Locate, Steer, and Improve: A Practical Survey of Actionable Mechanistic Interpretability in Large Language Models

Hengyuan Zhang[1] †, Zhihao Zhang[2] †, Ercong Nie[3], Mingyang Wang[3], Zunhai Su[4], Yiwei Wang[5], Qianli Wang[6], Shuzhou Yuan[7], Xufeng Duan[8], Qibo Xue[9], Zeping Yu[10], Chenming Shang[11], Xiao Liang[12], Jing Xiong[1], Hui Shen[13], Chaofan Tao[1], Zhengwu Liu[1], Senjie Jin[2], Zhiheng Xi[2], Dongdong Zhang[14], Sophia Ananiadou[11], Tao Gui[2], Ruobing Xie[15], Hayden Kwok-Hay So[1], Hinrich Schütze[3], Xuanjing Huang[2], Qi Zhang[2] *, Ngai Wong[1] *

[1]The University of Hong Kong [2]Fudan University [3]LMU Munich [4]Tsinghua University
[5]Technische Universität Darmstadt [6]Technische Universität Berlin
[7]Technische Universität Dresden [8]The Chinese University of Hong Kong [9]Nanjing University
[10]University of Manchester [11]Dartmouth College [12]University of California, Los Angeles
[13]University of Michigan [14]Microsoft [15]Tencent
hengyuan.zhang88@gmail.com

Abstract

Mechanistic Interpretability (MI) has emerged as a vital approach to demystify the opaque decision-making of Large Language Models (LLMs). However, existing reviews primarily treat MI as an observational science, summarizing analytical insights while lacking a systematic framework for *actionable intervention*. To bridge this gap, we present a guide structured around the pipeline: *"Locate, Steer, and Improve."* We formally categorize *Localizing* (diagnosis) and *Steering* (intervention) methods based on specific *Interpretable Objects* to establish a rigorous intervention protocol. Furthermore, we demonstrate how this framework enables tangible improvements in *Alignment*, *Capability*, and *Efficiency*, effectively operationalizing MI as a practical engineering toolkit for model optimization. The curated paper list of this work is available at https://github.com/rattlesnakey/Awesome-Actionable-MI-Survey.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success (Abdin et al., 2024; OpenAI et al., 2024; Qwen et al., 2025), yet their internal decision-making processes remain largely opaque. This lack of transparency poses significant risks and limits our ability to efficiently optimize them. Mechanistic Interpretability (MI) has emerged as a pivotal approach aiming to reverse-engineer these models into understandable components. Given the rapid growth of this field, recent reviews have systematized the literature from distinct perspectives.

A body of work focuses on the theoretical and foundational aspects of MI (Räuker et al., 2023; Allen-Zhu and Li, 2023; Ferrando et al., 2024; Zhao et al., 2024a; Geiger et al., 2025; Gantla, 2025), providing technical roadmaps for dissecting Transformer architectures. They primarily prioritize discovery—understanding the model's inner working mechanisms—and treat MI as an observational science, leaving practical applications underexplored. Recognizing the practical potential of MI, another line of works has begun to bridge the gap between understanding and utilization, discussing how MI techniques can aid downstream tasks or specific domains (Luo and Specia, 2024; Wu et al., 2024a; Rai et al., 2024; Bereska and Gavves, 2024; Lee et al., 2025).

However, these reviews face two main limitations. First, they lack a *sufficient categorization and clear definition* of MI methods in practical application contexts. Second, their coverage of applications is often incomprehensive, and the illustration is typically too general, making it difficult for readers to translate mechanistic insights into actionable interventions. This gap is particularly critical at the current stage of AI development. While the rapid progress of LLMs has been predominantly driven by *external scaling factors*, e.g., larger models, more data, and increased computational resources, this paradigm for model improvement is increasingly encountering bottlenecks. We argue that further advancements must be refined through a deeper understanding of internal mechanisms, shifting the focus from simply scaling the "outside"

† Equal contribution.
* Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 10317–10362
July 2-7, 2026

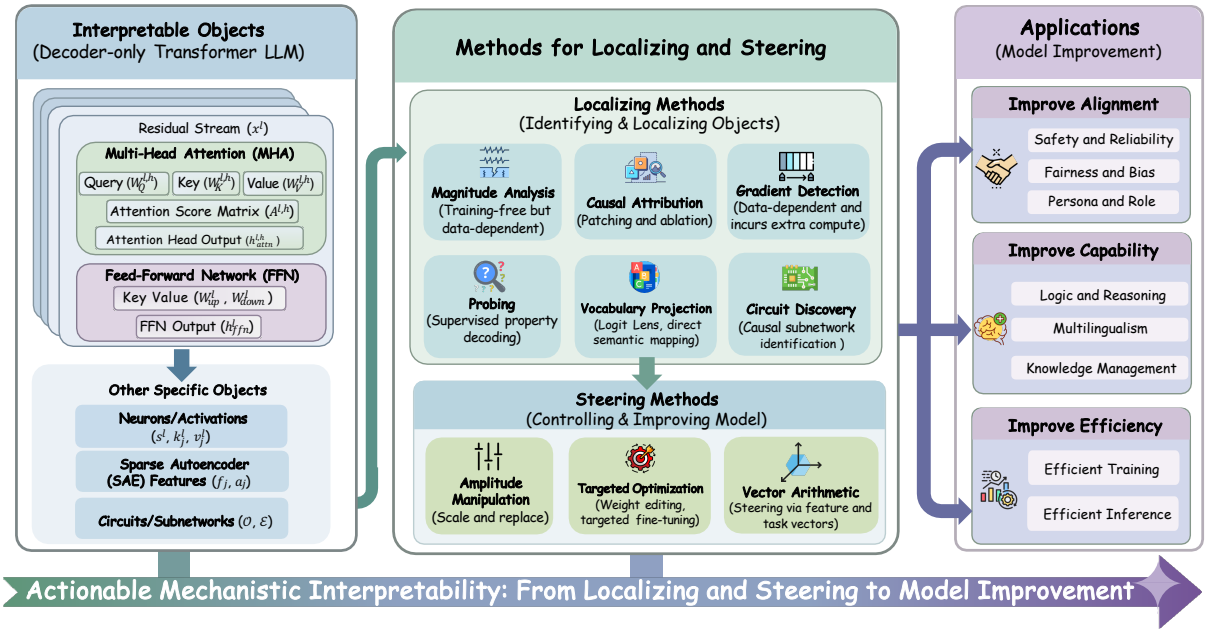

Figure 1: Overview of the paper structure. It progresses from interpretable objects (§2) to methodologies ranging from localizing (§3) to steering (§4), and finally illustrates applications for model improvement (§5). See Fig. 2 and 3 for the detailed outline.

to surgically improving the model from the "inside." However, the field currently lacks a unified guide that systematically *categorizes these internal methodologies* and presents a concrete *pipeline for active model improvement.*

To address these challenges, we propose the *"Locate, Steer, and Improve"* pipeline, which systematically transforms MI from passive observation to actionable engineering through the following contributions:

1) Pipeline-Driven Framework: We establish a rigorous framework for applying MI. We first define the **Interpretable Objects** within LLMs. Based on the application pipeline, we categorize methods into **Localizing** (diagnosis) and **Steering** (Intervention). Crucially, we provide the *Methodological Formulation*, *Applicable Objects and Scope* for each technique, helping readers quickly understand its actionable implementation.

2) Comprehensive Application Paradigms: We survey applications across three major themes: *Improve Alignment*, *Improve Capability*, and *Improve Efficiency*. Instead of providing a general overview, we summarize representative *MI paradigms* for each scenario. This approach allows readers to quickly capture the distinct usage patterns of MI in different application contexts.

3) Insights and Resources: We discuss challenges and future directions of applied MI research, providing the community with a curated, method-tagged collection of over 200 papers in Tab. 3.

2 Interpretable Objects of LLMs

Here, we formulate the interpretable objects, with a focus on the decoder-only architecture—the predominant framework for contemporary LLMs.

Token Embedding The model entry point maps discrete tokens to continuous vectors. We define the *Embedding Matrix* as $\mathbf{W}_E \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$, where $|\mathcal{V}|$ denotes the vocabulary size and d_{model} is the hidden dimension. For a given token t_i, its *Token Embedding* (or the initial residual stream state) is obtained by $\mathbf{x}_i^0 = \mathbf{W}_E[t_i] + \mathbf{p}_i$, where $\mathbf{p}_i$ denotes the positional embedding.[1]

Transformer Block Typically, an LLM is composed of L stacked layers, each consisting of an MHA and an FFN block. The primary communication channel is *Residual Stream State*. Its update dynamics at layer l ($\mathbf{x}^l$) are defined as[2]:

$$\mathbf{x}^{l,\text{mid}} = \mathbf{x}^l + \mathbf{h}_{\text{attn}}^l(\mathbf{x}^l) \quad (1)$$

$$\mathbf{x}^{l+1} = \mathbf{x}^{l,\text{mid}} + \mathbf{h}_{\text{ffn}}^l(\mathbf{x}^{l,\text{mid}}) \quad (2)$$

where $\mathbf{x}^{l,\text{mid}}$ is the intermediate state after the MHA block. This additive structure is crucial for MI

[1] Modern LLMs (e.g., Llama) typically apply Rotary Positional Embeddings (RoPE) directly to queries and keys.

[2] For simplicity, we omit normalization from the equations, as they are often ignored in high-level MI analyses.

analysis, enabling the decomposition of the final prediction into individual component contributions. See Appendix D.1 for the details of stream flow.

Multi-Head Attention (MHA) MHA allows tokens to contextualize information from other positions via H independent heads. For a head h, we define the learnable *Weight Matrices* as $\mathbf{W}_Q^{l,h}, \mathbf{W}_K^{l,h}, \mathbf{W}_V^{l,h} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$, and the output projection as $\mathbf{W}_O^{l,h} \in \mathbb{R}^{d_{\text{head}} \times d_{\text{model}}}$. The *Attention Score* $\mathbf{A}^{l,h} \in \mathbb{R}^{N \times N}$ is defined as:

$$\mathbf{A}^{l,h} = \text{softmax}\left(\frac{(\mathbf{x}^l \mathbf{W}_Q^{l,h})(\mathbf{x}^l \mathbf{W}_K^{l,h})^\top}{\sqrt{d_{\text{head}}}}\right) \quad (3)$$

The head output is $\mathbf{h}_{\text{attn}}^{l,h} = \mathbf{A}^{l,h}(\mathbf{x}^l \mathbf{W}_V^{l,h})\mathbf{W}_O^{l,h}$, and the total block output is the sum over all heads $\mathbf{h}_{\text{attn}}^l = \sum_{h=1}^{H} \mathbf{h}_{\text{attn}}^{l,h}$.

Feed-Forward Network (FFN) FFNs act as position-wise feature transformers:

$$\mathbf{h}_{\text{ffn}}^l = \sigma(\mathbf{x}^{l,\text{mid}} \mathbf{W}_{\text{up}}^l)\mathbf{W}_{\text{down}}^l \quad (4)$$

where $\mathbf{x}^{l,\text{mid}}$ is the input to FFN, $\mathbf{h}_{\text{ffn}}^l$ is output, σ is activation function.[3] The $\mathbf{W}_{\text{up}}^l \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ and $\mathbf{W}_{\text{down}}^l \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$ of FFN are viewed as a Key-Value dictionary (Geva et al., 2021, 2022). The **Neuron** j is defined as an atomic unit comprised of a pair of weights: *key weight* $\mathbf{k}_j^l$ (the j-th row of the $\mathbf{W}_{\text{up}}^l$) and *value weight* $\mathbf{v}_j^l$ (the j-th column of $\mathbf{W}_{\text{down}}^l$). The intermediate $\mathbf{s}^l = \sigma(\mathbf{x}^{l,\text{mid}} \mathbf{W}_{\text{up}}^l)$ denotes the *Neuron Activation*.

Sparse Autoencoder (SAE) Feature While individual internal objects (e.g., neuron activation $\mathbf{s}^l$, or residual stream state $\mathbf{x}^l$) in LLMs are often *polysemantic* due to the phenomenon of *superposition* (Elhage et al., 2022), SAEs provide a principled method to disentangle these representations into *monosemantic features* (Bricken et al., 2023).

SAEs are trained in a layer-wise manner. For instance, when applying an SAE to reconstruct $\mathbf{x}^l$, the forward pass is defined as:

$$\mathbf{a} = \sigma(\mathbf{x}^l \mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}}), \quad \hat{\mathbf{x}}^l = \mathbf{a}\mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}} \quad (5)$$

where $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{SAE}}}$ and $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{d_{\text{SAE}} \times d_{\text{model}}}$ are learnable weights ($d_{\text{SAE}} \gg d_{\text{model}}$). The objective minimizes reconstruction error with a sparsity penalty: $\mathcal{L} = \|\mathbf{x}^l - \hat{\mathbf{x}}^l\|_2^2 + \lambda\|\mathbf{a}\|_1$. Here,

[3] Modern LLMs often use SwiGLU with a gating matrix $\mathbf{W}_{\text{gate}}^l$. For simplicity, we present the standard FFN here.

Object		Notation
Token Embedding	Embedding Matrix	$\mathbf{W}_E$
	Token i Embedding (Input)	$\mathbf{x}_i^0$
Residual Stream	Residual Stream State	$\mathbf{x}^l$
	Intermediate State (Post-Attn)	$\mathbf{x}^{l,\text{mid}}$
Attention	Q, K, V, O Weight Matrices	$\mathbf{W}_Q^{l,h}, \mathbf{W}_K^{l,h}, \mathbf{W}_V^{l,h}, \mathbf{W}_O^{l,h}$
	Attention Score Matrix	$\mathbf{A}^{l,h}$
	Head Output	$\mathbf{h}_{\text{attn}}^{l,h}$
	MHA Block Output	$\mathbf{h}_{\text{attn}}^l$
FFN	Up Projection (Key) Matrix	$\mathbf{W}_{\text{up}}^l$
	Down Projection (Value) Matrix	$\mathbf{W}_{\text{down}}^l$
	FFN Block Output	$\mathbf{h}_{\text{ffn}}^l$
Neuron	Neuron Activation	$\mathbf{s}^l$
	j-th Neuron Key Weight	$\mathbf{k}_j^l$
	j-th Neuron Value Weight	$\mathbf{v}_j^l$
SAE Feature	j-th Feature	$\mathbf{f}_j$
	j-th Feature Activation	a_j

Table 1: The interpretable objects of LLMs and their mathematical notations in this paper.

the **SAE Feature** $\mathbf{f}_j$ (j-th row of $\mathbf{W}_{\text{dec}}$) represents a distinct semantic direction, while the **SAE Feature Activation** a_j (j-th element of $\mathbf{a}$) quantifies its activation strength in the input. See Appendix D.2 for more details about the SAE training procedure.

3 Localizing Methods

Localizing Methods aim to identify interpretable objects responsible for specific behaviors, laying the foundation for subsequent analysis and steering.

3.1 Magnitude Analysis

Methodological Formulation *Magnitude Analysis* scores objects $\{o_j\}$ (e.g., weights or activations) via a scalar function $s_j = f(o_j)$ to identify salient candidates. Common metrics include the L_p-norm ($s_j = \|o_j\|_p$) or max-value ($s_j = \max_k |(o_j)_k|$). High-scoring elements (e.g., $\arg\text{topk}_j\, s_j$) are selected for further inspection or intervention.

Applicable Objects and Scope This method applies broadly to both static parameters and dynamic representations. 1) For parameters, per-weight or per-row norms are widely used to highlight outliers (Zhang et al., 2025a; An et al., 2025; Su and Yuan, 2025). 2) For the latter, ranking Neurons or SAE Features by activation norms helps localize specialized features (Huben et al., 2024; Tang et al., 2024b; Su et al., 2025d), while vector norms of attention outputs ($\|\mathbf{h}_{attn}^{l,h}\|$) identify influential heads (Jin et al., 2025a; Bi et al., 2025). Furthermore, measuring *layer-wise distances* reveals structural roles: comparing representations across contrastive inputs (e.g., $\|\mathbf{x}^l - \mathbf{x}'^l\|$) localizes layers where task-specific information diverges most strongly (Chuang et al., 2024; Zhang

et al., 2024c), whereas comparing consecutive layers (e.g., $\|\mathbf{x}^l - \mathbf{x}^{l+1}\|$) identifies redundant computation (Elhoushi et al., 2024; Men et al., 2025).

The scope of *Magnitude Analysis* is characterized as **training-free but data-dependent**: it avoids training costs but relies on input data to filter salient objects for subsequent analysis.

3.2 Causal Attribution

Methodological Formulation *Causal Attribution* identifies objects *causally responsible* for behavior by measuring the effect of interventions (Vig et al., 2020). Let $F(\cdot)$ denote a scalar model output of interest (e.g., a target token logit), and let o be an internal object defined in §2. The causal effect of object o is defined as $\Delta F(o) = F(\mathrm{do}(o \leftarrow \tilde{o})) - F(o)$, where the intervention $\mathrm{do}(\cdot)$ typically takes the form of *Patching* (swapping with a counterfactual state) or *Ablation* (setting to zero). $\tilde{o}$ represents the modified state. Large $|\Delta F(o)|$ indicates that o is a critical mediator of the behavior.

