanding and removing gender bias in LLMs. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024a. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024b. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Fabian Hildebrandt, Andreas Maier, Patrick Krauss, and Achim Schilling. 2025. Refusal behavior in large language models: A nonlinear perspective. *arXiv preprint arXiv:2501.08145*.

Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2019. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693.

Miriam Horovicz and Roni Goldshmidt. 2024. TokenSHAP: Interpreting large language models with Monte Carlo shapley value estimation. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 1–8, Miami, FL, USA. Association for Computational Linguistics.

Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, et al. 2024. A comprehensive guide to explainable ai: from classical models to llms. *arXiv preprint arXiv:2412.00800*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023a. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024. Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2409.18169*.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gao Jin, Yizhen Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, André Freitas, and Mustafa A. Mustafa. 2023b. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artif. Intell. Rev.*, 57:175.

Yunpeng Huang, Jingwei Xu, Junyu Lai, Zixu Jiang, Taolue Chen, Zenan Li, Yuan Yao, Xiaoxing Ma, Lijuan Yang, Hao Chen, et al. 2023c. Advancing transformer architecture in long-context large language models: A comprehensive survey. *arXiv preprint arXiv:2311.12351*.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. 2024. Scar: Sparse conditioned autoencoders for concept detection and steering in llms. *Preprint*, arXiv:2411.07122.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Datamodels: Predicting predictions from training data. In *ArXiv preprint arXiv:2202.00622*.

Tatsuro Inaba, Kentaro Inui, Yusuke Miyao, Yohei Oseki, Benjamin Heinzerling, and Yu Takagi.

2025. How llms learn: Tracing internal representations with sparse autoencoders. *arXiv preprint arXiv:2503.06394*.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bin Ji, Huijun Liu, Mingzhe Du, and See-Kiong Ng. 2024a. Chain-of-thought improves text generation with citations in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18345–18353.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024b. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024a. On large language models' hallucination with regard to known facts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053, Mexico City, Mexico. Association for Computational Linguistics.

Weisen Jiang, Han Shi, Longhui Yu, Zhengying Liu, Yu Zhang, Zhenguo Li, and James Kwok. 2024b. Forward-backward reasoning in large language models for mathematical verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6647–6661, Bangkok, Thailand. Association for Computational Linguistics.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Shahar Katz and Yonatan Belinkov. 2023. VISIT: Visualizing and interpreting the semantic information flow of transformers. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. 2024. Backward lens: Projecting language model gradients into the vocabulary space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2390–2422, Miami, Florida, USA. Association for Computational Linguistics.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Matthew Khoriaty, Andrii Shportko, Gustavo Mercier, and Zach Wood-Doughty. 2025. Don't forget it! conditional sparse autoencoder clamping works for unlearning. *arXiv preprint arXiv:2503.11127*.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. 2024. Interpreting attention layer outputs with sparse autoencoders. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language

Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2024. Analyzing feed-forward blocks in transformers through the lens of attention maps. In *The Twelfth International Conference on Learning Representations*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Shuming Kong, Yanyan Shen, and Linpeng Huang. 2022. Resolving training biases via influence-based data relabeling. In *International Conference on Learning Representations*.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. Atp*: An efficient and scalable method for localizing llm behaviour to components. *arXiv preprint arXiv:2403.00745*.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Bum Chul Kwon and Nandana Mihindukulasooriya. 2023. Finspector: A human-centered visual inspection tool for exploring and comparing biases among foundation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 42–50, Toronto, Canada. Association for Computational Linguistics.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2024. Datainf: Efficiently estimating data influence in loRA-tuned LLMs and diffusion models. In *The Twelfth International Conference on Learning Representations*.

Biagio La Rosa, Graziano Blasilli, Romain Bourqui, David Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini. 2023. State of the art of visual analytics for explainable deep learning. In *Computer Graphics Forum*, volume 42, pages 319–355. Wiley Online Library.

Faisal Ladhak, Esin Durmus, and Tatsunori Hashimoto. 2023. Contrastive error attribution for finetuned language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11482–11498, Toronto, Canada. Association for Computational Linguistics.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *Preprint*, arXiv:2202.01875.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: a case study on dpo and toxicity. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Seongmin Lee, Zijie J Wang, Aishwarya Chakravarthy, Alec Helbling, ShengYun Peng, Mansi Phute, Duen Horng Polo Chau, and Minsuk Kahng. 2025a. Llm attributor: Interactive visual attribution for llm generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29655–29657.

Sujeong Lee, Hayoung Lee, Seongsoo Heo, and Wonik Choi. 2025b. Hudex: Integrating hallucination detection and explainability for enhancing the reliability of llm responses. *arXiv preprint arXiv:2502.08109*.

Chak Tou Leong, Yi Cheng, Kaishuai Xu, Jian Wang, Hanlin Wang, and Wenjie Li. 2024. No two devils alike: Unveiling distinct mechanisms of fine-tuning attacks. *arXiv preprint arXiv:2405.16229*.

Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27.

Chuang Li, Bingnan Xing, Dongdong Huo, Qihui Zhou, Zhen Xu, and Yu Wang. 2025a. Mixhd: A method for detecting hallucinations based on the internal state and output probability of large language models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

He Li, Haoang Chi, Mingyu Liu, and Wenjing Yang. 2024a. Look within, why llms hallucinate: A causal perspective. *arXiv preprint arXiv:2407.10153*.

Jing-Jing Li, Valentina Pyatkin, Max Kleiman-Weiner, Liwei Jiang, Nouha Dziri, Anne GE Collins, Jana Schaich Borg, Maarten Sap, Yejin Choi, and Sydney Levine. 2024b. Safetyanalyst: Interpretable, transparent, and steerable llm safety moderation. *arXiv preprint arXiv:2410.16665*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Raymond Li, Ruixin Yang, Wen Xiao, Ahmed Abu-Raed, Gabriel Murray, and Giuseppe Carenini. 2023b. Visual analytics for generative transformer models. *Preprint*, arXiv:2311.12418.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2025b. Safety layers in aligned large language models: The key to LLM security. In *The Thirteenth International Conference on Learning Representations*.

Tianlong Li, Zhenghua Wang, Wenhao Liu, Muling Wu, Shihan Dou, Changze Lv, Xiaohua Wang, Xiaoqing Zheng, and Xuanjing Huang. 2025c. Revisiting jailbreaking for large language models: A representation engineering perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3158–3178, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianyi Li, Erenay Dayanik, Shubhi Tyagi, and Andrea Pierleoni. 2024c. Hallucana: Fixing llm hallucination with a canary lookahead. *arXiv preprint arXiv:2412.07965*.

Xi Li, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. 2024d. Chain-of-scrutiny: Detecting backdoor attacks for large language models. *arXiv preprint arXiv:2406.05948*.

Xuying Li, Zhuo Li, Yuji Kosuga, Yasuhiro Yoshida, and Victor Bian. 2024e. Precision knowledge editing: Enhancing safety in large language models. *arXiv preprint arXiv:2410.03772*.

Yuxi Li, Zhibo Zhang, Kailong Wang, Ling Shi, and Haoyu Wang. 2024f. Model-editing-based jailbreak against safety-aligned large language models. *Preprint*, arXiv:2412.08201.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. 2024g. Do influence functions work on large language models? *arXiv preprint arXiv:2409.19998*.

Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5). Https://hdsr.mitpress.mit.edu/pub/aelql9qy.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Johnny Lin. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. Software available from neuronpedia.org.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. 2024. Towards understanding jailbreak attacks in LLMs: A representation space analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7067–7085, Miami, Florida, USA. Association for Computational Linguistics.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the biology of a large language model. https://transformer-circuits.pub/2025/attribution-graphs/biology.html. Transformer Circuits Thread.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. 2024. Sparse crosscoders for cross-layer features and model diffing. https://transformer-circuits.pub/2024/crosscoders/index.html. Transformer Circuits Thread.

Fuxiao Liu, Paiheng Xu, Zongxia Li, Yue Feng, and Hyemi Song. 2023. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*.

Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. 2024a. On the universal truthfulness hyperplane inside LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18199–18224, Miami, Florida, USA. Association for Computational Linguistics.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024b. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *ICLR*.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. 2025a. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*.

Zhining Liu, Rana Ali Amjad, Ravinarayana Adkathimar, Tianxin Wei, and Hanghang Tong. 2025b. Selfelicit: Your language model secretly knows where is the relevant evidence. *arXiv preprint arXiv:2502.08767*.

Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. 2023. Interpretable-by-design text understanding with iteratively generated concept bottleneck. *arXiv preprint arXiv:2310.19660*.