Applicable Objects and Scope This analysis primarily targets **dynamic objects** (e.g., $\mathbf{x}^l$, $\mathbf{h}^l_{\text{ffn}}$, $\mathbf{h}^{l,h}_{\text{attn}}$) and circuit edges. **Patching**-based attribution replaces an object computed from the original input with one computed from a *counterfactual input* to isolate specific information pathways. By systematically patching across layers, one can localize where task-specific information (e.g., factual knowledge) is introduced (Meng et al., 2022, 2023; Yeo et al., 2025b; Ravindran, 2025). See Appendix E.3 for more details. **Ablation**-based attribution explicitly "zeros out" objects—such as specific heads $\mathbf{h}^{l,h}_{attn}$ (Geva et al., 2023) or circuit edges (Yao et al., 2024a)—and measures the resulting performance drop to determine their necessity.

While *Causal Attribution* offers rigorous **causal responsibility**, it is computationally expensive (scaling linearly with object count). This inefficiency necessitates *Gradient Detection* (§3.3) as a rapid approximation for initial screening.

3.3 Gradient Detection

Methodological Formulation *Gradient Detection* scores an object o_j via the local sensitivity of a scalar target $F(x)$ (e.g., a logit or loss) to that object: $s_j(x) = \phi(\nabla_{o_j} F(x), o_j)$. Common instantiations include $s_j = \|\nabla_{o_j} F(x)\|$ and $s_j = \nabla_{o_j} F(x)^\top o_j$ (Sundararajan et al., 2017). The latter is motivated by the first-order Taylor approximation $F(o_j + \Delta o_j) \approx F(o_j) + \nabla_{o_j} F(x)^\top \Delta o_j$. See Appendix E.4 for more details.

Applicable Objects and Scope Since F is differentiable w.r.t. any object o_j in Tab. 1, *Gradient Detection* applies to inputs, activations, and parameters. For input embeddings ($\mathbf{x}^0_i$) and residual states ($\mathbf{x}^l$), it ranks influential tokens/positions/layers (Nguyen et al., 2025a). For Neuron objects (e.g., $\mathbf{s}^l$, $\mathbf{v}^l_j$), it localizes knowledge- or context-sensitive neurons and traces upstream dependencies (Dai et al., 2022; Shi et al., 2024). The same principle extends to Attention outputs $\mathbf{h}^{l,h}_{attn}$ and parameter blocks (e.g., $\mathbf{W}^{l,h}_{Q/K/V/O}$, $\mathbf{W}^l_{\text{up}}$, $\mathbf{W}^l_{\text{down}}$), where gradients rank salient attention/FFN modules (Jafari et al., 2025; Azarkhalili and Libbrecht, 2025) and identify individual parameters (Li et al., 2025b) or block granularity (Zhang et al., 2024e).

Gradient Detection is **data-dependent** and incurs **extra compute** from backpropagation; it is therefore often used as a lightweight ranking step before *Causal Attribution* (§3.2). In addition, pairing F with *Causal Attribution* yields a contrastive signal via the counterfactual logit gap $|\mathrm{logit}(\mathbf{x}) - \mathrm{logit}(\mathbf{x}^{cf})|$ (Yin and Neubig, 2022).

3.4 Probing

Methodological Formulation *Probing* trains a predictor g_ψ to decode a label y from an internal vector $\mathbf{z} \in \mathbb{R}^{d_{\text{model}}}$ (e.g., $\mathbf{x}^l$), e.g., $\hat{y} = g_\psi(\mathbf{z}) = \mathrm{softmax}(\mathbf{W}_P \mathbf{z})$ on $\mathcal{D} = \{(\mathbf{z}, y)\}$ (Belinkov, 2022). For *object detection*, probes are trained on candidate objects (layers/heads/FFNs) and ranked by decoding performance, typically followed by *Causal Attribution* for validation.

Applicable Objects and Scope *Probing* is defined on internal vectors/activations, and is commonly applied to the Residual Stream ($\mathbf{x}^l$, $\mathbf{x}^{l,\text{mid}}$), Attention outputs ($\mathbf{h}^{l,h}_{attn}$), FFN outputs ($\mathbf{h}^l_{ffn}$), and Neuron activations ($\mathbf{s}^l$). Recent LLM studies use layer-wise probes to track where contextual knowledge and conflict signals become decodable in $\mathbf{x}^l$ (Ju et al., 2024; Zhao et al., 2024c), and head-/module-level probes to localize capabilities or risks by comparing decoding strength across $\mathbf{h}^{l,h}_{attn}$ and module outputs (Du et al., 2024; Zhao et al., 2025b; Kim et al., 2025b). *Probing* also extends to SAE Feature by training probes on feature activations a_j, enabling concept decoding and interpretable SAE analysis (Kantamneni et al., 2025; Chanin et al., 2024).

The scope of *Probing* is **supervised decoding**:

it measures how well y is predictable from an internal object, requiring **labeled data** and **additional training**; in practice it is often used for object ranking and paired with *Causal Attribution* (§3.2) to test functional necessity.

3.5 Vocabulary Projection

Methodological Formulation The canonical technique in this category is the *Logit Lens* (nostalgebraist, 2020), which treats the unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{d_{\text{model}} \times |\mathcal{V}|}$ as a universal decoder for intermediate states. For a generic object $\mathbf{z} \in \mathbb{R}^{d_{\text{model}}}$ (e.g., the residual stream state $\mathbf{x}^l$), it computes a vocabulary distribution $\mathbf{p} = \text{softmax}(\mathbf{z}\mathbf{W}_U)$.

Applicable Objects and Scope This method can be applied to multiple objects in §2. Applied to $\mathbf{x}^l$, it traces layer-wise prediction evolution and highlights *crucial layers* where traits emerge (Wendler et al., 2024). Applied to $\mathbf{h}^{l,h}_{\text{attn}}$, it exposes the information a head injects into the residual stream, enabling characterization of functional heads such as "Name Mover Head" (Wang et al., 2023b) and "Copy Suppression Heads" (McDougall et al., 2024). Similarly, projecting Neuron Value Weights (Geva et al., 2021) or SAE Features (Bricken et al., 2023) supports interpreting their semantics via the tokens they promote in vocabulary space. See Appendix E.1 for more details.

The scope of this method is **direct semantic mapping**. Unlike *Probing* (§3.4), it provides a zero-shot view of internal representations without auxiliary training.

3.6 Circuit Discovery

Methodological Formulation Circuit discovery seeks *mechanistic pathways*: structured, directed dependencies among internal objects that implement a target behavior (Hanna et al., 2023). Let $(\mathcal{O}, \mathcal{E})$ be the computational graph over objects $\mathcal{O}$ and directed edges $\mathcal{E}$, where $e_{ij} \in \mathcal{E}$ denotes signal flow from o_i to o_j. A circuit $\mathcal{C} \subseteq \mathcal{E}$ is *faithful* if restricting computation to $\mathcal{C}$ (e.g., by patching/ablating all other edges) preserves the target output $F(x)$ or task performance.

Applicable Objects and Scope Circuit discovery targets *edges between objects*, ranging over directed dependencies among interpretable objects in Tab. 1 (e.g., between $\mathbf{x}^l$, $\mathbf{h}^{l,h}_{attn}$ and $\mathbf{h}^l_{ffn}$, among Attention/FFN modules, or among SAE Features $\mathbf{f}_j$). In practice, it is operationalized via (i) *intervention*-based search for causally necessary edges (activation patching/ablation search) (Conmy et al., 2023; Stolfo et al., 2023; Wang et al., 2023b), (ii) *attribution*-based edge scoring that ranks edges by estimated influence (e.g., gradient-based attributions and position-aware refinements) (Hanna et al., 2024; Haklay et al., 2025; Mueller et al., 2025), or (iii) *feature*-based replacement models (e.g., SAE/transcoder variants) that lift *Circuit Discovery* to sparse features, enabling attribution graphs and prompt-specific tracing in SAE feature space (Bricken et al., 2023; Ameisen et al., 2025; Hanna et al., 2025). See Appendix E.5 for more details.

The scope of *Circuit Discovery* is **causal subgraph identification**: extracting a minimally sufficient subnetwork **across layers** that mediates a behavior. It is closely coupled with *Causal Attribution* (§3.2) for validation, and often uses gradient-based signals to propose candidate edges.

4 Steering Methods

While localization methods (§3) identify the specific objects responsible for model behaviors, this section focuses on techniques to manipulate these objects to *steer* model outputs.

4.1 Amplitude Manipulation

Methodological Formulation *Amplitude Manipulation* steers behavior by modifying the magnitude of a target object o to $\tilde{o}$. This involves either **Ablation/Patching** where $\tilde{o} \in \{0, \mathbb{E}[o], o_{\text{tgt}}\}$ (zeroing, mean, or target replacement) to suppress or alter information flow through o, or **Scaling**, where $\tilde{o} = \alpha \cdot o$, adjusting strength via coefficient α to amplify or attenuate their downstream influence during forward pass. While ablation and patching here is operationally identical to that in §3.2, the objective differs: attribution employs them to *diagnose* causality, whereas steering targets pre-localized objects to *intervene* on behavior.

Applicable Objects and Scope **Ablation and Patching** typically target object outputs (e.g., $\mathbf{x}^l$, $\mathbf{h}^l_{ffn}$, and $\mathbf{h}^{l,h}_{\text{attn}}$) or activations ($\mathbf{s}^l$, a_j) and circuit edges. This eliminates or swaps their influence during the forward pass, thereby mitigating unwanted behaviors such as language or knowledge confusion (Nie et al., 2025; Huang et al., 2025a; Niu et al., 2025) or reducing computation to accelerate inference (Liu et al., 2024b; Men et al., 2025). **Scaling** predominantly targets $\mathbf{s}^l$, a_j or steering vectors $\mathbf{v}$ within *Vector Arithmetic* (§4.3) to adjust their contributions, enabling fine-grained control

over specific model attributes (Stoehr et al., 2024; Wang et al., 2025a; Pach et al., 2025; Liu et al., 2025c).

This method is characterized as **inference-time activation control**. It provides a training-free mechanism to modulate model behavior by directly adjusting the activation amplitude of target objects.

4.2 Targeted Optimization

Methodological Formulation *Targeted Optimization* permanently updates a weight subset w to alter behavior. Given a steering dataset $\mathcal{D}_{\text{steer}}$, it minimizes a task-specific loss $\mathcal{L}$ via $w_{new} \leftarrow w - \eta \cdot \nabla_w \mathcal{L}(\mathcal{D}_{\text{steer}})$. Crucially, focused adaptation is enforced by freezing remaining weights $\theta_{\setminus w}$ or assigning w a higher learning rate η.

Applicable Objects and Scope The subset of weights w is often distributed across multiple layers. The weights within w generally operate at two granularity: **1) Fine-grained Objects:** This category comprises neuron's key/value weights ($\mathbf{k}_j^l/\mathbf{v}_j^l$) or specific weight matrices (e.g., FFN's $\mathbf{W}_{\text{down}}^l$). Interventions at this level are motivated by empirical findings that these weights densely encode specific factual knowledge and functional capabilities (Meng et al., 2022, 2023; Zhang et al., 2024a; Zhu et al., 2024). **2) Coarse-grained Objects:** Optimization can also target all weight matrices within a specific functional component (e.g., a specific Attention Head, FFN, or an entire Transformer Block) (Zhang et al., 2024d; Li et al., 2025d). Updating entire components effectively adapts general processing mechanisms, necessary for adjusting broad capabilities that involve complex behaviors.

This method is characterized by **persistence** and **surgical precision**. Unlike *Amplitude Manipulation* (§4.1), it involves gradient-based training to rewrite parameters, which enables precise memory editing and focused capability enhancement while minimizing collateral damage to unrelated traits.

4.3 Vector Arithmetic

Methodological Formulation Positing that concepts or skills are encoded linearly, this method steers a target object $\mathbf{z}$ (hidden states or paramters) via $\hat{\mathbf{z}} = \mathbf{z} + \alpha \cdot \mathbf{v}$. Here, $\mathbf{v}$ is the *steering vector*, which denotes the *direction of target attribute* and α controls the intervention strength.

Applicable Objects and Scope The target object $\mathbf{z}$ can be either dynamic hidden states or static parameters. **1) Hidden States:** The primary targets are $\mathbf{x}^l$ and $\mathbf{h}_{\text{attn}}^{l,h}$). For these, $\mathbf{v}$ is typically derived from *contrastive activation means* (calculating the difference between average activations of counterfactual inputs (Rimsky et al., 2024)) or extracted **SAE features** that are corresponding to target concepts (Shu et al., 2025). See Appendix E.2 for more details. **2) Parameters:** For static weights, the steering vector $\mathbf{v}$ is explicitly defined as a *Task Vector* in Model Merging (Liu et al., 2025c). This vector is computed as the element-wise difference between a fine-tuned model and its pre-trained base, encapsulating a transferable skill or behavior.

The scope of this method is characterized by **additive directionality**. Unlike *Targeted Optimization* (§4.2), *Vector Arithmetic* acts as a steering force, dynamically pushing the model towards a target attribute without permanently altering weights.

5 Applications

5.1 Improve Alignment

5.1.1 Safety and Reliability

Summary. MI enhances model safety and reliability primarily via two steering paradigms: 1) Component-level Manipulation, which targets safety-critical neurons or structures via *Amplitude Manipulation* (§4.1) and *Targeted Optimization* (§4.2); 2) *Vector Arithmetic* (§4.3), which modulates high-level representations of *truthfulness*, *refusal*, or *instruction following*.

1) Works adopting the first paradigm focus on localizing and intervening on specific safety-related objects. Researchers identify specialized attention heads (Zhou et al., 2025), neurons (Chen et al., 2025b; Suau et al., 2024; Gao et al., 2025a), and SAE features (Templeton et al., 2024; Goyal et al., 2025; Yeo et al., 2025a) that encode toxicity, harmfulness, or hallucination. Once localized, mitigation is typically achieved via *Amplitude Manipulation* to ablate or downscale these activations during the forward pass. Complementarily, *Targeted Optimization* offers a persistent solution by updating parameters within identified safety circuits (Huang et al., 2025a), neurons (Zhao et al., 2025d; Chen et al., 2025b; Li et al., 2024), layer (Li et al., 2025d), or vector levels (Lee et al., 2024).

2) The second paradigm operates by steering latent space representations via *Vector Arithmetic*. Studies have established that high-level concepts such as truthfulness, refusal, and instruction following are mediated by specific directions in the

latent space (Zhao et al., 2025c; Yin et al., 2025; Wang et al., 2025f,g; Huang et al., 2025b). Based on this insight, targeted interventions inject steering vectors to inhibit hallucinations, correct failed refusals (Chuang et al., 2024; Chen et al., 2024a; Zhang et al., 2024c; Orgad et al., 2025), or improve instruction following (He et al., 2025b; Stolfo et al., 2025; Jiang et al., 2024c; Li et al., 2025c). This approach effectively bolsters safety and reliability while preserving the model's general capabilities.

5.1.2 Fairness and Bias

Summary. MI improves model fairness and mitigates biases predominantly through a two-step paradigm: utilizing *Causal Attribution* (§3.2) and *Magnitude Analysis* (§3.1) to identify bias-carrying objects, followed by *Amplitude Manipulation* (§4.1) or *Targeted Optimization* (§4.2) to suppress their detrimental effects.

Strategies addressing *data-centric* biases (e.g., gender, demographic, and cultural biases) typically employ *Causal Attribution* techniques, such as zero ablation and activation patching, to trace how biased signals are mediated by specific subsets of attention heads, FFNs, or residual stream states (Vig et al., 2020; Cai et al., 2024a; Ahsan et al., 2025). Some studies extend this approach to broader bias categories by first scoring internal structures (e.g., circuit edges) via *Magnitude Analysis* and subsequently validating them through *Causal Attribution* (Chandna et al., 2025; Yu et al., 2025a; Kim et al., 2025b). Once these bias-critical units are localized, mitigation is achieved either through inference-time *Amplitude Manipulation* (suppressing or down-weighting activations (Vig et al., 2020; Liu et al., 2024c; Chandna et al., 2025; Guan et al., 2025)) or through *Targeted Optimization*, which persistently updates the parameters of the identified subsets (Chintam et al., 2023; Cai et al., 2024a; Yu and Ananiadou, 2025).

5.1.3 Persona and Role

Summary. MI facilitates the control of LLM personas and roles through two steering paradigms: 1) *Vector Arithmetic* (§4.3), which modulates global persona representations, and 2) Component-level Manipulation, which targets persona-specific neurons or weights via *Amplitude Manipulation* (§4.1) and *Targeted Optimization* (§4.2).

1) Works adopting the first paradigm typically construct counterfactual inputs with opposing personas to extract steering vectors. These vectors are then applied via *Vector Arithmetic* to the *Residual Stream State* to amplify or suppress specific attributes during the forward pass, effectively controlling sycophancy, role-playing capabilities, and character traits (Rimsky et al., 2024; Potertì et al., 2025; Pai et al., 2025; Chen et al., 2025d; Handa et al., 2025; Lu et al., 2026).

2) The second paradigm focuses on pinpointing precise model components responsible for specific values or personalities. Once identified, researchers apply *Amplitude Manipulation* to selectively activate or suppress these neurons to alter the model's persona (Su et al., 2025b; Deng et al., 2025). Complementarily, *Targeted Optimization* is employed to fine-tune these identified persona-specific components, embedding objective-aligned behaviors persistently into the model (Chen et al., 2024b).