Niclas Luick. 2024. Universal response and emergence of induction in llms. *arXiv preprint arXiv:2411.07071*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and

Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Chiyu Ma, Lin Shi, Ollie Liu, Wenhua Liang, Jiaqi Gan, Ming Cheng, Willie Neiswanger, and Soroush Vosoughi. 2024. Mechanistic insights: Circuit transformations across input and fine-tuning landscapes. *OpenReview*.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. 2023. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11328–11345, Singapore. Association for Computational Linguistics.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, Hanxun Huang, Yige Li, Jiaming Zhang, Xiang Zheng, Yang Bai, Zuxuan Wu, Xipeng Qiu, Jingfeng Zhang, Yiming Li, Xudong Han, Haonan Li, Jun Sun, Cong Wang, Jindong Gu, Baoyuan Wu, Siheng Chen, Tianwei Zhang, Yang Liu, Mingming Gong, Tongliang Liu, Shirui Pan, Cihang Xie, Tianyu Pang, Yinpeng Dong, Ruoxi Jia, Yang Zhang, Shiqing Ma, Xiangyu Zhang, Neil Gong, Chaowei Xiao, Sarah Erfani, Tim Baldwin, Bo Li, Masashi Sugiyama, Dacheng Tao, James Bailey, and Yu-Gang Jiang. 2025. Safety at scale: A comprehensive survey of large model safety. *Preprint*, arXiv:2502.05206.

Mohammad Reza Ghasemi Madani, Aryo Pradipta Gema, Gabriele Sarti, Yu Zhao, Pasquale Minervini, and Andrea Passerini. 2025. Noiser: Bounded input perturbations for attributing large language models. *arXiv preprint arXiv:2504.02911*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis Bach. 2008. Supervised dictionary learning. *Advances in neural information processing systems*, 21.

Alireza Makhzani and Brendan Frey. 2013. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Amanda McGrath and Alexandra Jonker. 2024. What is ai interpretability?

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *The Twelfth International Conference on Learning Representations*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Aditi Mishra, Bretho Danzy, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2025. Promptaid: Visual prompt exploration, perturbation, testing and iteration for large language models. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–14.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada. Association for Computational Linguistics.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States. Association for Computational Linguistics.

Behnam Mohammadi. 2024. Explaining large language models decisions using shapley values. *SSRN*.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. 2024. A glitch in the matrix? locating and detecting language model grounding with fakepedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6828–6844, Bangkok, Thailand. Association for Computational Linguistics.

Kyle Moore, Jesse Roberts, Thao Pham, and Douglas Fisher. 2024. Reasoning beyond bias: A study on counterfactual prompting and chain of thought reasoning. *arXiv preprint arXiv:2408.08651*.

Yutao Mou, Yuxiao Luo, Shikun Zhang, and Wei Ye. 2025. Saro: Enhancing llm safety through reasoning-based alignment. *arXiv preprint arXiv:2504.09420*.

Maximilian Mozes, Tolga Bolukbasi, Ann Yuan, Frederick Liu, Nithum Thain, and Lucas Dixon. 2023. Gradient-based automated iterative recovery for parameter-efficient tuning. *Preprint*, arXiv:2302.06598.

Anish Mudide, Joshua Engels, Eric J Michaud, Max Tegmark, and Christian Schroeder de Witt. 2025. Efficient dictionary learning with switch sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Aashiq Muhamed, Mona T. Diab, and Virginia Smith. 2025. Decoding dark matter: Specialized sparse autoencoders for interpreting rare concepts in foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1604–1635, Albuquerque, New Mexico. Association for Computational Linguistics.

Neel Nanda. 2022. Neuroscope: A website for mechanistic interpretability of language models. https://neuroscope.io.

Neel Nanda. 2024. Attribution patching: Activation patching at industrial scale. https://www.neelnanda.io/mechanistic-interpretability/attribution-patching. Neel Nanda.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023a. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Neel Nanda, Senthooran Rajamanoharan, János Kramár, and Rohin Shah. 2023b. Fact finding: Attempting to reverse-engineer factual recall on the neuron level (post 1). https://www.lesswrong.com/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall. LessWrong.

nostalgebraist. 2020. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangde. 2024. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Charles O'Neill, Christine Ye, Kartheik Iyer, and John F Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 548–560, Singapore. Association for Computational Linguistics.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. 2025a. The hidden dimensions of llm alignment: A multi-dimensional safety analysis. *arXiv preprint arXiv:2502.09674*.

Yijun Pan, Taiwei Shi, Jieyu Zhao, and Jiaqi W Ma. 2025b. Detecting and filtering unsafe training data via data attribution. *arXiv preprint arXiv:2502.11411*.

Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. 2019. Sanvis: Visual analytics for understanding self-attention networks. In *2019 IEEE Visualization Conference (VIS)*, pages 146–150.

Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep S Lubana, and Hidenori Tanaka. 2024a. Emergence of hidden capabilities: Exploring learning dynamics in concept space. *Advances in Neural Information Processing Systems*, 37:84698–84729.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024b. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Mądry. 2023. Trak: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. 2024. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420.