Beyond active steering, MI techniques are further applied to evaluate and analyze psychological traits in LLMs. For instance, *Probing* is utilized to locate persona-specific representations (Tak et al., 2025; Yuan et al., 2025; Ju et al., 2025), while measuring the similarity between internal states and persona vectors serves as a metric for assessing the model's alignment with specific psychological profiles (Karny et al., 2025; Banayeeanzade et al., 2025; Bas and Novak, 2025).

5.2 Improve Capability

5.2.1 Multilingualism

Summary. MI enables targeted control and enhancement of language behavior in LLMs mainly through two steering paradigms: 1) *Amplitude Manipulation* (§4.1), which targets language-specific features; 2) *Vector Arithmetic* (§4.3), which shifts representations along language-related directions.

1) Works adopting the first paradigm leverage language data to pinpoint language-specific neurons (Zhao et al., 2024b; Gurgurov et al., 2025a; Tang et al., 2024b; Liu et al., 2025d; Jing et al., 2025) or SAE features (Andrylie et al., 2025; Brinkmann et al., 2025) via *Magnitude Analysis* on their activations. Interventions typically involve zeroing or scaling these activations to enhance multilingual performance and control output language.

2) The second paradigm typically operates by

steering the residual stream state via *Vector Arithmetic*. To determine the optimal intervention site, works often employ *Vocab Projection* (Wendler et al., 2024; Zhao et al., 2024b) and *Magnitude Analysis* (Philippy et al., 2023; Mousi et al., 2024) to identify crucial layers—often located in the late layers of the model. Applying steering vectors in these layers effectively projects representations along language-related directions, enabling better multilingual understanding (Chi et al., 2023; Hinck et al., 2024; Zhang et al., 2025d) and mitigating language inconsistency issue (Wang et al., 2025c,d; Nie et al., 2025; Liu et al., 2025e). Regarding the construction of steering vectors, while most works derive them from parallel language data, recent studies also explore constructing vectors directly from language-specific neurons (Gurgurov et al., 2025a) or SAE features (Andrylie et al., 2025).

5.2.2 Knowledge Management

Summary. MI supports knowledge management mainly through three paradigms: 1) Precise knowledge *updating* and memory rewriting; 2) Knowledge *retention* mitigates interference under continual updates; 3) Knowledge *consolidation* combines skills/models. Across them, interventions use *Targeted Optimization* (§4.2), *Amplitude Manipulation* (§4.1), and *Vector Arithmetic* (§4.3).

1) Knowledge updating localizes editable carriers via *Causal Attribution* / *Probing* (optionally aided by *Gradient Detection* or *Vocabulary Projection*) and performs persistent rewrites with *Targeted Optimization* (Meng et al., 2022, 2023; Chen et al., 2025f,c; Zhang et al., 2025f; Katz et al., 2024). Alternatively, activation-level updates use *Amplitude Manipulation*: Lai et al. (2025) edits components through gated activation control, while SAE-based methods intervene in activation space for unlearning/detoxification by identifying relevant features via *Magnitude Analysis* (Muhamed et al., 2025a).

2) Retention identifies interference circuits via *Circuit Discovery* or *Causal Attribution*, then applies Head-/FFN-level *Amplitude Manipulation* to suppress conflict sources (e.g., pruning or controlled attention) (Jin et al., 2024; Li et al., 2025a; Niu et al., 2025; Jin et al., 2025a). To limit drift, continual adaptation restricts *Targeted Optimization* to selected modules or uses frozen-backbone representation edits (Zhang et al., 2024e,a; Wu et al., 2024b; Yang et al., 2025b; Wang et al., 2025e; Yang et al., 2025a), while *Probing* diagnostics help forecast and evaluate side effects (Du et al., 2024).

3) Consolidation combines skills/models by using lightweight *Probing* to identify compatible carriers (e.g., aligned parameter subspaces or transferable feature bases), then persistently composes them via *Vector Arithmetic*, aggregating complementary skills while mitigating interference (Yadav et al., 2023; Liu et al., 2025c; Chen et al., 2025a; Zhao et al., 2025e; Li et al., 2026).

5.2.3 Logic and Reasoning

Summary. MI improves logic and reasoning through three paradigms: 1) localizing critical carriers and strengthening them via *Targeted Optimization* (§4.2); 2) extracting strategy-relevant directions/features and steering them with *Vector Arithmetic* (§4.3) or *Amplitude Manipulation* (§4.1); 3) diagnosing step-level failures with lightweight monitors for selective self-correction.

1) For arithmetic reasoning, MI localizes operator/reasoning carriers via *Causal Attribution* or *Magnitude Analysis*, and then improves performance by restricting *Targeted Optimization* to the identified heads, FFNs, or layers (Zhang et al., 2024d; Tan et al., 2025).

2) To steer reasoning strategies, MI identifies strategy-relevant directions or sparse features via *Magnitude Analysis* or cross-model feature discovery, then modulates inference with *Vector Arithmetic* (injecting strategy vectors) or *Amplitude Manipulation* (scaling/steering sparse features) to modulate reasoning behaviors (Venhoff et al., 2025a; Højer et al., 2025; Hong et al., 2025; Tang et al., 2025; Ward et al., 2025; Troitskii et al., 2025; Galichin et al., 2025; Cywiński et al., 2025; Zhang and Viteri, 2025; Wang et al., 2025h; Liu et al., 2025b; Sinii et al., 2025; Li et al., 2025g; Kim et al., 2026; Zhang et al., 2025g; Nguyen and Le, 2026; Zhang et al., 2025h).

3) For reasoning reliability, MI uses lightweight diagnosis: *Probing* or stepwise verification flags erroneous intermediate steps and triggers targeted re-generation before errors propagate (Sun et al., 2025b; You et al., 2025). Complementarily, quantifying token influence in CoT trajectories via *Gradient Detection* can also serves as a potential diagnostic tool to verify whether the model is attending to relevant logic or spurious correlations during generation (Wu et al., 2023).

5.3 Improve Efficiency

5.3.1 Efficient Training

Summary. MI enhances training efficiency primarily through two paradigms: 1) Sparse Fine-tuning, which isolates and updates only critical subnetworks via *Targeted Optimization* (§4.2); 2) Training Dynamics Monitoring, which leverages *Magnitude Analysis* (§3.1) to design internal metrics for tracking training status and avoiding unnecessary computations.

1) Works adopting the first paradigm, unlike PEFT methods that introduce external modules, achieve efficiency by fine-tuning intrinsic subsets—often matching or exceeding the performance of full fine-tuning. Specifically, studies employ *Gradient Detection* (Zhu et al., 2024; Song et al., 2024) or *Magnitude Analysis* (Xu et al., 2025a; Mondal et al., 2025; Gurgurov et al., 2025b) to pinpoint task-specific neurons, updating only their corresponding key/value weights ($\mathbf{k}_j^l/\mathbf{v}_j^l$). More granular approaches achieve massive efficiency by isolating extremely sparse subsets, such as 0.1% of neurons via *Causal Attribution* (Zhao et al., 2024b), attention heads based on $\mathbf{A}^{l,h}$ (Sergeev and Kotelnikov, 2025; Lai et al., 2025) or subgraphs via *Circuit Discovery* (Li et al., 2025e), demonstrating that substantial gains are attainable by targeting the correct mechanistic components.

2) The second paradigm predominantly leverages *Magnitude Analysis* to monitor the state evolution of internal objects, addressing the limitations of traditional validation loss in capturing critical phase transitions. For instance, some studies track the $\mathbf{A}^{l,h}$ of induction heads to signal the onset of in-context learning (Hoogland et al., 2024; Singh et al., 2024; Minegishi et al., 2025), while others monitor weight norms, gradients, or activations of internal objects to predict generalization phases, cautioning against premature stopping (Liu et al., 2023b; Furuta et al., 2024; Qiye et al., 2024; Li et al., 2025h).

5.3.2 Efficient Inference

Summary. MI enhances inference efficiency primarily through two paradigms: 1) Selective Computation, which utilizes *Magnitude Analysis* (§3.1) or *Circuit Discovery* (§3.6) to identify redundant calculations; 2) Adaptive Quantization, which optimizes bit-width allocation based on object sensitivity.

1) Works adopting the first paradigm effectively induce sparsity by pruning dispensable objects at both data and model levels. At the *data level*, researchers employ *Gradient Detection* (Xia et al., 2025; Lei et al., 2025) or *Magnitude Analysis* to prune tokens and KV cache (Guo et al., 2024; Ye et al., 2025b; He et al., 2024a; Cai et al., 2024b). Advanced methods further utilize *Circuit Discovery* to preserve attention heads critical for performance (Tang et al., 2024a; Xiao et al., 2024). At the *model level*, MI guides the skipping of architectural objects. For instance, applying *Magnitude Analysis* or *Probing* to residual streams or router activations enables the dynamic bypassing of redundant layers (Laitenberger et al., 2025; Valade, 2024; Elhoushi et al., 2024; Shelke et al., 2024; Lawson and Aitchison, 2025; Men et al., 2025) and inefficient MoE experts (Lu et al., 2024; Su et al., 2025d). Furthermore, identifying specialized neurons (Liu et al., 2024b; Tan et al., 2024a) allows for the execution of only critical subnetworks.

2) For the second paradigm, a primary application is *Layer-wise Mixed-Precision Quantization*, which dictates bit-width allocation based on layer sensitivity. To quantify this importance, studies leverage diverse mechanistic signals, including *Magnitude Analysis* of outliers or $\mathbf{x}^l$ (Dumitru et al., 2024; Zhang et al., 2025a; Xiao et al., 2025b; Zhang et al., 2026), *Gradient Detection* (Ranjan and Savakis, 2025), and *Vocab Projection* (Zeng et al., 2024). Guided by these insights, lower bit-widths are assigned to robust layers to maximize efficiency, while higher precision is retained for sensitive layers.

6 Conclusion

In this survey, we systematically reframe MI from a primarily observational endeavor into a practical, actionable paradigm. Under the unified pipeline of "*Locate, Steer, and Improve*", we clarify how interpretable objects can be precisely localized, causally manipulated, and ultimately leveraged to enhance alignment, capability, and efficiency in LLMs. Our analysis highlights that many recent advances—ranging from safety and persona alignment, to knowledge editing, and further to sparse fine-tuning—are most effective when grounded in explicit mechanistic intervention. We further discuss challenges and future directions in Appendix A, aiming to provide a solid foundation to advance more powerful and reliable LLMs.

Limitation

This survey focuses on mechanistic interpretability for dense large language models and does not systematically cover methods specific to other architectures and modalities. In particular, Mixture-of-Experts (MoE) models introduce routing and sparsely activated experts, while vision models rely on modality-specific representations and structures that pose distinct interpretability challenges. Although many of the methods discussed in this work are conceptually general, a comprehensive treatment of these architectures is left to future work.

In addition, the field lacks unified benchmarks or standardized evaluation protocols for localization methods, making it difficult to compare approaches or to determine whether the identified model components are causally optimal. This limitation also affects downstream applications, where interventions often rely on a single localization method without formal guarantees. Some works partially mitigate this issue by using multiple localization techniques and observing convergence on similar model regions, but principled evaluation remains an open challenge.

Acknowledgement

The authors thank the anonymous reviewers for their helpful comments. This work was supported in part by the Theme-based Research Scheme (TRS) project T45-701/22-R of the Research Grants Council of Hong Kong, the AVNET-HKU Emerging Microelectronics and Ubiquitous Systems (EMUS) Lab, the Henan Province Major Industrial "Challenge-Based Innovation" program (No. 251000210300), and the National Natural Science Foundation of China (Nos. 62476061, 62376061, 62576106).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Hiba Ahsan, Arnab Sen Sharma, Silvio Amir, David Bau, and Byron C. Wallace. 2025. Elucidating Mechanisms of Demographic Bias in LLMs for Healthcare. http://arxiv.org/abs/2502.13319. *Preprint*, arXiv:2502.13319.

Mst. Shapna Akter, Hossain Shahriar, Alfredo Cuzzocrea, and Fan Wu. 2024. Uncovering the interpretation of large language models. In *48th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2024, Osaka, Japan, July 2-4, 2024*, pages 1057–1066. IEEE.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 1, learning hierarchical language structures. *arXiv preprint arXiv:2305.13673*.

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, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. 2025. Systematic outliers in large language models. *arXiv preprint arXiv:2502.06415*.

Lyzander Marciano Andrylie, Inaya Rahmanisa, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. Sparse autoencoders can capture language-specific concepts across diverse languages. *Preprint*, arXiv:2507.11230.

Anthropic. 2024. Introducing Claude 3.5 Sonnet. Announcement of Claude 3.5 Sonnet model release, featuring improved intelligence, vision capabilities, and new Artifacts feature.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. 2025. Saes are good for steering–if you select the right features. *arXiv preprint arXiv:2505.20063*.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L Croci, Bo Li, Pashmina Cameron, Martin Jaggi,

Dan Alistarh, Torsten Hoefler, and James Hensman. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37:100213–100240.

Behrooz Azarkhalili and Maxwell W. Libbrecht. 2025. Generalized attention flow: Feature attribution for transformer models via maximum flow. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19954–19974. Association for Computational Linguistics.

Nooshin Bahador. 2025. Mechanistic interpretability of fine-tuned vision transformers on distorted images: Decoding attention head behavior for transparent and trustworthy ai. *arXiv preprint arXiv:2503.18762*.

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

Amin Banayeeanzade, Ala N Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. 2025. Psychological steering in llms: An evaluation of effectiveness and trustworthiness. *arXiv preprint arXiv:2510.04484*.

Tetiana Bas and Krystian Novak. 2025. Steering latent traits, not learned facts: An empirical study of activation control limits. *arXiv preprint arXiv:2511.18284*.

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*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Comput. 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*.

Sumanta Bhattacharyya and Pedram Rooshenas. 2025. Steered generation via gradient descent on sparse features. *arXiv preprint arXiv:2502.18644*.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4135–4144.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14113–14130. Association for Computational Linguistics.

Joseph Bloom. 2024. Open source sparse autoencoders for all residual stream layers of gpt2 small.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, volume 36, pages 75067–75096. Curran Associates, Inc.

Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. 2025. Beyond multiple choice: Evaluating steering vectors for adaptive free-form summarization. *arXiv preprint arXiv:2505.24859*.

Trenton Bricken and Cengiz Pehlevan. 2021. Attention approximates sparse distributed memory. *Advances in Neural Information Processing Systems*, 34:15301–15315.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. 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 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Bart Bussmann, Patrick Leask, and Neel Nanda. 2024. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. 2024a. Locating and Mitigating Gender Bias in Large Language Models. *Preprint*, arXiv:2403.14409.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and 1 others. 2024b. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808.

Bhavik Chandna, Zubair Bashir, and Procheta Sen. 2025. Dissecting Bias in LLMs: A Mechanistic Interpretability Perspective. *Preprint*, arXiv:2506.05166.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15662–15693.

David Chanin, James Wilken-Smith, Tomás Dulka, Hardik Bhatnagar, and Joseph Bloom. 2024. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *CoRR*, abs/2409.14507.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. 2025a. Transferring linear features across language models with model stitching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. 2025b. Towards understanding safety alignment: A mechanistic perspective from safety neurons.

Lihu Chen, Adam Dejl, and Francesca Toni. 2025c. Identifying query-relevant neurons in large language models for long-form texts. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 23595–23604. AAAI Press.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025d. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024a. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *Forty-first International Conference on Machine Learning*.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. 2024b. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning. In *Forty-first International Conference on Machine Learning*.

Xinrong Chen, Xu Chu, Yingmin Qiu, Hengyuan Zhang, Jing Xiong, Shiyu Tang, Shuai Liu, Shaokang Yang, Cheng Yang, Hayden Kwok-Hay So, and 1 others. 2026. Residual decoding: Mitigating hallucinations in large vision-language models via history-aware residual guidance. *arXiv preprint arXiv:2602.01047*.

Yilong Chen, Junyuan Shang, Zhenyu Zhang, Yanxi Xie, Jiawei Sheng, Tingwen Liu, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. 2025e. Inner thinking transformer: Leveraging dynamic depth scaling to foster adaptive internal thinking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28241–28259, Vienna, Austria. Association for Computational Linguistics.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024c. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17817–17825. AAAI Press.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025f. Knowledge localization: Mission not accomplished? enter query localization! In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zewen Chi, Heyan Huang, and Xian-Ling Mao. 2023. Can cross-lingual transferability of multilingual transformers be activated without end-task data? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12572–12584, Toronto, Canada. Association for Computational Linguistics.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar Van Der Wal. 2023. Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore. Association for Computational Linguistics.

Hakaze Cho, Haolin Yang, Brian M Kurkoski, and Naoya Inoue. 2025. Binary autoencoder for mechanistic interpretability of large language models. *arXiv preprint arXiv:2509.20997*.

Ikhyun Cho and Julia Hockenmaier. 2025. Toward efficient sparse autoencoder-guided steering for improved in-context learning in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28949–28961.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Henry Conklin and Kenny Smith. 2024. Representations as language: An information-theoretic

framework for interpretability. *arXiv preprint arXiv:2406.02449*.

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.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. OR-Bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Bartosz Cywiński, Bart Bussmann, Arthur Conmy, Josh Engels, Neel Nanda, and Senthooran Rajamanoharan. 2025. Can we interpret latent reasoning using current mechanistic interpretability tools?

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off course: Reliability challenges in steering language models. *arXiv preprint arXiv:2504.04635*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Adam Davies and Ashkan Khakzar. 2024. The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms. *arXiv preprint arXiv:2408.05859*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao yang, Xin Zhao, and Ji-Rong Wen. 2025. Neuron based personality trait induction in large language models. In *The Thirteenth International Conference on Learning Representations*.