Sheng Y Peng, Pin-Yu Chen, Matthew Hull, and Duen H Chau. 2024. Navigating the safety landscape: Measuring risks in finetuning large language models. *Advances in Neural Information Processing Systems*, 37:95692–95715.

Lavanya Prahallad and Radhika Mamidi. 2024. Significance of chain of thought in gender bias mitigation for english-dravidian machine translation. *arXiv preprint arXiv:2405.19701*.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024a. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *ICLR*.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024b. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *Proceedings of the 2024 International Conference on Learning Representations*. ArXiv:2402.14811.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, Miami, Florida, USA. Association for Computational Linguistics.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. 2024. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4864–4888, Bangkok, Thailand. Association for Computational Linguistics.

Hanhao Qu, Yu Cao, Jun Gao, Liang Ding, and Ruifeng Xu. 2022. Interpretable proof generation via iterative backward reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2968–2981, Seattle, United States. Association for Computational Linguistics.

Melissa Kazemi Rad, Huy Nghiem, Andy Luo, Sahil Wadhwa, Mohammad Sorower, and Stephen Rawls. 2025. Refining input guardrails: Enhancing llm-as-a-judge efficiency through chain-of-thought fine-tuning and alignment. *arXiv preprint arXiv:2501.13080*.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, Rohin Shah, and Neel Nanda. 2024a. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024b. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. 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.

Zhongzheng Ren, Raymond Yeh, and Alexander Schwing. 2020. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. *Advances in Neural Information Processing Systems*, 33:21786–21797.

Farnoush Rezaei Jafari, Grégoire Montavon, Klaus-Robert Müller, and Oliver Eberle. 2024. Mambalrp: Explaining selective state space sequence models. *Advances in Neural Information Processing Systems*, 37:118540–118570.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Nikta Gohari Sadr, Sangmitra Madhusudan, and Ali Emami. 2025. Think or step-by-step? unzipping the black box in zero-shot prompts. *arXiv preprint arXiv:2502.03418*.

Anastasiia V Sadybekov and Vsevolod Katritch. 2023. Computational approaches streamlining drug discovery. *Nature*, 616(7958):673–685.

Iqbal H Sarker. 2024. Llm potentiality and awareness: a position paper from the perspective of trustworthy and responsible ai modeling. *Discover Artificial Intelligence*, 4(1):40.

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations*.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. 2022. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8179–8186.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Lee Sharkey, Dan Braun, and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders. In *AI Alignment Forum*, volume 8, pages 15–16.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open problems in mechanistic interpretability. *Preprint*, arXiv:2501.16496.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. 2024. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. *arXiv preprint arXiv:2410.02298*.

Claudia Shi, Nicolas Beltran Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David Blei. 2024a. Hypothesis testing the circuit hypothesis in llms. *Advances in Neural Information Processing Systems*, 37:94539–94567.

Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, et al. 2024b. Large language model safety: A holistic survey. *arXiv preprint arXiv:2412.17686*.

Wei Shi, Sihang Li, Tao Liang, Mingyang Wan, Guojun Ma, Xiang Wang, and Xiangnan He. 2025. Route sparse autoencoder to interpret large language models. *arXiv preprint arXiv:2503.08200*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Anthony Sicilia and Malihe Alikhani. 2024. Eliciting uncertainty in chain-of-thought to mitigate bias against forecasting harmful user behaviors. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 211–223, Miami, Florida, USA. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

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

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024b. Representation surgery: theory and practice of affine steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Charlotte Siska and Anush Sankaran. 2025. Attention-defense: Leveraging system prompt attention for explainable defense against novel jailbreaks. *arXiv preprint arXiv:2504.12321*.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence*, 5(8):873–883.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 3607–3625, Singapore. Association for Computational Linguistics.

Linxin Song, Yan Cui, Ao Luo, Freddy Lecue, and Irene Li. 2024. Better explain transformers by illuminating important information. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2048–2062, St. Julian's, Malta. Association for Computational Linguistics.

Samuel Soo, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Ming YAN. 2025. Interpretable steering of large language models with feature guided activation additions. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M. Rush. 2019. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):353–363.

Jingbo Su. 2024. Enhancing adversarial attacks through chain of thought. *arXiv preprint arXiv:2410.21791*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024a. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*.

Yiheng Su, Junyi Jessy Li, and Matthew Lease. 2024b. Wrapper boxes for faithful attribution of model predictions to training data. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 551–576, Miami, Florida, US. Association for Computational Linguistics.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. 2025. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xue Tan, Hao Luan, Mingyu Luo, Xiaoyan Sun, Ping Chen, and Jun Dai. 2025. Revprag: Revealing poisoning attacks in retrieval-augmented generation through llm activation analysis. *arXiv preprint arXiv:2411.18948*.