Fabrizio Dimino, Krati Saxena, Bhaskarjit Sarmah, and Stefano Pasquali. 2025. Tracing Positional Bias in Financial Decision-Making: Mechanistic Insights from Qwen2.5. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 96–104.

Maximilian Dreyer, Lorenz Hufe, Jim Berend, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2025. From what to how: Attributing clip's latent components reveals unexpected semantic reliance. *arXiv preprint arXiv:2505.20229*.

Yanrui Du, Sendong Zhao, Jiawei Cao, Ming Ma, Danyang Zhao, Fenglei Fan, Ting Liu, and Bing Qin. 2024. Towards secure tuning: Mitigating security risks arising from benign instruction fine-tuning. *CoRR*, abs/2410.04524.

Xufeng Duan, Xinyu Zhou, Bei Xiao, and Zhenguang Cai. 2025. Unveiling language competence neurons: A psycholinguistic approach to model interpretability. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10148–10157, Abu Dhabi, UAE. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Razvan-Gabriel Dumitru, Vikas Yadav, Rishabh Maheshwary, Paul-Ioan Clotan, Sathwik Tejaswi Madhusudhan, and Mihai Surdeanu. 2024. Layer-wise quantization: A pragmatic and effective method for quantizing llms beyond integer bit-levels. *arXiv preprint arXiv:2406.17415*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger B. Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *CoRR*, abs/2209.10652.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. 2024. LayerSkip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12622–12642, Bangkok, Thailand. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter

models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Jiahai Feng and Jacob Steinhardt. 2023. How do language models bind entities in context? In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural 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. *arXiv preprint arXiv:2403.00824*.

Pedro Ferreira, Wilker Aziz, and Ivan Titov. 2025. Truthful or fabricated? using causal attribution to mitigate reward hacking in explanations. In *Workshop on Actionable Interpretability at ICML 2025*.

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Towards empirical interpretation of internal circuits and properties in grokked transformers on modular polynomials. *Transactions on Machine Learning Research*.

Andrey Galichin, Alexey Dontsov, Polina Druzhinina, Anton Razzhigaev, Oleg Y Rogov, Elena Tutubalina, and Ivan Oseledets. 2025. I have covered all the bases here: Interpreting reasoning features in large language models via sparse autoencoders. *arXiv preprint arXiv:2503.18878*.

Sandeep Reddy Gantla. 2025. Exploring mechanistic interpretability in large language models: Challenges, approaches, and insights. In *2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, pages 1–8. IEEE.

Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. 2025a. H-neurons: On the existence, impact, and origin of hallucination-associated neurons in llms. *Preprint*, arXiv:2512.01797.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V Govande, Bowen Baker, and Dan Mossing. 2025b. Weight-sparse transformers have interpretable circuits. *arXiv preprint arXiv:2511.13653*.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and 1 others. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

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, Kevin Wang, and Yoav Goldberg. 2022. 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.

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, Ann Yuan, Marius Guerard, Emily Reif, Michael A. Lepori, and Lucas Dixon. 2024. Who's asking? user personas and the mechanics of latent misalignment. In *Advances in Neural Information Processing Systems*, volume 37, pages 125967–126003. Curran Associates, Inc.

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*.

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

Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, and Hari Sundaram. 2025. Breaking Bad Tokens: Detoxification of LLMs Using Sparse Autoencoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12702–12720, Suzhou, China. Association for Computational Linguistics.

Oliver Gruber and Thomas Goschke. 2004. Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes. *Acta psychologica*, 115(2-3):105–121.

Aleksandra Gruszka and Gerald Matthews. 2010. *Handbook of individual differences in cognition: Attention, memory, and executive control*. Springer.

Xin Guan, PeiHsin Lin, Zekun Wu, Ze Wang, Ruibo Zhang, Emre Kazim, and Adriano Koshiyama. 2025. MPF: Aligning and Debiasing Language Models post Deployment via Multi Perspective Fusion. *Preprint*, arXiv:2507.02595.

Zhiyu Guo, Hidetaka Kamigaito, and Taro Watanabe. 2024. Attention score is not all you need for token importance indicator in KV cache reduction: Value also matters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21158–21166, Miami, Florida, USA. Association for Computational Linguistics.

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. 2025. Enhancing automated interpretability with output-centric feature descriptions. *arXiv preprint arXiv:2501.08319*.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025a. Language arithmetics: Towards systematic language neuron identification and manipulation. *Preprint*, arXiv:2507.22608.

Daniil Gurgurov, Josef van Genabith, and Simon Ostermann. 2025b. Sparse subnetwork enhancement for underrepresented languages in large language models. *Preprint*, arXiv:2510.13580.

Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, and Yonatan Belinkov. 2025. Position-aware automatic circuit discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2792–2817. Association for Computational Linguistics.

Gunmay Handa, Zekun Wu, Adriano Koshiyama, and Philip Colin Treleaven. 2025. Personality as a probe for LLM evaluation: Method trade-offs and downstream effects. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

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 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*.

Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. 2025. Circuit-tracer: A new library for finding feature circuits. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 239–249, Suzhou, China. Association for Computational Linguistics.

Yefei He, Luoming Zhang, Weijia Wu, Jing Liu, Hong Zhou, and Bohan Zhuang. 2024a. Zipcache: Accurate and efficient kv cache quantization with salient token identification. *Advances in Neural Information Processing Systems*, 37:68287–68307.

Zhengfu He, Xuyang Ge, Qiong Tang, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024b. 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, and 1 others. 2024c. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. 2025a. Sae-ssv: Supervised steering in sparse representation spaces for reliable control of language models. *arXiv preprint arXiv:2505.16188*.

Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, and Mengnan Du. 2025b. Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models. *arXiv preprint arXiv:2502.11356*.

Amr Hegazy, Mostafa Elhoushi, and Amr Alanwar. 2025. Guiding giants: Lightweight controllers for weighted activation steering in llms. *Preprint*, arXiv:2505.20309.

Musashi Hinck, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shao-Yen Tseng, and Vasudev Lal. 2024. Why do LLaVA vision-language models reply to images in English? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13402–13421, Miami, Florida, USA. Association for Computational Linguistics.

Yihuai Hong, Meng Cao, Dian Zhou, Lei Yu, and Zhijing Jin. 2025. The reasoning-memorization interplay in language models is mediated by a single direction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21565–21585, Vienna, Austria. Association for Computational Linguistics.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. 2024. The developmental landscape of in-context learning, 2024. *URL https://arxiv.org/abs/2402.02364*.

Cheng-Hsun Hsueh, Paul Kuo-Ming Huang, Tzu-Han Lin, Che-Wei Liao, Hung-Chieh Fang, Chao-Wei Huang, and Yun-Nung Chen. 2024. Editing the mind of giants: An in-depth exploration of pitfalls of knowledge editing in large language models. *arXiv preprint arXiv:2406.01436*.

Lijie Hu, Chenyang Ren, Zhengyu Hu, Hongbin Lin, Cheng-Long Wang, Hui Xiong, Jingfeng Zhang, and Di Wang. 2025. Editable concept bottleneck models. *Preprint*, arXiv:2405.15476.

Haoming Huang, Yibo Yan, Jiahao Huo, Xin Zou, Xinfeng Li, Kun Wang, and Xuming Hu. 2025a. Pierce

the mists, greet the sky: Decipher knowledge overshadowing via knowledge circuit analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15471–15490, Suzhou, China. Association for Computational Linguistics.

Jing Huang, Junyi Tao, Thomas Icard, Diyi Yang, and Christopher Potts. 2025b. Internal causal mechanisms robustly predict language model out-of-distribution behaviors. *Preprint*, arXiv:2505.11770.

Yufei Huang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Unified view of grokking, double descent and emergent abilities: A comprehensive study on algorithm task. In *First Conference on Language Modeling*.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6816.

Bertram Højer, Oliver Jarvis, and Stefan Heinrich. 2025. Improving reasoning performance in large language models via representation engineering. *Preprint*, arXiv:2504.19483.

Aya Abdelsalam Ismail, Tuomas Oikarinen, Amy Wang, Julius Adebayo, Samuel Stanton, Taylor Joren, Joseph Kleinhenz, Allen Goodman, Héctor Corrada Bravo, Kyunghyun Cho, and 1 others. 2024. Concept bottleneck language models for protein design. *arXiv preprint arXiv:2411.06090*.

Farnoush Rezaei Jafari, Oliver Eberle, Ashkan Khakzar, and Neel Nanda. 2025. Relp: Faithful and efficient circuit discovery via relevance patching. *CoRR*, abs/2508.21258.

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.

Gangwei Jiang, Zhaoyi Li, Caigao Jiang, Siqiao Xue, Jun Zhou, Linqi Song, Defu Lian, and Ying Wei. 2024b. Interpretable catastrophic forgetting of large language model fine-tuning via instruction vector. *arXiv e-prints*, pages arXiv–2406.

Gangwei Jiang, Zhaoyi Li, Defu Lian, and Ying Wei. 2024c. Refine large language model fine-tuning via instruction vector. *arXiv preprint arXiv:2406.12227*.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025a. Massive values in self-attention modules are the key to contextual knowledge understanding. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025b. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 558–573. Association for Computational Linguistics.

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, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1193–1215. Association for Computational Linguistics.

Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. LinguaLens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28220–28239, Suzhou, China. Association for Computational Linguistics.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359, Miami, Florida, USA. Association for Computational Linguistics.

Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. 2025. Probing then editing response personality of large language models. In *Second Conference on Language Modeling*.

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. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, pages 8235–8246. ELRA and ICCL.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7721–7735. Association for Computational Linguistics.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. Are sparse autoencoders useful? A case study in sparse probing. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Amir Hossein Kargaran, Yihong Liu, François Yvon, and Hinrich Schütze. 2025. How programming concepts and neurons are shared in code language models. *arXiv preprint arXiv:2506.01074*.

Sheer Karny, Anthony Baez, and Pat Pataranutaporn. 2025. Neural transparency: Mechanistic interpretability interfaces for anticipating model behaviors for personalized ai. *arXiv preprint arXiv:2511.00230*.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, and 1 others. 2025. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint arXiv:2503.09532*.

Aly M. Kassem, Zhuan Shi, Negar Rostamzadeh, and Golnoosh Farnadi. 2025. Reviving your MNEME: predicting the side effects of LLM unlearning and fine-tuning via sparse model diffing. *CoRR*, abs/2507.21084.

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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2390–2422. Association for Computational Linguistics.

Ihor Kendiukhov. 2025. A review of developmental interpretability in large language models. *arXiv preprint arXiv:2508.15841*.

Dmitrii Kharlapenko, Stepan Shabalin, Fazl Barez, Arthur Conmy, and Neel Nanda. 2025. Scaling sparse feature circuit finding for in-context learning. *arXiv preprint arXiv:2504.13756*.

Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. 2025a. Interpreting attention heads for image-to-text information flow in large vision–language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Junsol Kim, James Evans, and Aaron Schein. 2025b. Linear representations of political perspective emerge in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Junsol Kim, Shiyang Lai, Nino Scherrer, James Evans, and 1 others. 2026. Reasoning models generate societies of thought. *arXiv preprint arXiv:2601.10825*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. 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 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Wen Lai, Alexander Fraser, and Ivan Titov. 2025. Joint localization and activation editing for low-resource fine-tuning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Filipe Laitenberger, Dawid Kopiczko, Cees G. M. Snoek, and Yuki M. Asano. 2025. What layers when: Learning to skip compute in llms with residual gates. *Preprint*, arXiv:2510.13876.

Tim Lawson and Laurence Aitchison. 2025. Learning to skip the middle layers of transformers. *Preprint*, arXiv:2506.21103.

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. 2024. Residual stream analysis with multi-layer saes. In *The Thirteenth International Conference on Learning Representations*.

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. *Preprint*, arXiv:2401.01967.

Seongmin Lee, Aeree Cho, Grace C. Kim, ShengYun Peng, Mansi Phute, and Duen Horng Chau. 2025. Interpretation meets safety: A survey on interpretation methods and tools for improving LLM safety. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21514–21545, Suzhou, China. Association for Computational Linguistics.

Lei Lei, Jie Gu, Xiaokang Ma, Chu Tang, Jingmin Chen, and Tong Xu. 2025. Generic token compression in multimodal large language models from an explainability perspective. *Preprint*, arXiv:2506.01097.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025a. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Haoling Li, Xin Zhang, Xiao Liu, Yeyun Gong, Yifan Wang, Qi Chen, and Peng Cheng. 2025b. Enhancing large language model performance with gradient-based parameter selection. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 24431–24439. AAAI Press.

Jiaming Li, Haoran Ye, Yukun Chen, Xinyue Li, Lei Zhang, Hamid Alinejad-Rokny, Jimmy Chih-Hsien Peng, and Min Yang. 2025c. Training superior sparse autoencoders for instruct models. *arXiv preprint arXiv:2506.07691*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 19730–19742. PMLR.

Ruizhe Li, Chen Chen, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. 2026. Attributing response to context: A jensen-shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. In *The Fourteenth International Conference on Learning Representations*.

Ruizhe Li and Yanjun Gao. 2025. Anchored Answers: Unravelling Positional Bias in GPT-2's Multiple-Choice Questions. *Preprint*, arXiv:2405.03205.

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

Xuying Li, Zhuo Li, Yuji Kosuga, Yasuhiro Yoshida, and Victor Bian. 2024. Precision Knowledge Editing: Enhancing Safety in Large Language Models. *Preprint*, arXiv:2410.03772.

Yueyan Li, Wenhao Gao, Caixia Yuan, and Xiaojie Wang. 2025e. Fine-tuning is subgraph search: A new lens on learning dynamics. *Preprint*, arXiv:2502.06106.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025f. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Zihao Li, Xu Wang, Yuzhe Yang, Ziyu Yao, Haoyi Xiong, and Mengnan Du. 2025g. Feature extraction and steering for enhanced chain-of-thought reasoning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10913, Suzhou, China. Association for Computational Linguistics.

Ziyue Li, Chenrui Fan, and Tianyi Zhou. 2025h. Where to find grokking in llm pretraining? monitor memorization-to-generalization without test. *arXiv preprint arXiv:2506.21551*.

Xiao Liang, Zhong-Zhi Li, Yeyun Gong, Yang Wang, Hengyuan Zhang, Yelong Shen, Ying Nian Wu, and Weizhu Chen. 2025. Sws: Self-aware weakness-driven problem synthesis in reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.08989*.

Xiao Liang, Zhong-Zhi Li, Zhenghao Lin, Eric Hancheng Jiang, Hengyuan Zhang, Yelong Shen, Kai-Wei Chang, Ying Nian Wu, Yeyun Gong, and Weizhu Chen. 2026. Training llms for divide-and-conquer reasoning elevates test-time scalability. *arXiv preprint arXiv:2602.02477*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, and 1 others. 2025. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*.

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, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qi Liu, Haozhe Duan, Jiaxin Mao, and Ji-Rong Wen. 2025a. How do large language models understand relevance? a mechanistic interpretability perspective. *ACM Transactions on Information Systems*.

Sheng Liu, Tianlang Chen, Pan Lu, Haotian Ye, Yizheng Chen, Lei Xing, and James Zou. 2025b. Fractional reasoning via latent steering vectors improves inference time compute. *Preprint*, arXiv:2506.15882.

Shuqi Liu, Han Wu, Bowei He, Xiongwei Han, Mingxuan Yuan, and Linqi Song. 2025c. Sens-merging: Sensitivity-guided parameter balancing for merging large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19243–19255. Association for Computational Linguistics.

Tianlin Liu, Mathieu Blondel, Carlos Riquelme Ruiz, and Joan Puigcerver. 2024a. Routers in vision mixture of experts: An empirical study. *Trans. Mach. Learn. Res.*, 2024.

Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024b. Unraveling babel: Exploring multilingual activation patterns within large language models. *arXiv*.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024c. The Devil is in the Neurons: Interpreting and Mitigating Social Biases in Pre-trained Language Models. *Preprint*, arXiv:2406.10130.

Yihong Liu, Runsheng Chen, Lea Hirlimann, Ahmad Dawar Hakimi, Mingyang Wang, Amir Hossein Kargaran, Sascha Rothe, François Yvon, and Hinrich Schuetze. 2025d. On relation-specific neurons in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 992–1022, Suzhou, China. Association for Computational Linguistics.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schuetze. 2025e. Tracing multilingual factual knowledge acquisition in pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2121–2146, Suzhou, China. Association for Computational Linguistics.

Ziming Liu, Eric J Michaud, and Max Tegmark. 2023b. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations*.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. 2026. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*.

Dawn Lu and Nina Rimsky. 2024. Investigating bias representations in llama 2 chat via activation steering. *arXiv preprint arXiv:2402.00402*.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. 2024. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. *arXiv preprint arXiv:2402.14800*.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*.

Ang Lv, Kaiyi Zhang, Yuhan Chen, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. 2024. Interpreting key mechanisms of factual recall in transformer-based language models. *CoRR*, abs/2403.19521.

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, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2024. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 337–363, Miami, Florida, US. Association for Computational Linguistics.

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7713–7724.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2025. ShortGPT: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20192–20204, Vienna, Austria. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Joseph Miller, Bilal Chughtai, and William Saunders. 2024. Transformer circuit faithfulness metrics are not robust. *arXiv preprint arXiv:2407.08734*.