Zhen Tan, Lu Cheng, Song Wang, Bo Yuan, Jundong Li, and Huan Liu. 2024. Interpreting pretrained language models via concept bottlenecks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 56–74. Springer.

Qiyao Tang and Xiangyang Li. 2025. An investigation of large language models and their vulnerabilities in spam detection. *arXiv preprint arXiv:2504.09776*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html. Transformer Circuits Thread.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Justin Theodorus, V Swaytha, Shivani Gautam, Adam Ward, Mahir Shah, Cole Blondin, and Kevin Zhu. 2025. Finding sparse autoencoder representations of errors in cot prompting. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. 2023. Sample based explanations via generalized representers. In *Proceedings of the 37th International Conference on Neural Information Processing*

Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yamei Tu, Jiayi Xu, and Han-Wei Shen. 2021. Keywordmap: Attention-based visual exploration for keyword analysis. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pages 206–215. IEEE.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vinod Veeramachaneni. 2025. Large language models: A comprehensive survey on architectures, applications, and challenges. *Advanced Innovations in Computer Programming Languages*, 7(1):20–39.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Huanqian Wang, Yang Yue, Rui Lu, Jingxin Shi, Andrew Zhao, Shenzhi Wang, Shiji Song, and Gao Huang. 2024a. Model surgery: Modulating llm's behavior via simple parameter editing. *arXiv preprint arXiv:2407.08770*.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. 2025a. Data shapley in one training run. In *The Thirteenth International Conference on Learning Representations*.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. 2024b. Helpful or harmful data? fine-tuning-free shapley attribution for explaining language model predictions. In *ICML*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024c. Detoxifying large language models via knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3093–3118, Bangkok, Thailand. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024d. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 89–104, Mexico City, Mexico. Association for Computational Linguistics.

Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2024e. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3(1):133.

Yi Wang, Fenghua Weng, Sibei Yang, Zhan Qin, Minlie Huang, and Wenjie Wang. 2025b. Delman: Dynamic defense against large language model jailbreaking with model editing. *arXiv preprint arXiv:2502.11647*.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024f. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*.

Yunchao Wang, Guodao Sun, Zihang Fu, and Ronghua Liang. 2025c. Human-computer interaction and visualization in natural language generation models: Applications, challenges, and opportunities. *Preprint*, arXiv:2410.08723.

Zijie J. Wang, Robert Turko, and Duen Horng Chau. 2021. Dodrio: Exploring transformer models with interactive visualization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 132–141, Online. Association for Computational Linguistics.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Winninger, Boussad Addad, and Katarzyna Kapusta. 2025. Using mechanistic interpretability to craft adversarial attacks against large language models. *Preprint*, arXiv:2503.06269.

Kangxi Wu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024a. Enhancing training data attribution for large language models with fitting error consideration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14131–14143, Miami, Florida, USA. Association for Computational Linguistics.

Xuansheng Wu, Jiayi Yuan, Wenlin Yao, Xiaoming Zhai, and Ninghao Liu. 2025. Interpreting and steering llms with mutual information-based explanations on sparse autoencoders. *arXiv preprint arXiv:2502.15576*.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. 2024b. Usable xai: 10 strategies towards exploiting explainability in the llm era. *Preprint*, arXiv:2403.08946.

Zhaoxuan Wu, Yao Shu, and Bryan Kian Hsiang Low. 2022. Davinz: Data valuation using deep neural networks at initialization. In *International Conference on Machine Learning*, pages 24150–24176. PMLR.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*.

Chen Xiong, Xiangyu Qi, Pin-Yu Chen, and Tsung-Yi Ho. 2024. Defensive prompt patch: A robust and interpretable defense of llms against jailbreak attacks. *arXiv preprint arXiv:2405.20099*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024a. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

Yang Xu, Yi Wang, and Hao Wang. 2024b. Tracking the feature dynamics in llm training: A mechanistic study. *arXiv preprint arXiv:2412.17626*.

Zhihao Xu, Ruixuan HUANG, Changyu Chen, and Xiting Wang. 2024c. Uncovering safety risks of large language models through concept activation vector. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jingyuan Yang, Dapeng Chen, Yajing Sun, Rongjun Li, Zhiyong Feng, and Wei Peng. 2024a. Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3343–3353, Bangkok, Thailand. Association for Computational Linguistics.