Gouki Minegishi, Hiroki Furuta, Shohei Taniguchi, Yusuke Iwasawa, and Yutaka Matsuo. 2025. In-context meta learning induces multi-phase circuit emergence. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.

John Jacob Brooke Morgan and Adam Raymond Gilliland. 1927. *An introduction to psychology*. Macmillan.

Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. 2024. Exploring alignment in shared cross-lingual spaces. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6326–6348, Bangkok, Thailand. Association for Computational Linguistics.

Anish Mudide, Joshua Engels, Eric J Michaud, Max Tegmark, and Christian Schroeder de Witt. 2024. Efficient dictionary learning with switch sparse autoencoders. *arXiv preprint arXiv:2410.08201*.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, and 4 others. 2025. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Aashiq Muhamed, Jacopo Bonato, Mona T. Diab, and Virginia Smith. 2025a. Saes *Can* improve unlearning: Dynamic sparse autoencoder guardrails for precision unlearning in llms. *CoRR*, abs/2504.08192.

Aashiq Muhamed, Mona Diab, and Virginia Smith. 2025b. 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.

Aashiq Muhamed and Virginia Smith. 2025. The geometry of forgetting: Analyzing machine unlearning through local learning coefficients. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. *URL: https://www. neelnanda. io/mechanistic-interpretability/attribution-patching*.

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

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. Grains: Gradient-based attribution for inference-time steering of llms and vlms. *CoRR*, abs/2507.18043.

Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. 2025b. Probing and steering evaluation awareness of language models, july 2025. *URL http://arxiv. org/abs/2507.01786*.

Trang Nguyen, Jackson Michaels, Madalina Fiterau, and David Jensen. 2025c. Challenges in understanding modality conflict in vision-language models. *arXiv preprint arXiv:2509.02805*.

Tuc Nguyen and Thai Le. 2026. Atlas: Adaptive test-time latent steering with external verifiers for enhancing llms reasoning. *arXiv preprint arXiv:2601.03093*.

Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2025. Mechanistic understanding and mitigation of language confusion in English-centric large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 690–706, Suzhou, China. Association for Computational Linguistics.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. 2025a. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025b. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. 2025. Llama see, llama do: A mechanistic perspective on contextual entrainment and distraction in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16218–16239, Vienna, Austria. Association for Computational Linguistics.

nostalgebraist. 2020. Interpreting GPT: the logit lens.

Pascal Notsawo Jr, Hattie Zhou, Mohammad Pezeshki, Irina Rish, Guillaume Dumas, and 1 others. 2023. Predicting grokking long before it happens: A look into the loss landscape of models which grok. *arXiv preprint arXiv:2306.13253*.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. 2025. Probing semantic routing in large mixture-of-expert models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 18263–18278. Association for Computational Linguistics.

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, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. 2025. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tsung-Min Pai, Jui-I Wang, Li-Chun Lu, Shao-Hua Sun, Hung-Yi Lee, and Kai-Wei Chang. 2025. Billy: Steering large language models via merging persona vectors for creative generation. *arXiv preprint arXiv:2510.10157*.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C 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.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. 2025. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Forty-second International Conference on Machine Learning*.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. Task-specific skill localization in fine-tuned language models. *Preprint*, ICML:2302.06600.

Avi Parrack, Carlo Leonardo Attubato, and Stefan Heimersheim. 2025. Benchmarking deception probes via black-to-white performance boosts. *arXiv preprint arXiv:2507.12691*.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29, Dubrovnik, Croatia. Association for Computational Linguistics.

Anirudh Phukan, Divyansh, Harshit Kumar Morj, Vaishnavi, Apoorv Saxena, and Koustava Goswami. 2025. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in VLMs. 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 9661–9675, Albuquerque, New Mexico. Association for Computational Linguistics.

Joris Postmus and Steven Abreu. 2024. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv preprint arXiv:2410.16314*.

Daniele Potertì, Andrea Seveso, and Fabio Mercorio. 2025. Can role vectors affect LLM behaviour? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17735–17747, Suzhou, China. Association for Computational Linguistics.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *arXiv preprint arXiv:2402.14811*.

Hu Qiye, Zhou Hao, and Yu RuoXi. 2024. Exploring grokking: Experimental and mechanistic investigations. *arXiv preprint arXiv:2412.10898*.

Philip Quirke and Fazl Barez. 2024. Understanding addition in transformers. In *The Twelfth International Conference on Learning Representations*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Bianca Raimondi, Daniela Dalbagno, and Maurizio Gabbrielli. 2025. Analysing Moral Bias in Finetuned LLMs through Mechanistic Interpretability. *Preprint*, arXiv:2510.12229.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. 2024a. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*.

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*.

Navin Ranjan and Andreas Savakis. 2025. Mix-qvit: Mixed-precision vision transformer quantization driven by layer importance and quantization sensitivity. *arXiv preprint arXiv:2501.06357*.

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.

Santhosh Kumar Ravindran. 2025. Adversarial activation patching: A framework for detecting and mitigating emergent deception in safety-aligned transformers. *arXiv preprint arXiv:2507.09406*.

Yuqi Ren, Renren Jin, Tongxuan Zhang, and Deyi Xiong. 2025. Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2988–3001.

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.

Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2023. Attention lens: A tool for mechanistically interpreting the attention head information retrieval mechanism. *arXiv preprint arXiv:2310.16270*.

Alexander Sergeev and Evgeny Kotelnikov. 2025. Optimizing multimodal language models through attention-based interpretability. *Preprint*, arXiv:2511.23375.

Adam Shai, Lucas Teixeira, Alexander Oldenziel, Sarah Marzen, and Paul Riechers. 2024. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034.

Chenming Shang, Hengyuan Zhang, Hao Wen, and Yujiu Yang. 2024a. Understanding multimodal deep neural networks: A concept selection view. *arXiv preprint arXiv:2404.08964*.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. 2024b. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11040.

Jiandong Shao, Yao Lu, and Jianfei Yang. 2025. Benford's curse: Tracing digit bias to numerical hallucination in llms. *arXiv preprint arXiv:2506.01734*.

Vansh Sharma and Venkat Raman. 2025. Steering conceptual bias via transformer latent-subspace activation. *Preprint*, arXiv:2506.18887.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Anushka Shelke, Riya Savant, and Raviraj Joshi. 2024. Towards building efficient sentence bert models using layer pruning. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 720–725.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. IRCAN: mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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 (ICML)*, volume 70 of *Proceedings of Machine Learning Research*.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1690–1712, Suzhou, China. Association for Computational Linguistics.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. 2024. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 45637–45662.

Viacheslav Sinii, Alexey Gorbatovski, Artem Cherepanov, Boris Shaposhnikov, Nikita Balagansky, and Daniil Gavrilov. 2025. Steering llm reasoning through bias-only adaptation. *Preprint*, arXiv:2505.18706.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

Weixi Song, Zuchao Li, Lefei Zhang, Hai Zhao, and Bo Du. 2024. Sparse is enough in fine-tuning pre-trained large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. 2025. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint arXiv:2501.09929*.

Divyansh Srivastava, Ge Yan, and Lily Weng. 2024. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37:79057–79094.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. 2024. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8189–8200, Miami, Florida, USA. Association for Computational Linguistics.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*.

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.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025a. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974, Vienna, Austria. Association for Computational Linguistics.

Yi Su, Jiayi Zhang, Shu Yang, Xinhai Wang, Lijie Hu, and Di Wang. 2025b. Understanding how value neurons shape the generation of specified values in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9433–9452, Suzhou, China. Association for Computational Linguistics.

Zunhai Su, Zhe Chen, Wang Shen, Hanyu Wei, Linge Li, Huangqi Yu, and Kehong Yuan. 2025c. Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations. *arXiv preprint arXiv:2501.16383*.

Zunhai Su, Qingyuan Li, Hao Zhang, YuLei Qian, Yuchen Xie, and Kehong Yuan. 2025d. Unveiling super experts in mixture-of-experts large language models. *arXiv preprint arXiv:2507.23279*.

Zunhai Su and Kehong Yuan. 2025. Kvsink: Understanding and enhancing the preservation of attention sinks in kv cache quantization for llms. *arXiv preprint arXiv:2508.04257*.

Xavier Suau, Pieter Delobelle, Katherine Metcalf, Armand Joulin, Nicholas Apostoloff, Luca Zappella, and Pau Rodriguez. 2024. Whispering experts: Neural interventions for toxicity mitigation in language models. In *Forty-first International Conference on Machine Learning*.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. 2024a. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations*.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024b. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025a. Personality vector: Modulating personality of large language models by model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24667–24688, Suzhou, China. Association for Computational Linguistics.

Yucheng Sun, Alessandro Stolfo, and Mrinmaya Sachan. 2025b. Probing for arithmetic errors in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8122–8139, Suzhou, China. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 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.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120, Vienna, Austria. Association for Computational Linguistics.

Ryota Takatsuki, Sonia Joseph, Ippei Fujisawa, and Ryota Kanai. 2025. Decoding vision transformers: the

diffusion steering lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4819–4824.

Shaomu Tan, Di Wu, and Christof Monz. 2024a. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

Yuqiao Tan, Minzheng Wang, Shizhu He, Huanxuan Liao, Chengfeng Zhao, Qiunan Lu, Tian Liang, Jun Zhao, and Kang Liu. 2025. Bottom-up policy optimization: Your language model policy secretly contains internal policies. *Preprint*, arXiv:2512.19673.

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

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Shikuan Hong, Yiwu Yao, and Gongyi Wang. 2024a. Razorattention: Efficient kv cache compression through retrieval heads. *arXiv preprint arXiv:2407.15891*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024b. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Wayne Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6832–6849, Vienna, Austria. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua M Susskind. 2022. The slingshot mechanism: An empirical study of adaptive optimizers and the\emph {Grokking Phenomenon}. In *Has it Trained Yet? NeurIPS 2022 Workshop*.

Dmitrii Troitskii, Koyena Pal, Chris Wendler, and Callum Stuart McDougall. 2025. Internal states before wait modulate reasoning patterns. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18640–18649, Suzhou, China. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Florian Valade. 2024. Accelerating large language model inference with self-supervised early exits. *Preprint*, arXiv:2407.21082.

Teun van der Weij, Massimo Poesio, and Nandi Schoots. 2024. Extending activation steering to broad skills and multiple behaviours. *arXiv preprint arXiv:2403.05767*.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. 2023. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025a. Understanding reasoning in thinking language models via steering vectors. *Preprint*, arXiv:2506.18167.

Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. 2025b. How visual representations map to language feature space in multimodal llms. *arXiv preprint arXiv:2506.11976*.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024a. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. *Advances in Neural Information Processing Systems*, 37:95238–95265.

Haoyu Wang, Yaqing Wang, Tianci Liu, Tuo Zhao, and Jing Gao. 2023a. HadSkip: Homotopic and adaptive layer skipping of pre-trained language models for efficient inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4283–4294, Singapore. Association for Computational Linguistics.

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

Mengru Wang, Xingyu Chen, Yue Wang, Zhiwei He, Jiahao Xu, Tian Liang, Qiuzhi Liu, Yunzhi Yao, Wenxuan Wang, Ruotian Ma, Haitao Mi, Ningyu Zhang, Zhaopeng Tu, Xiaolong Li, and Dong Yu. 2025a. Two experts are all you need for steering thinking: Reinforcing cognitive effort in moe reasoning models without additional training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, and Ningyu Zhang. 2025b. Beyond prompt engineering: Robust behavior control in llms via steering target atoms. *arXiv preprint arXiv:2505.20322*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025c. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. 2025d. Language mixing in reasoning language models: Patterns, impact, and internal causes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2637–2665, Suzhou, China. Association for Computational Linguistics.

Mingze Wang, Ruoxi Yu, Lei Wu, and 1 others. 2024b. How transformers implement induction heads: Approximation and optimization analysis. *arXiv e-prints*, pages arXiv–2410.

Runyu Wang, Peng Ping, Zhengyu Guo, Xiaoye Zhang, Quan Shi, Liting Zhou, and Tianbo Ji. 2025e. Loki: Low-damage knowledge implanting of large language models. *Preprint*, arXiv:2505.22120.

Sizhe Wang, Yongqi Tong, Hengyuan Zhang, Dawei Li, Xin Zhang, and Tianlong Chen. 2024c. Bpo: Towards balanced preference optimization between knowledge breadth and depth in alignment. *arXiv preprint arXiv:2411.10914*.

Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. 2025f. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *The Thirteenth International Conference on Learning Representations*.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. 2025g. Refusal direction is universal across safety-aligned languages. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024d. Gradient based feature attribution in explainable AI: A technical review. *CoRR*, abs/2403.10415.

Zhenyu Wang. 2025. Logitlens4llms: Extending logit lens analysis to modern large language models. *arXiv preprint arXiv:2503.11667*.

Zijian Wang, Yanxiang Ma, and Chang Xu. 2025h. Eliciting chain-of-thought in base llms via gradient-based representation optimization. *Preprint*, arXiv:2511.19131.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2025i. Eliminating Position Bias of Language Models: A Mechanistic Approach. *Preprint*, arXiv:2407.01100.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. 2025. Reasoning-finetuning repurposes latent representations in base models. *CoRR*, abs/2507.12638.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Jiaqi Weng, Han Zheng, Hanyu Zhang, Qinqin He, Jialing Tao, Hui Xue, Zhixuan Chu, and Xiting Wang. 2025. Safe-sail: Towards a fine-grained safety landscape of large language models via sparse autoencoder interpretation framework. *arXiv preprint arXiv:2509.18127*.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *arXiv preprint arXiv:2307.13339*.

Xuansheng Wu, Jiayi Yuan, Wenlin Yao, Xiaoming Zhai, and Ninghao Liu. 2025a. 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, Lijie Hu, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, and 1 others. 2024a. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv preprint arXiv:2403.08946*.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025b. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025c. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024b. Reft: Representation finetuning for language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Dirk U Wulff and Rui Mata. 2025. Advancing cognitive science with llms. *arXiv preprint arXiv:2511.00206*.

xAI. 2025. Grok 3 beta — the age of reasoning agents. Blog post announcing Grok 3 Beta, describing improvements in reasoning capabilities and performance benchmarks.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. TokenSkip: Controllable chain-of-thought compression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3351–3363, Suzhou, China. Association for Computational Linguistics.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023a. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023b. Efficient streaming language models with attention sinks. *arXiv*.

Hanqi Xiao, Yi-Lin Sung, Elias Stengel-Eskin, and Mohit Bansal. 2025a. Task-circuit quantization: Leveraging knowledge localization and interpretability for compression. In *Second Conference on Language Modeling*.

He Xiao, Qingyao Yang, Dirui Xie, Wendong Xu, Wenyong Zhou, Haobo Liu, Zhengwu Liu, and Ngai Wong. 2025b. Exploring layer-wise information effectiveness for post-training quantization in small language models. *arXiv preprint arXiv:2508.03332*.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2025. SORRY-Bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations (ICLR)*.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. 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 5725–5737, Online. Association for Computational Linguistics.

Jing Xiong, Liyang Fan, Hui Shen, Zunhai Su, Min Yang, Lingpeng Kong, and Ngai Wong. 2025. Dope: Denoising rotary position embedding. *arXiv preprint arXiv:2511.09146*.

Jing Xiong, Qi Han, Yunta Hsieh, Hui Shen, Huajian Xin, Chaofan Tao, Chenyang Zhao, Hengyuan Zhang, Taiqiang Wu, Zhen Zhang, and 1 others. 2026a. Mmformalizer: Multimodal autoformalization in the wild. *arXiv preprint arXiv:2601.03017*.

Jing Xiong, Chengming Li, Min Yang, Xiping Hu, and Bin Hu. 2026b. Expression syntax information bottleneck for math word problems. *Preprint*, arXiv:2310.15664.

Jing Xiong, Hui Shen, Shansan Gong, Yuxin Cheng, Jianghan Shen, Chaofan Tao, Haochen Tan, Haoli Bai, Lifeng Shang, and Ngai Wong. 2026c. Ovd: On-policy verbal distillation. *arXiv preprint arXiv:2601.21968*.

Haoyun Xu, Runzhe Zhan, Yingpeng Ma, Derek F. Wong, and Lidia S. Chao. 2025a. Let's focus on neuron: Neuron-level supervised fine-tuning for large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9393–9406, Abu Dhabi, UAE. Association for Computational Linguistics.

Ziwen Xu, Shuxun Wang, Kewei Xu, Haoming Xu, Mengru Wang, Xinle Deng, Yunzhi Yao, Guozhou Zheng, Huajun Chen, and Ningyu Zhang. 2025b. Easyedit2: An easy-to-use steering framework for editing large language models. *arXiv preprint arXiv:2504.15133*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems 36:*

Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Hao Yang, Qianghua Zhao, and Lei Li. 2024. Chain-of-thought in large language models: Decoding, projection, and activation. *Preprint*, arXiv:2412.03944.

Jiayu Yang, Yuxuan Fan, Songning Lai, Shengen Wu, Jiaqi Tang, Chun Kang, Zhijiang Guo, and Yutao Yue. 2025a. Ace: Attribution-controlled knowledge editing for multi-hop factual recall. *arXiv preprint arXiv:2510.07896*.

Wanli Yang, Fei Sun, Rui Tang, Hongyu Zang, Du Su, Qi Cao, Jingang Wang, Huawei Shen, and Xueqi Cheng. 2025b. Fine-tuning done right in model editing. *arXiv preprint arXiv:2509.22072*.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024a. Knowledge circuits in pretrained transformers. In *Advances in Neural Information Processing Systems*, volume 37, pages 118571–118602. Curran Associates, Inc.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024b. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems*, 37:118571–118602.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2025a. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025b. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22128–22136.

Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Ranjan Satapathy, Roy Ka-Wei Lee, and Erik Cambria. 2025a. Understanding refusal in language models with sparse autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6377–6399, Suzhou, China. Association for Computational Linguistics.

Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. 2025b. Towards faithful natural language explanations: A study using activation patching in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10436–10458.

Fangcong Yin, Xi Ye, and Greg Durrett. 2024. Lofit: Localized fine-tuning on llm representations. *Advances in Neural Information Processing Systems*, 37:9474–9506.

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.

Qingyu Yin, Chak Tou Leong, Linyi Yang, Wenxuan Huang, Wenjie Li, Xiting Wang, Jaehong Yoon, YunXing, XingYu, and Jinjin Gu. 2025. Refusal falls off a cliff: How safety alignment fails in reasoning? *Preprint*, arXiv:2510.06036.

Weiqiu You, Anton Xue, Shreya Havaldar, Delip Rao, Helen Jin, Chris Callison-Burch, and Eric Wong. 2025. Probabilistic soundness guarantees in llm reasoning chains. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7517–7536, Suzhou, China. Association for Computational Linguistics.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. 2025a. Entangled in Representations: Mechanistic Investigation of Cultural Biases in Large Language Models. *Preprint*, arXiv:2508.08879.

Mengxia Yu, De Wang, Qi Shan, Colorado J Reed, and Alvin Wan. 2024. The super weight in large language models. *arXiv preprint arXiv:2411.07191*.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2025b. Mitigate position bias in LLMs via scaling a single hidden states channel. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6092–6111, Vienna, Austria. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024a. How do large language models learn in-context? query and key matrices of in-context heads are two towers for metric learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3281–3292. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024b. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3293–3306, Miami, Florida, USA. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024c. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3267–3280. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024d. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*.

Zeping Yu and Sophia Ananiadou. 2025. Understanding and Mitigating Gender Bias in LLMs via Interpretable Neuron Editing. *Preprint*, arXiv:2501.14457.

Zeping Yu, Yonatan Belinkov, and Sophia Ananiadou. 2025c. Back attention: Understanding and enhancing multi-hop reasoning in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11268–11283, Suzhou, China. Association for Computational Linguistics.

Shuzhou Yuan, Zhan Qu, Mario Tawfelis, and Michael Färber. 2025. From monolingual to bilingual: Investigating language conditioning in large language models for psycholinguistic tasks. *arXiv preprint arXiv:2508.02502*.

Dharunish Yugeswardeenoo, Harshil Nukala, Cole Blondin, Sean O'Brien, Vasu Sharma, and Kevin Zhu. 2025. Interpreting the latent structure of operator precedence in language models. In *The First Workshop on the Interplay of Model Behavior and Model Internals*.

Binrui Zeng, Bin Ji, Xiaodong Liu, Jie Yu, Shasha Li, Jun Ma, Xiaopeng Li, Shangwen Wang, Xinran Hong, and Yongtao Tang. 2024. Lsaq: Layer-specific adaptive quantization for large language model deployment. *arXiv preprint arXiv:2412.18135*.

Feng Zhang, Yanbin Liu, Weihua Li, Jie Lv, Xiaodan Wang, and Quan Bai. 2025a. Towards superior quantization accuracy: A layer-sensitive approach. *arXiv preprint arXiv:2503.06518*.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

Hengyuan Zhang, Xinrong Chen, Yingmin Qiu, Xiao Liang, Ziyue Li, Guanyu Wang, Weiping Li, Tong Mo, Hayden Kwok-Hay So, and Ngai Wong. 2025b. Guilomo: Allocating expert number and rank for loramoe via bilevel optimization with guidedselection vectors. *Preprint*.

Hengyuan Zhang, Xinrong Chen, Zunhai Su, Xiao Liang, Jing Xiong, Wendong Xu, He Xiao, Chaofan Tao, Wei Zhang, Ruobing Xie, and 1 others. 2026. Beyond outliers: A data-free layer-wise mixed-precision quantization approach driven by numerical and structural dual-sensitivity. *arXiv preprint arXiv:2603.17354*.

Hengyuan Zhang, Zitao Liu, Chenming Shang, Dawei Li, and Yong Jiang. 2025c. A question-centric multi-experts contrastive learning framework for improving the accuracy and interpretability of deep sequential knowledge tracing models. *ACM Transactions on Knowledge Discovery from Data*, 19(2):1–25.

Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Yiyao Yu, Feng Yao, Renliang Sun, Yujiu Yang, and Furu Wei. 2025d. ShifCon: Enhancing non-dominant language capabilities with a shift-based multilingual contrastive framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4818–4841, Vienna, Austria. Association for Computational Linguistics.

Hengyuan Zhang, Yanru Wu, Dawei Li, Sak Yang, Rui Zhao, Yong Jiang, and Fei Tan. 2024a. Balancing speciality and versatility: a coarse to fine framework for supervised fine-tuning large language model. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7467–7509. Association for Computational Linguistics.

Hengyuan Zhang, Shiping Yang, Xiao Liang, Chenming Shang, Yuxuan Jiang, Chaofan Tao, Jing Xiong, Hayden Kwok-Hay So, Ruobing Xie, Angel X Chang, and 1 others. 2025e. Find your optimal teacher: Personalized data synthesis via router-guided multi-teacher distillation. *arXiv preprint arXiv:2510.10925*.

Jason Zhang and Scott Viteri. 2025. Uncovering latent chain of thought vectors in language models. *Preprint*, arXiv:2409.14026.

Jiawei Zhang. 2019. Cognitive functions of the brain: perception, attention and memory. *arXiv preprint arXiv:1907.02863*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, and 3 others. 2024b. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

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.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. 2024d. Interpreting and improving large language models in arithmetic calculation. In *Forty-first International Conference on Machine Learning*.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025f. Multilingual knowledge editing with language-agnostic factual neurons. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 5775–5788. Association for Computational Linguistics.

Zhenyu Zhang, Xiaoxia Wu, Zhongzhu Zhou, Qingyang Wu, Yineng Zhang, Pragaash Ponnusamy, Harikaran Subbaraj, Jue Wang, Shuaiwen Leon Song, and Ben Athiwaratkun. 2025g. Understanding and steering the cognitive behaviors of reasoning models at test-time. *arXiv preprint arXiv:2512.24574*.

Zhenyu Zhang, Shujian Zhang, John Lambert, Wenxuan Zhou, Zhangyang Wang, Mingqing Chen, Andrew Hard, Rajiv Mathews, and Lun Wang. 2025h. Fantastic reasoning behaviors and where to find them: Unsupervised discovery of the reasoning process. *arXiv preprint arXiv:2512.23988*.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025i. Cross-modal information flow in multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19781–19791. Computer Vision Foundation / IEEE.

Zhihao Zhang, Qiaole Dong, Qi Zhang, Jun Zhao, Enyu Zhou, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Mingqi Wu, Yanwei Fu, Tao Ji, Tao Gui, Xuanjing Huang, and Kai Chen. 2025j. Why reinforcement fine-tuning enables mllms preserve prior knowledge better: A data perspective. *Preprint*, arXiv:2506.23508.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024e. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6228–6247. Association for Computational Linguistics.

Zhong Zhang, Bang Liu, and Junming Shao. 2023. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1713, Toronto, Canada. Association for Computational Linguistics.

Delong Zhao, Qiang Huang, Di Yan, Yiqun Sun, and Jun Yu. 2025a. Partially shared concept bottleneck models. *arXiv preprint arXiv:2511.22170*.

Dezhi Zhao, Xin Liu, Xiaocheng Feng, Hui Wang, and Bing Qin. 2025b. Probing and boosting large language models capabilities via attention heads. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28518–28532, Suzhou, China. Association for Computational Linguistics.

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

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. 2025c. LLMs encode harmfulness and refusal separately. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, pages 15296–15319. Curran Associates, Inc.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. 2025d. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations*.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025e. Steering knowledge selection behaviours in llms via sae-based representation engineering. 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 5117–5136.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024c. Analysing the residual stream of language models under knowledge conflicts. *CoRR*, abs/2410.16090.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. 2022. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024. LANDeRMT: Dectecting and routing language-aware neurons for selectively finetuning LLMs to machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148, Bangkok, Thailand. Association for Computational Linguistics.

A Discussion and Challenges

While the actionable applications discussed in previous sections demonstrate the encouraging potential of MI, it is crucial to recognize that it is not a universal solution. There remain important challenges and boundary conditions that may limit its future scalability, reliability, and broader practical impact. We discuss these fundamental points below.

Scalability Constraints and Granularity MI remains difficult to scale beyond low-level components (Kharlapenko et al., 2025; Nikankin et al., 2025a). While individual neurons or learned features are increasingly well-characterized (Duan et al., 2025; Bricken et al., 2023), identifying higher-level computational structures still relies heavily on manual inspection (He et al., 2024b; Marks et al., 2025; Yao et al., 2024a; Lindsey et al., 2025; Nguyen et al., 2025c). Although recent work has made progress toward automation (Conmy et al., 2023; Hanna et al., 2024), current methods often require substantial human intervention and do not robustly generalize across prompts, tasks, or models (Prakash et al., 2024; Hanna et al., 2025; Li et al., 2025h). Crucially, as model sizes scale beyond 100B parameters, the computational cost of fine-grained causal localization grows prohibitively, forcing a fundamental trade-off between localization granularity and computational feasibility. Prominent approaches such as SAEs rely on training surrogate models, introducing costs that grow with model size and feature dimensionality. Precisely attributing behavior to individual components would in principle require exhaustive interventions, but modern LLM scale renders such causal tracing computationally infeasible (Zhang and Nanda, 2023; Hanna et al., 2024). As a result, most analyses (Nanda, 2023; Syed et al., 2024; Yu and Ananiadou, 2024c; Ameisen et al., 2025) operate at coarser granularities or rely on heuristic approximations.

Distributed Mechanisms vs. Localized Sparsity Current mechanistic analyses often face a fundamental trade-off between sparsity and completeness of representation (Gao et al., 2025b; Pach et al., 2025). Many interpretability methods aim to force the model's internal representations into a small set of monosemantic components to make interpretation tractable. However, aggressively enforcing sparsity may prune or obscure components that are genuinely part of the true mechanism but do not fit a sparse pattern. This leads to a tension: methods that induce sparsity can improve interpretability but risk overlooking distributed or "inactive" subcomponents of genuine mechanisms. Acknowledging scenarios where localization may be fundamentally limited, and developing metrics that balance sparsity and mechanistic completeness, remains an open challenge.

Intervention Robustness and Side Effects Interventions informed by MI, such as model editing or steering, often lack robustness and predictability (Yin et al., 2024; Wang et al., 2025b). Changes intended to modify a specific behavior can introduce unintended side effects on other tasks or domains, raising concerns about generalization and reliability (Jiang et al., 2024b; Zhang et al., 2024a, 2025e; Hsueh et al., 2024; Xu et al., 2025b; Da Silva et al., 2025; Braun et al., 2025; Zhang et al., 2025j). For instance, Yu and Ananiadou (2025) demonstrate that modifying a very small number of neurons can lead to substantial degradation in overall language performance. The need for accurate target localization and steering methods that avoid collateral behavioral disruption remains a central technical challenge.

Evaluation Protocols for Actionable MI A critical gap exacerbating these challenges is the lack of robust evaluation frameworks to assess the faithfulness of localization and explanation methods (Miller et al., 2024). Although some benchmarks (Mueller et al., 2025; Parrack et al., 2025; Nguyen et al., 2025b; Wu et al., 2025c; Karvonen et al., 2025) have been proposed, there remains no consensus on metrics that can determine whether an identified component truly corresponds to the underlying causal mechanism. In the absence of reliable ground truth at the mechanism level, rigorous validation and comparison of MI methods is inherently challenging.

To address the need for standardized actionable-MI evaluation, we preliminarily propose a minimal evaluation framework (as shown in Table 2) that covers three distinct task settings: math reasoning, safety, and knowledge editing. For each setting, we explicitly define (1) the features to localize or steer, (2) the primary evaluation metric, (3) essential side-effect checks, and (4) representative benchmark datasets.

This framework facilitates rigorous, multi-faceted evaluation. For example, to evaluate a

Capability	Feature to Localize/ Steer	Primary Metric	Side Effect Check	Example Datasets
Math Reasoning	Arithmetic and Reflection Features	Accuracy, Reflection Token Ratio	General Capabilities	GSM8K, AIME25
Safety	Refusal Features	Refusal Rate, Attack Success Rate	Over-Refusal	SORRY-Bench, OR-Bench
Knowledge	Factual Associations	Edit Success Rate	Locality, Fluency	CounterFact, ZsRE

Table 2: A minimal evaluation benchmark framework for actionable MI. This framework combines primary success metrics with essential side effect checks to ensure a holistic assessment of any intervention.

localizing method targeting reflection features in mathematical reasoning, one might first identify candidate features and then apply targeted ablation. The resulting change in reflection token frequency on benchmarks like AIME25 can then serve as a quantitative measure of causal faithfulness. Similarly, to rigorously compare steering interventions, one can apply different steering methods to the exact same localized features and measure differences in task accuracy or behavioral shifts. For instance, researchers can directly compare the efficacy of (1) amplifying a reflection feature via *Amplitude Manipulation*, (2) injecting it as a steering vector via *Vector Arithmetic*, and (3) performing localized weight updates via *Targeted Optimization*.

Furthermore, measuring unintended consequences is a first-class requirement. For each task setting, we design dedicated side-effect checks to evaluate the collateral impact of different steering methods:

- **Math Reasoning:** Interventions aimed at enhancing reflection or arithmetic features should not degrade performance in other domains. Therefore, after steering a reflection-related feature, the model must also be evaluated on general knowledge and safety benchmarks to ensure its broader capabilities remain intact.
- **Safety:** The objective is not only to increase the refusal rate on harmful prompts (measured by benchmarks such as SORRY-Bench (Xie et al., 2025)), but also to avoid over-refusal—that is, rejecting benign or harmless prompts. This represents a critical trade-off, as overly aggressive safety interventions may significantly harm the model's general helpfulness. Benchmarks such as OR-Bench (Cui et al., 2024) are specifically designed to quantify this negative side effect.
- **Knowledge Editing:** For knowledge editing, *locality* is the primary side-effect metric. It measures whether editing a specific fact (e.g., modifying a president's name) unintentionally alters unrelated knowledge in the model (Zhang et al., 2024b). *Fluency* is additionally essential, as it evaluates whether the model's generative quality degrades post-edit, such as by becoming repetitive or producing ungrammatical outputs.

B Future Directions

Looking forward, several directions appear particularly promising for advancing MI.

Broadening the Architectural Scope While our formalization in §2 is primarily centered on decoder-only Transformer LLMs, an important future direction is to extend the "Locate, Steer, and Improve" framework to broader model architectures. Although our pipeline is defined in the decoder-only setting, many of the localizing and steering methodologies we categorize are general in spirit and can, in principle, be transferred to other architectures once their corresponding interpretable objects are identified. For example, compared with decoder-only dense models, MoE architectures introduce an additional routing mechanism, making router states, routing activations, routing decisions, and expert activations natural functional objects beyond standard neural activations (Shazeer et al., 2017; Fedus et al., 2022; Zhou et al., 2022; Olson et al., 2025). This suggests that localization tools developed for dense models, such as *Magnitude Analysis*, can be extended from ordinary neural activations to router activations, in order to identify influential routes or experts and subsequently steer behavior through interventions on the corresponding expert modules (Wang et al., 2025a; Olson et al., 2025; Liu et al., 2024a). Similarly, MLLMs augment text-only LLMs with vision encoders and cross-modal interaction modules (Liu et al., 2023a; Li et al., 2023; Alayrac et al., 2022; Lin et al., 2025). Their interpretable objects

therefore go beyond language-side activations to include visual token representations, image-patch attention maps, cross-modal attention patterns, and modality-level information-flow pathways, which can likewise be analyzed using existing tools in our taxonomy. For instance, *Magnitude Analysis* can be applied to visual or multimodal activations, including those in the vision encoder and multimodal attention stack, while *Vocabulary Projection* can be adapted to inspect how visual representations align with linguistic concepts (Neo et al., 2025; Yu and Ananiadou, 2024d; Bi et al., 2025; Zhang et al., 2025i). At the same time, fully realizing actionable MI in these architectures will likely require more than direct adaptation: future work should develop architecture-specific localizing and steering techniques that explicitly account for distinctive structural properties such as sparse routing in MoE models and tightly entangled cross-modal representations in MLLMs (Lin et al., 2025).

Integration with Cognitive Science A key priority for mechanistic interpretability is to move from isolated, low-level analyses toward integrated, system-level explanations. Most existing MI work focuses on task-specific and localized mechanisms, such as knowledge neurons, safety-related neurons, arithmetic heads, or specific task circuits (Yao et al., 2024b; Chen et al., 2024a; Xiao et al., 2025a; Zhang et al., 2024d; Gurgurov et al., 2025a; Li et al., 2025d). While informative, these approaches offer limited insight into how models organize computation more broadly (Zhao et al., 2024a). In contrast, cognitive science conceptualizes cognition through higher-order organizations. This includes broad processing paradigms like System 1 and System 2 reasoning (Li et al., 2025f), as well as distinct functional subsystems governing attention, memory, language, and executive control (Morgan and Gilliland, 1927; Gruber and Goschke, 2004; Gruszka and Matthews, 2010; Zhang, 2019). Future research should draw explicit parallels between MI-discovered mechanisms and these established cognitive architectures. For instance, exploring whether specific localized circuits function analogously to working memory or attention control, could reveal if LLM internal structures exhibit organizational principles analogous to human cognitive systems (Geiger et al., 2025). Furthermore, this integration can foster a bidirectional synergy: MI findings might inform cognitive theories of human reasoning and language processing, while established cognitive frameworks provide a robust blueprint for understanding and evaluating LLMs.