Sen Yang, Shujian Huang, Wei Zou, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2023a. Local interpretation of transformer based on linear decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10270–10287, Toronto, Canada. Association for Computational Linguistics.

Yi Yang, Hanyu Duan, Ahmed Abbasi, John P Lalor, and Kar Yan Tam. 2023b. Bias a-head? analyzing bias in transformer-based language model attention heads. *arXiv preprint arXiv:2311.10395*.

Zhou Yang, Zhensu Sun, Terry Zhuo Yue, Premkumar Devanbu, and David Lo. 2024b. Robustness, security, privacy, explainability, efficiency, and usability of large language models for code. *arXiv preprint arXiv:2403.07506*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. 2024. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):262–272.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Chih-Kuan Yeh, Ankur Taly, Mukund Sundararajan, Frederick Liu, and Pradeep Ravikumar. 2022. First is

better than last for language data influence. *Advances in Neural Information Processing Systems*, 35:32285–32298.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gal Yona, Or Honovich, Itay Laish, and Roee Aharoni. 2023. Surfacing biases in large language models using contrastive input decoding. *arXiv preprint arXiv:2305.07378*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024a. Natural language reasoning, a survey. *ACM Comput. Surv.*, 56(12).

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024b. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations*.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. 2024a. Prompt-guided internal states for hallucination detection of large language models. *arXiv preprint arXiv:2411.04847*.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024b. Tell your model where to attend: Post-hoc attention steering for LLMs. In *The Twelfth International Conference on Learning Representations*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024c. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949, Bangkok, Thailand. Association for Computational Linguistics.

Yuyou Zhang, Miao Li, William Han, Yihang Yao, Zhepeng Cen, and Ding Zhao. 2025. Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable llm safety. *arXiv preprint arXiv:2503.05021*.

Chongwen Zhao, Zhihao Dou, and Kaizhu Huang. 2024a. Eeg-defender: Defending against jailbreak through early exit generation of large language models. *arXiv preprint arXiv:2408.11308*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024b. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Haiyan Zhao, Fan Yang, Bo Shen, Himabindu Lakkaraju, and Mengnan Du. 2024c. Towards uncovering how large language model works: An explainability perspective. *arXiv preprint arXiv:2402.10688*.

Jiachen Zhao, Zonghai Yao, zhichao Yang, and hong yu. 2023a. SELF-EXPLAIN: Teaching large language models to reason complex questions by themselves. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023b. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

Wei Zhao, Zhe Li, Yige Li, Ye Zhang, and Jun Sun. 2024d. Defending large language models against jailbreak attacks via layer-specific editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5094–5109, Miami, Florida, USA. Association for Computational Linguistics.

Weixiang Zhao, Yulin Hu, Zhuojun Li, Yang Deng, Jiahe Guo, Xingyu Sui, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2024e. Towards comprehensive post safety alignment of large language models via safety patching. *Preprint*, arXiv:2405.13820.

Dylan Zhou, Kunal Patil, Yifan Sun, Karthik lakshmanan, Senthooran Rajamanoharan, and Arthur Conmy. 2025a. LLM neurosurgeon: Targeted knowledge removal in LLMs using sparse autoencoders. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Yilun Zhou and Julie Shah. 2023. The solvability of interpretability evaluation metrics. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2399–2415, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024. How alignment and jailbreak work: Explain LLM safety

through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2461–2488, Miami, Florida, USA. Association for Computational Linguistics.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025b. On the role of attention heads in large language model safety. In *The Thirteenth International Conference on Learning Representations*.

Minjun Zhu, Linyi Yang, Yifan Wei, Ningyu Zhang, and Yue Zhang. 2024. Locking down the finetuned llms safety. *arXiv preprint arXiv:2410.10343*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Appendix

Table 2 and Table 3 provide an overview of interpretation-informed safety enhancement techniques (§4) and tools that facilitate understanding and application of interpretation (§5). This table extends Table 1 in the main text to include safety-oriented interpretation methods not yet leveraged for safety enhancements or tool use. Each row is one work; each column corresponds to a technique or tool. Safety issues, techniques, and tools addressed by a work are indicated by a colored cell.

Table 2: Overview of representative works at the intersections of safety-focused interpretation (§3), safety enhancements they inform (§4), and tools operationalizing them (§5), extending Table 1. Continued in Table 3.