Theoretical Foundations In parallel, stronger theoretical foundations are needed. Connecting internal representations to principles from cognitive science (Davies and Khakzar, 2024; Wulff and Mata, 2025; Zhang et al., 2025c; Ren et al., 2025) or information theory (Conklin and Smith, 2024) may help unify disparate MI findings and reduce reliance on ad-hoc interpretations. A principled framework could also clarify what kinds of internal structures should be expected in large-scale models and why (Kendiukhov, 2025).

From Interpretation to Interpretable Design Finally, an emerging direction is the progression from interpretation to intervention and, ultimately, model design. Insights from MI are increasingly used not only to explain behavior, but also to edit, steer, or modularize models. This direction connects naturally to earlier work on intrinsically interpretable models, such as Concept Bottleneck Models (Ismail et al., 2024; Sun et al., 2024a; Shang et al., 2024a,b; Tan et al., 2024b; Hu et al., 2025; Zhao et al., 2025a) and Weight-sparse transformers (Gao et al., 2025b), which enforce transparency through architectural constraints. However, despite their interpretability benefits, such models typically underperform black-box architectures on large-scale, complex tasks (Srivastava et al., 2024). Looking forward, a key challenge is to bridge this gap by designing interpretable backbone architectures that can serve as viable alternatives to transformers, achieving interpretability by construction while maintaining performance comparable to state-of-the-art black-box models. In this sense, interpretability-informed design may move beyond post-hoc analysis toward fundamentally more controllable, customizable, and transparent model architectures.

C Paper Outline

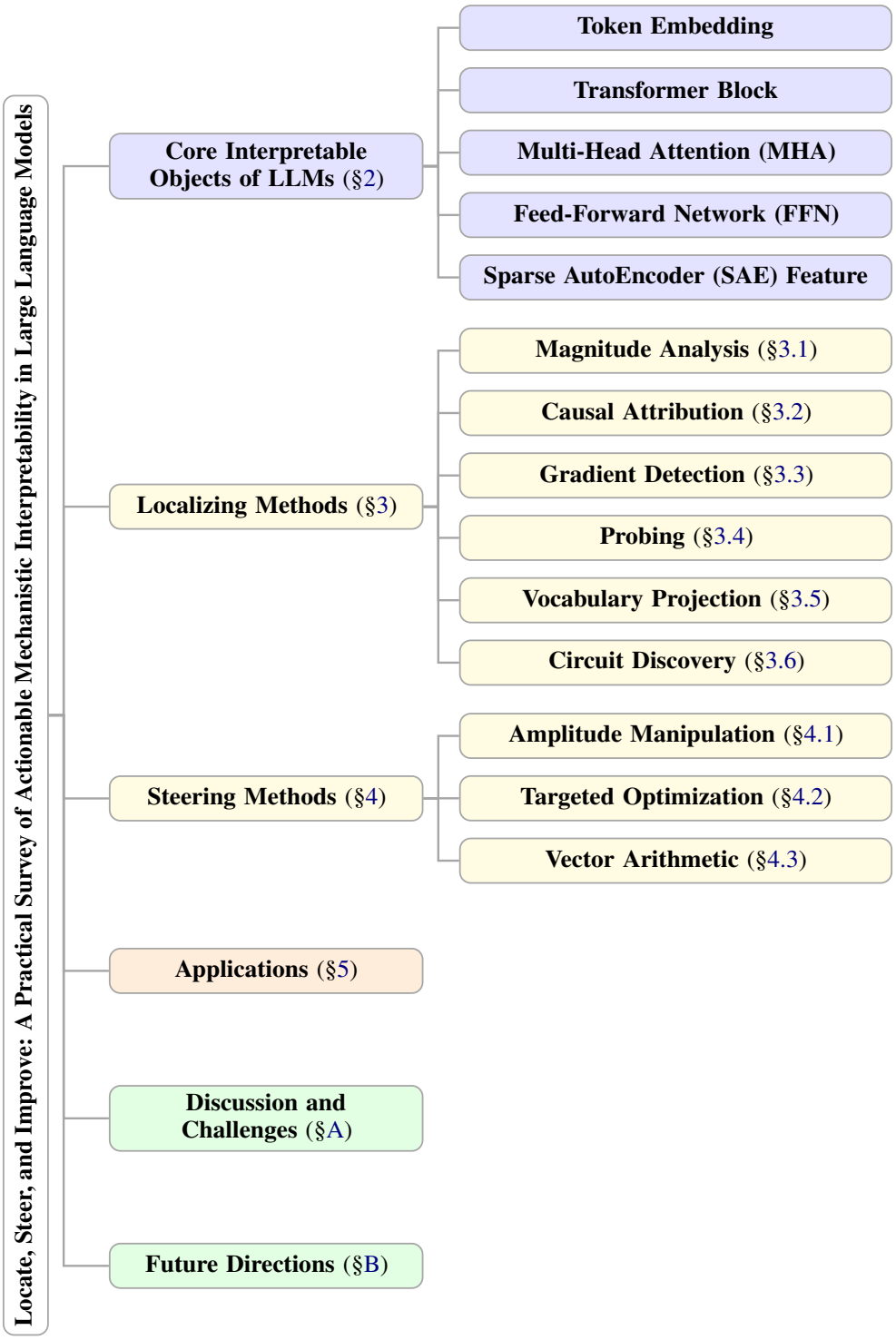

Figure 2: Outline of this survey (Part I). Applications are shown separately in Fig. 3.

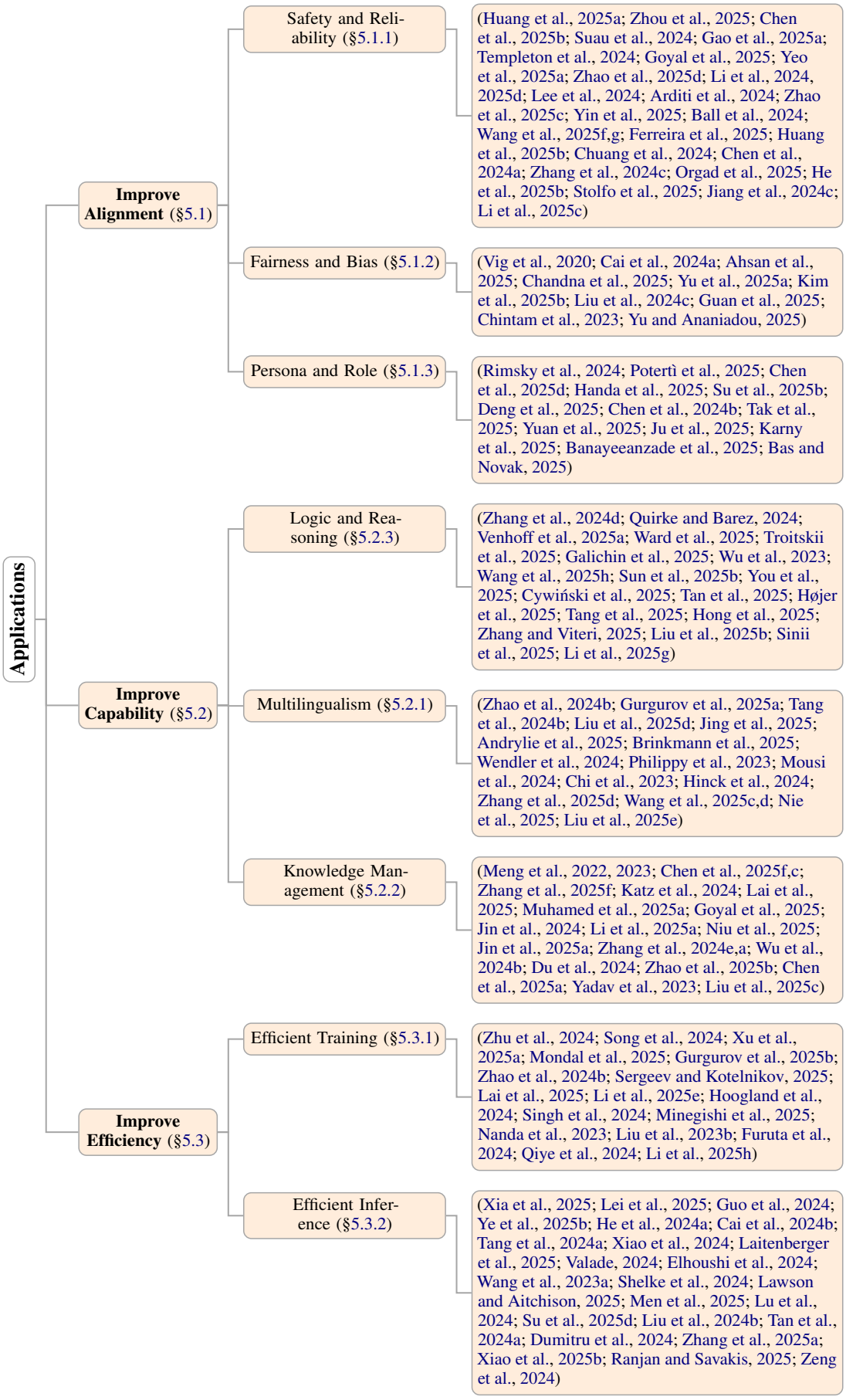

Figure 3: Outline of this survey (Part II: Applications). The remaining sections are shown in Fig. 2.

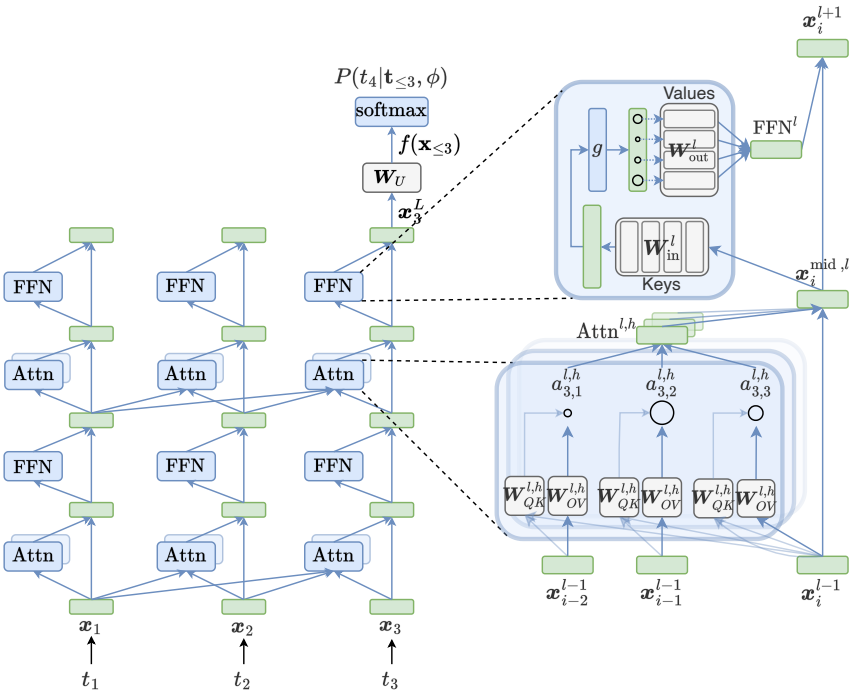

Figure 4: The schematic of information flow within a standard Transformer block. The Residual Stream ($\mathbf{x}_l$) serves as the backbone, while Attention and FFN act as additive branches that read from and write to this stream. Based on the figure from Ferrando et al. (2024).

D Details of Interpretable Objects of LLM

In this section, we further describe the details of the internal objects, with a focus on the decoder-only architecture—the predominant framework for contemporary LLMs (Dubey et al., 2024; Team et al., 2024; Abdin et al., 2024; Anthropic, 2024; xAI, 2025; DeepSeek-AI et al., 2025). We do this by providing a mechanistic breakdown of the information flow within Transformer blocks and illustrating the theoretical and practical aspects of Sparse Autoencoders (SAEs).

D.1 Information Flow in Transformer Blocks

To effectively locate and steer internal objects, one must first understand how information propagates through the model. As illustrated in Fig. 4, the Transformer architecture is fundamentally built around the **Residual Stream**, acting as the central communication channel or "highway" (Elhage et al., 2021; Bricken and Pehlevan, 2021; Wang et al., 2023b; Ferrando and Voita, 2024; Voita et al., 2024; Liu et al., 2025a).

Mechanistically, the residual stream $\mathbf{x}_l$ at layer l is updated iteratively by sub-layers, we view the components as performing specific operations on this stream:

- **Attention Heads (The Router):** Attention heads "read" information from the residual stream of previous tokens (via subspace projections $\mathbf{W}_Q^{l,h}, \mathbf{W}_K^{l,h}$) and "write" attended information to the current position (via $\mathbf{W}_V^{l,h}, \mathbf{W}_O^{l,h}$). They primarily manage information routing and contextual dependencies (Elhage et al., 2021; Olsson et al., 2022; Voita et al., 2019; Feng and Steinhardt, 2023; Men et al., 2024).

- **FFN (The Processor):** FFN operate independently on each token position. They act as "Key-Value" memories where the first layer projects the stream into a high-dimensional state (Knowledge Keys) and the second layer writes the retrieved knowledge back to the stream (Geva et al., 2021, 2022; Dai et al., 2022).

This additive structure, $\mathbf{x}_{l+1} = \mathbf{x}_l + \text{MHA}(\mathbf{x}_l) + \text{FFN}(\mathbf{x}_l)$, implies that features in the residual stream are linear combinations of outputs from all previous components, justifying the use of methods like SAEs to disentangle them.

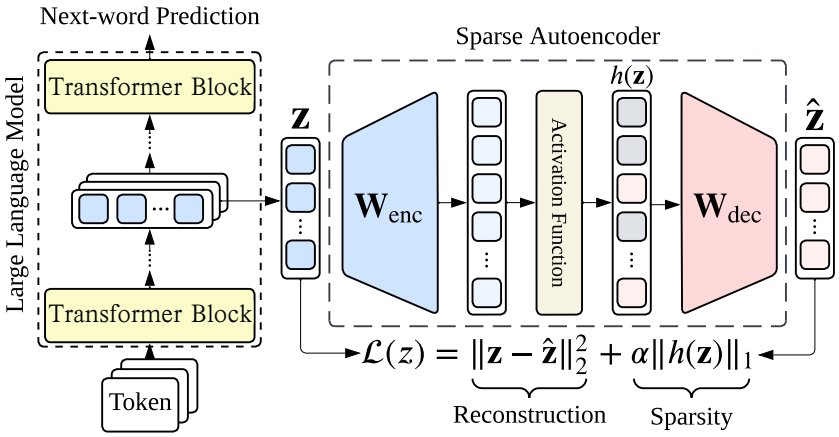

Figure 5: The framework of Sparse Autoencoders (SAEs). The SAE acts as a "microscope" for the LLM, expanding dense representation into a sparse, overcomplete set of interpretable features via an encoder-decoder architecture. Based on the figure from Shu et al. (2025).

D.2 Extended Analysis of Sparse Autoencoders (SAEs)

While the main text introduces the definition of SAEs, here we discuss the underlying motivation and practical training considerations.

The Necessity: Resolving Superposition. The motivation for SAEs stems from the challenge of *superposition*, a phenomenon where neural networks represent more features than they have physical neurons by encoding them as nearly orthogonal directions in the activation space (Elhage et al., 2022). This causes *polysemanticity*, where a single neuron activates for unrelated concepts. SAEs resolve this by projecting low-dimensional dense activations into a higher-dimensional sparse latent space, effectively "unpacking" the superposition. Crucially, this decomposition allows researchers to steer model behavior by targeting these granular features (Templeton et al., 2024; Lieberum et al., 2024; He et al., 2025b; Cho and Hockenmaier, 2025; Li et al., 2025c), transforming opaque vectors into an actionable vocabulary.

Framework and Visualization. Fig. 5 illustrates the standard SAE framework applied to a target LLM. The SAE is an independent module attached to a specific object of a frozen LLM—in fact, SAEs can be applied to nearly all internal objects within LLMs, including $\mathbf{s}^l$, $\mathbf{x}^l$, $\mathbf{x}^{l,mid}$, $\mathbf{h}^{l,h}_{attn}$, $\mathbf{h}^l_{attn}$, and $\mathbf{h}^l_{ffn}$.

The **Expansion Factor** (ratio of d_{SAE} to d_{model}) is a critical hyperparameter. To capture the vast number of features hidden in superposition, d_{SAE} is typically set to be $16\times$ to $128\times$ larger than the model dimension (Cunningham et al., 2023; Templeton et al., 2024; Bloom, 2024; Ghilardi et al., 2024; Mudide et al., 2024; Lieberum et al., 2024; He et al., 2024c).