	SAFETY TYPE				§3 INTERPRET FOR SAFETY						§4 ENHANCE SAFETY						§5 PRACTICAL TOOLS					VENUE
Work	Hallucination	Jailbreak & Harm	Bias	Privacy Leakage	(3.1) Training Attrib.	(3.2) Input Token	(3.3.1) Probe Latent	(3.3.2) Perturb Comp	(3.3.3) Decipher Latent	(3.4) Self-Reason	(4.1) Attn. to Rel. Token	(4.2.1) Steer Latent Vec	(4.2.2) Modulate Neuron	(4.2.3) Edit Model	(4.3) Verify & Output	(4.4) Output w. Reason	Ease Impl.	(5.1) TDA Vis	(5.2) Token Vis	(5.3) Latent Vec Vis	(5.4) Neuron Vis	Publication
Su et al. (2024b)		■			■																	BlackboxNLP
He et al. (2024b)		■			■																	COLM
Pan et al. (2025b)		■	■		■																	ArXiv
Wu et al. (2024a)	■				■																	EMNLP
Chen et al. (2024a)		■			■																	ArXiv
Hazra et al. (2024)		■			■									■								EMNLP
Zhao et al. (2024e)		■			■									■								ArXiv
Lee et al. (2025a)	■				■													■				AAAI
Qian et al. (2024)	■	■	■	■	■		■					■										ACL
Lee et al. (2024)		■			■		■		■													ICML
Ferrando et al. (2022a)	■					■																EMNLP
Yuksekgonul et al. (2024)	■					■																ICLR
Mohammadi (2024)			■			■																SSRN
Cohen-Wang et al. (2024)		■				■																NeurIPS
Yona et al. (2023)			■			■																ArXiv
Eberle et al. (2023)			■			■																EMNLP
Vig (2019)			■			■													■			ACL
Sarti et al. (2023)			■			■													■			ACL
Mishra et al. (2025)			■			■													■			TVCG
Wang et al. (2025b)		■				■								■								ArXiv
Dale et al. (2023)	■					■									■							ACL
Chuang et al. (2024)	■					■									■							EMNLP
Pan et al. (2025a)		■				■	■				■											ArXiv
Zhang et al. (2024b)			■			■		■			■		■									ICLR
Li et al. (2023b)	■					■		■											■	■		ArXiv
Tenney et al. (2020)			■			■		■											■	■		EMNLP
Feng et al. (2024)		■				■			■													ICLR
Halawi et al. (2024)	■					■			■													ICLR
Liu et al. (2024a)	■						■															EMNLP
Arditi et al. (2024)		■					■															ArXiv
Lin et al. (2024)		■					■															EMNLP
Abdelnabi et al. (2024)		■					■															ArXiv
Slobodkin et al. (2023)	■						■															EMNLP
Ashok and May (2025)	■	■					■															ArXiv
Orgad et al. (2024)	■		■				■															ICLR
Li et al. (2025a)	■						■															ICASSP
Zhang et al. (2024a)	■						■															ArXiv
Su et al. (2024a)	■						■															ACL
Azaria and Mitchell (2023)	■						■															EMNLP
Ji et al. (2024b)	■						■															BlackboxNLP
Bürger et al. (2024)	■						■															NeurIPS
Zhou et al. (2024)		■					■															EMNLP
Xu et al. (2024c)		■					■															NeurIPS
He et al. (2024c)		■					■															ArXiv
Tan et al. (2025)	■						■															ArXiv
He et al. (2024a)	■						■															ACL
Beigi et al. (2024)	■						■															EMNLP
CH-Wang et al. (2024)	■						■															ACL
Winninger et al. (2025)		■					■															ArXiv
Arditi et al. (2024)		■					■															ArXiv
Li et al. (2024f)		■					■															ArXiv
Zhu et al. (2024)		■					■				■											ArXiv
Li et al. (2025c)		■					■					■										COLING
Duan et al. (2024)	■						■					■										ArXiv
Yang et al. (2024a)	■						■					■										ACL
Ball et al. (2024)		■					■					■										ArXiv
Wang et al. (2024a)		■					■					■										ArXiv
Bhattacharjee et al. (2024)		■					■					■										SafeGenAI
Chu et al. (2024)	■	■	■				■					■										CCS
Rimsky et al. (2024)	■	■					■					■										ACL
Singh et al. (2024b)		■	■				■					■										ICML
Zhang et al. (2024c)	■						■					■										ACL
Li et al. (2023a)	■						■					■										NeurIPS
Turner et al. (2023)		■					■					■										ArXiv
Gao et al. (2024)		■					■					■										ArXiv
Shen et al. (2024)		■					■					■										ICLR
Han et al. (2025)		■					■					■										ArXiv
Hernandez et al. (2024a)			■				■					■										COLM
Chen et al. (2024b)			■				■					■								■		ArXiv
Wang et al. (2024c)		■					■							■								ACL
Li et al. (2024e)		■					■							■								ArXiv
Chen et al. (2025)	■						■								■							AAAI
Zhao et al. (2024a)		■					■								■							ArXiv
Burns et al. (2022)	■						■								■							ArXiv
Li et al. (2024c)	■						■								■							ArXiv
Monea et al. (2024)	■							■														ACL
Vig et al. (2020)			■					■														NeurIPS
Wei et al. (2024)		■						■														ICML
Yang et al. (2023b)			■					■														ArXiv
Zhou et al. (2025b)		■						■														ICLR
Deng et al. (2025)	■							■					■									AAAI
Liu et al. (2024b)			■					■					■									ICLR
Li et al. (2024a)	■							■						■								ArXiv
Ma et al. (2023)			■					■						■								EMNLP
Li et al. (2025b)		■						■						■								ICLR
Jin et al. (2024)	■							■	■													ACL
Zhao et al. (2024d)		■						■	■					■								EMNLP

Table 3: Overview of representative works at the intersections of safety-focused interpretation (§3), safety enhancements they inform (§4), and tools operationalizing them (§5), extending Table 1 and continuing Table 2.

	SAFETY TYPE				§3 INTERPRET FOR SAFETY						§4 ENHANCE SAFETY						§5 PRACTICAL TOOLS					VENUE
Work	Hallucination	Jailbreak & Harm	Bias	Privacy Leakage	§3.1 Training Attrib.	§3.2 Input Token	§3.3.1 Probe Latent	§3.3.2 Perturb Comp	§3.3.3 Decipher Latent	§3.4 Self-Reason	§4.1 Attn. to Rel. Token	§4.2.1 Steer Latent Vec	§4.2.2 Modulate Neuron	§4.2.3 Edit Model	§4.3 Verify & Output	§4.4 Output w. Reason	Ease Impl.	§5.1 TDA Vis	§5.2 Token Vis	§5.3 Latent Vec Vis	§5.4 Neuron Vis	Publication
Ferrando et al. (2025)	■								■													ICLR
Theodorus et al. (2025)	■								■													ICLR
Muhamed et al. (2025)		■							■													NAACL
Härle et al. (2024)		■							■													ArXiv
Gallifant et al. (2025)		■							■													ArXiv
Marks et al. (2025)			■						■													ICLR
Jiang et al. (2024a)	■								■													NAACL
Hernandez et al. (2024b)			■						■											■		ICLR
Ameisen et al. (2025)									■												■	Anthropic
Lindsey et al. (2025)	■	■	■						■												■	Anthropic
Zhou et al. (2025a)			■						■				■									ICLR
Frikha et al. (2025)				■					■				■									ArXiv
Hegde (2024)			■						■				■									SciForDL
He et al. (2025)	■	■	■						■				■									ArXiv
Abdaljalil et al. (2025)	■								■				■									ArXiv
Bayat et al. (2025)	■								■				■									ArXiv
Wu et al. (2025)		■							■				■									ArXiv
O'Brien et al. (2024)		■							■				■									ArXiv
Geva et al. (2022a)	■		■						■				■							■		EMNLP
Khoriaty et al. (2025)		■							■				■									ArXiv
Geva et al. (2022b)		■							■				■									EMNLP
Yu et al. (2024b)	■								■					■								EMNLP
Lee et al. (2025b)	■									■												ArXiv
Akbar et al. (2024)	■									■												EMNLP
Li et al. (2024b)		■	■	■						■												ArXiv
Betley et al. (2025)		■		■						■												ICLR
Weng et al. (2023)	■									■					■							EMNLP
Cheng et al. (2025b)	■									■					■							ArXiv
Dhuliawala et al. (2024)	■									■					■							ACL
Liu et al. (2025a)		■								■						■						ArXiv
Jiang et al. (2025)	■									■						■						ICASSP
Ji et al. (2024a)	■									■						■						AAAI
Zhang et al. (2025)		■								■						■						ArXiv
Kaneko et al. (2024)			■							■						■						ArXiv
Prahallad and Mamidi (2024)			■							■						■						ArXiv
Li et al. (2024d)		■								■						■						ArXiv
Cao et al. (2024)		■								■						■						NaNA
Rad et al. (2025)		■								■						■						ArXiv
Moore et al. (2024)			■							■						■						ArXiv
Sicilia and Alikhani (2024)			■							■						■						NLP4PI
Mou et al. (2025)		■								■						■						ArXiv
Liu et al. (2025a)		■								■						■						ArXiv
Kwon and Mihindukulasooriya (2023)			■																	■		IUI