Training Challenges and Solutions. Training SAEs presents unique challenges discussed in recent surveys (Shu et al., 2025):

- **Dead Latents:** A major issue where many feature neurons never activate during training, wasting capacity. Techniques such as *ghost gradients* or periodic *resampling* of dead neurons are commonly employed to mitigate this (Bricken et al., 2023).
- **Feature Absorption:** Sometimes broadly activating features suppress more specific ones. Advanced architectures like *Gated SAEs*, *Top-K SAEs*, *BatchTopK SAEs*, *JumpReLU-SAEs* and *Binary-SAEs* have been proposed to improve feature quality and reconstruction fidelity (Gao et al., 2024; Rajamanoharan et al., 2024a; Bussmann et al., 2024; Rajamanoharan et al., 2024b; Cho et al., 2025).

Pre-trained Resources. To reduce barriers to interpretability research, several high-quality pre-trained SAE suites have been made publicly available. For instance, Lieberum et al. (2024) released *Gemma Scope* (which provides SAEs for all layers of the Gemma2 model), while He et al. (2024c) introduced *Llama Scope* (covering every layer of Llama3 model). Similarly, Templeton et al. (2024) demonstrated "Golden Gate Claude" features.

E Details of Methods

E.1 Details of Vocabulary Projection

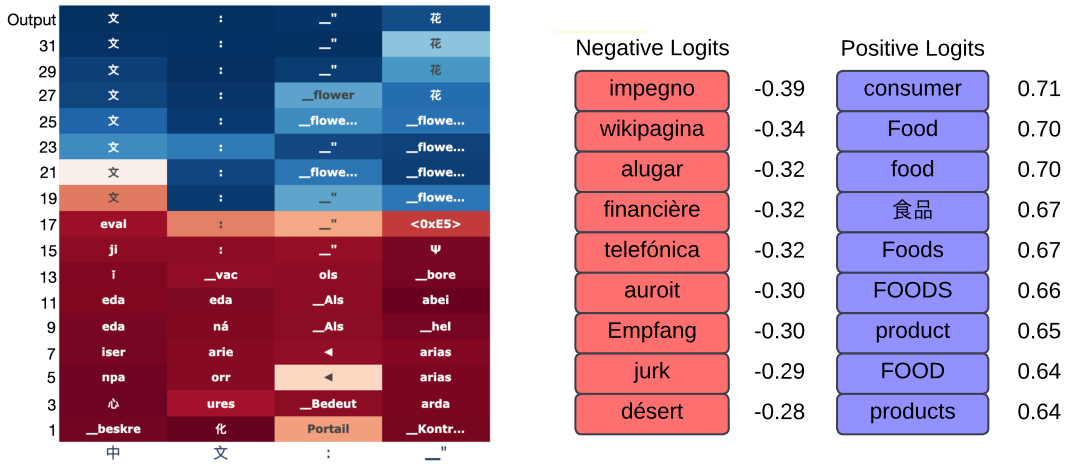

Figure 6: (a) Projecting residual streams reveals the layer-wise evolution of latent concepts, showing an English-centric bottleneck in multilingual tasks (Wendler et al., 2024). (b) Projecting SAE decoder weights identifies the semantic meaning of sparse features (e.g., a "food" feature) by identifying top-ranked tokens (Shu et al., 2025). Based on figures from Wendler et al. (2024) and Shu et al. (2025).

In this appendix, we detail classic case studies demonstrating how Vocab Projection (Logit Lens) serves as a versatile tool for interpreting various internal components of LLMs, ranging from global residual streams (Pal et al., 2023; Belrose et al., 2023; Wendler et al., 2024; Shai et al., 2024; Jiang et al., 2024a; Cancedda, 2024; Takatsuki et al., 2025; Yugeswardeenoo et al., 2025; Phukan et al., 2025) to specific attention heads (Sakarvadia et al., 2023; Wang et al., 2023b; Yu and Ananiadou, 2024d; Jiang et al., 2025; Kim et al., 2025a; Wang, 2025; Bahador, 2025), neurons (Geva et al., 2021; Huo et al., 2024; Yu and Ananiadou, 2025; Kargaran et al., 2025; Shao et al., 2025) and sparse features (Bricken et al., 2023; Lawson et al., 2024; Shu et al., 2025; Arad et al., 2025; Dreyer et al., 2025; Venhoff et al., 2025b; Muhamed et al., 2025b; Wu et al., 2025a; Gur-Arieh et al., 2025).

Layer-wise Analysis Example: The Latent Language of Transformers Vocab Projection applied to residual streams ($\mathbf{x}^l$) reveals the evolution of latent concepts. A prominent example is the study by Wendler et al. (2024) on multilingual models. As illustrated in Fig. 6 (a), analyzing a translation task reveals three distinct processing phases: initial layers focus on the surface form of the input language; middle layers process semantics in an abstract, "English-centric" concept space (even for non-English tasks); and final layers rotate back to the target language output. This confirms that English serves as an internal pivot for abstract reasoning in multilingual LLMs.

Head-wise Analysis Example: Mechanisms of Indirect Object Identification Applying Vocab Projection to the output of individual attention heads ($\mathbf{h}_{\text{attn}}^{l,h}$) allows researchers to determine exactly what information a head "writes" into the residual stream. Wang et al. (2023b) utilized this technique to reverse-engineer the Indirect Object Identification (IOI) task (e.g., completing sentences like "When Mary and John went to the store, John gave a drink to *Mary*"). The authors identified specific functional heads by checking if their outputs projected strongly to the correct name ("Mary"): "Name Mover Heads" (e.g., Head 9.9), which explicitly write the correct name ("Mary") vector, and "Negative Name Mover Heads" (e.g., Head 10.7), which attend to the name but write a negative vector to suppress it. This mechanistic distinction is only visible through Vocab Projection.

Neuron-wise Analysis Example: FFNs as Key-Value Memories Geva et al. (2021) demonstrate that FFNs operate as key-value memories, where value weights ($\mathbf{W}_{\text{down}}$) represent output distributions. By

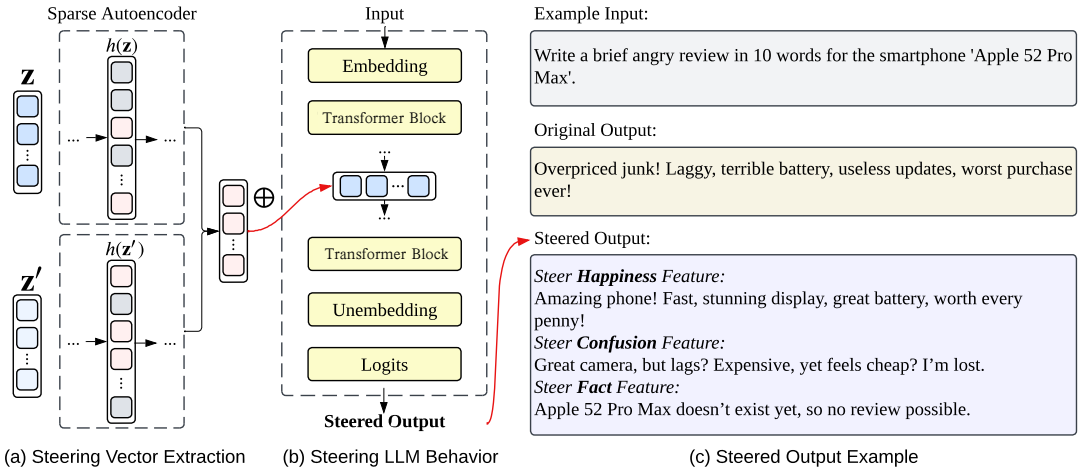

Figure 7: The pipeline for steering LLMs using SAE Features. (a) **Steering Vector Extraction**: The target steering vector is derived by analyzing a set of prompts to identify features that distinguish a concept-rich state $\mathbf{z}'$ from a neutral state $\mathbf{z}$. The steering vector is computed as the weighted sum of these identified SAE Decoder columns. (b) **Steering LLM Behavior**: This aggregated vector is injected into the Transformer's residual stream via vector addition. (c) **Steered Output Example**: Empirical results showing how steering specific features (e.g., Happiness, Confusion) drastically alters the model's generation style even when the original prompt implies a negative sentiment. Based on the figure from Shu et al. (2025).

projecting specific column vectors of $\mathbf{W}_{\text{down}}$ into the vocabulary, they reveal that individual neurons often promote semantically related clusters of tokens (e.g., a single neuron boosting "press", "news", and "media"). This suggests that FFN updates refine model predictions by composing these pre-learned semantic distributions layer by layer.

Feature-wise Analysis Example: Interpreting SAE Features For Sparse Autoencoders (SAEs), output-based explanations leverage the decoder weights to interpret monosemantic features (Shu et al., 2025). By computing the logits contribution $\mathbf{l}_j = \mathbf{f}_j \mathbf{W}_U$ for a feature vector $\mathbf{f}_j$, one can identify the top-ranked tokens. As shown in Fig. 6 (b), a feature whose projection yields high positive logits for tokens like "Food" and "food" is interpreted as encoding a "food" concept, directly grounding the sparse feature in human-understandable semantics.

E.2 Details of Vector Arithmetic

In the main text, we introduced *Vector Arithmetic* as a method to steer model behavior by injecting a direction $\mathbf{v}$ into hidden states. Here, we detail the two primary mechanisms for deriving this steering vector: *Contrastive Activation Means* (Rimsky et al., 2024; van der Weij et al., 2024; Lu and Rimsky, 2024; Postmus and Abreu, 2024; Turner et al., 2024; Su et al., 2025a; Hegazy et al., 2025; Sharma and Raman, 2025), and *SAE Features* (Bayat et al., 2025; Shu et al., 2025; Weng et al., 2025; He et al., 2025b; Soo et al., 2025; He et al., 2025a; Bhattacharyya and Rooshenas, 2025; Goyal et al., 2025).

Deriving Vectors via Contrastive Activation Means This method, often referred to as "Mass-Mean Shift" or "Activation Addition", assumes that a concept can be isolated by comparing the model's internal states across opposing contexts. Formally, let $\mathcal{D}^+$ be a set of prompts eliciting the target behavior (e.g., sycophantic responses) and $\mathcal{D}^-$ be a set of prompts eliciting the opposing behavior (e.g., non-sycophantic responses). We collect the intermediate activations $\mathbf{x}^l$ at a chosen layer l for both sets. The steering vector $\mathbf{v}$ is calculated as the difference between the centroids of these two distributions:

$$\mathbf{v} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^- = \frac{1}{|\mathcal{D}^+|} \sum_{\mathbf{x}_i \in \mathcal{D}^+} \mathbf{x}_i^l - \frac{1}{|\mathcal{D}^-|} \sum_{\mathbf{x}_j \in \mathcal{D}^-} \mathbf{x}_j^l \tag{6}$$

During inference, this vector represents the "direction of sycophancy." By adding $\alpha \cdot \mathbf{v}$ to the residual stream, we shift the model's current state towards the centroid of the positive behavior, effectively inducing the target trait without altering the model's weights.

Deriving Vectors via SAE Features Sparse Autoencoders (SAEs) provide a more precise method to isolate steering vectors by disentangling the residual stream into monosemantic features. As illustrated in Fig. 7, the process begins with **Feature Identification via Comparison** (Fig. 7a). Normally, researchers employ a set of prompts containing the target concept (e.g., "Happiness") versus a set lacking it, producing two aggregated latent states: a concept-rich state $\mathbf{z}'$ and a baseline state $\mathbf{z}$. By comparing their sparse activations $h(\mathbf{z}')$ and $h(\mathbf{z})$, a subset of relevant features $\mathcal{J}$ that exhibit significant activation differences is identified. The steering vector $\mathbf{v}$ is then constructed as the weighted sum of the corresponding SAE decoder columns: $\mathbf{v} = \sum_{j \in \mathcal{J}} \delta_j \cdot \mathbf{W}_{\text{dec}}[:, j]$, where the scalar weight δ_j represents the activation difference of feature j between the two states. Finally, during the **Steering Intervention** (Fig. 7b), this aggregated vector is injected into the residual stream ($\hat{\mathbf{x}} = \mathbf{x} + \alpha \cdot \mathbf{v}$). As shown in Fig. 7(c), this precise manipulation can dramatically alter generation styles, such as transforming a negative review into a positive one by amplifying the "Happiness" features.

E.3 Details of Causal Attribution

In the main text, we introduced Patching as a method to isolate information pathways by swapping internal states (Vig et al., 2020; Meng et al., 2022, 2023; Goldowsky-Dill et al., 2023; Yeo et al., 2025b; Ravindran, 2025). Here, we detail a representative implementation of this technique: *Causal Tracing*, as proposed by Meng et al. (2022).

Example: Locating Factual Associations via Restoration Patching Meng et al. (2022) utilize a specific form of patching, often called "denoising" or "restoration" patching, to localize where factual knowledge (e.g., knowing that "The Space Needle" is located in "Seattle") is stored within the model. As illustrated in Fig. 8, the process involves three key steps:

1. **Corrupted Run (Intervention):** First, the authors effectively erase the specific knowledge from the model's computation. They create a *corrupted input* by adding Gaussian noise to the embeddings of the subject tokens (e.g., "Space Needle"). In this corrupted state, the model's probability of predicting the correct object ("Seattle") drops significantly.

2. **Patched Run (Restoration):** The core operation is to systematically restore specific internal states from the original *clean run* into this corrupted run. For a specific layer l and token position i, the method copies the clean hidden activation $\mathbf{h}_i^l$ and pastes it into the corrupted computation graph.

3. **Effect Measurement:** The causal effect of the patched state is quantified by the **Indirect Effect (IE)**, which measures how much of the original target probability ("Seattle") is recovered by this restoration. A high IE score indicates that the patched state at (l, i) carries critical information for the fact.

Insights from Causal Tracing By applying this patching method across all layers and token positions, Meng et al. (2022) generated the heatmap results shown in Fig 8. This analysis revealed two distinct localized mechanisms responsible for factual recall:

- **Early Site (Subject Tokens):** A strong causal effect is found in the *early MLP layers* corresponding to the subject tokens (e.g., "Space Needle"). This suggests that the model decodes the specific entity and retrieves its properties early in the network.

- **Late Site (Last Token):** A second peak of causal importance appears in the *late Attention layers* at the final token position (e.g., "downtown"). This indicates that the retrieved information is transported and processed by attention heads just before prediction.

E.4 Details of Gradient Detection

Gradient Detection can be viewed as a first-order (Taylor) approximation to intervention effects, and Integrated Gradients (IG) as a path-integrated extension that satisfies completeness relative to a baseline. This appendix follows canonical gradient saliency and attribution work and related variants/caveats

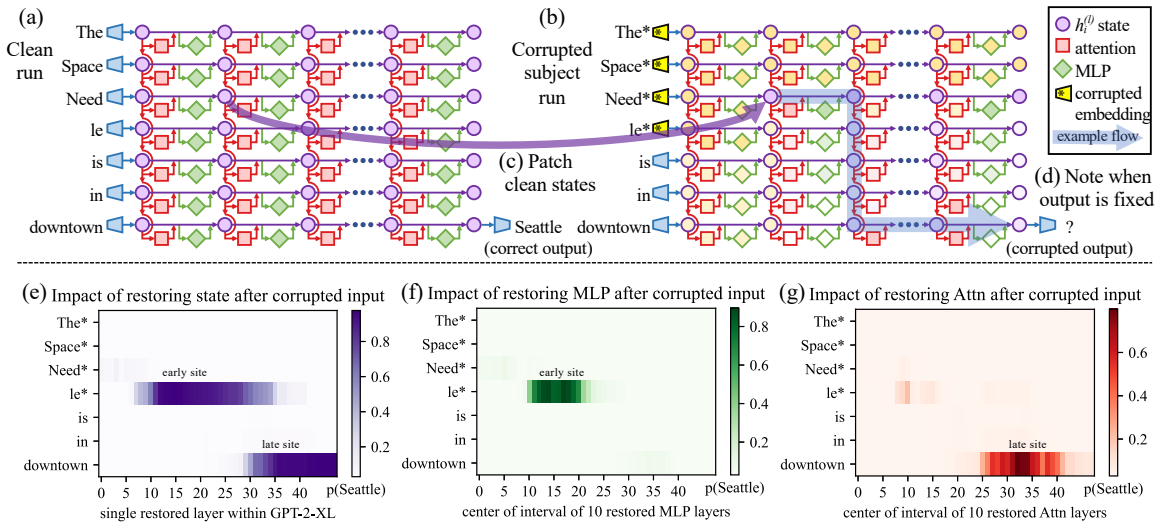

Figure 8: Overview of Causal Tracing via Patching (Meng et al., 2022). The method identifies critical internal states by creating a corrupted run (noising the subject "Space Needle") and systematically restoring clean states to see which ones recover the prediction "Seattle". The heatmap results reveal that factual information is processed in early MLP layers at the subject position and later transferred to the final token via attention. Based on the figure from Meng et al. (2022).

(Sundararajan et al., 2017; Smilkov et al., 2017; Shrikumar et al., 2017; Yin and Neubig, 2022; Wang et al., 2024d).

E.4.1 Taylor Expansion Derivation of $s_j = \nabla_{o_j} F(x)^\top o_j$

Fix an input x and an internal object $o_j = o_j(x) \in \mathbb{R}^d$ (e.g., an embedding, residual vector, head output, or parameter block). Consider an additive modification $o_j \longrightarrow o_j + \Delta o_j$, and study the local output change of a scalar target $F(\cdot)$ with respect to o_j.

Taylor Expansion With Remainder Assume F is twice differentiable in a neighborhood of o_j. The second-order Taylor expansion around o_j yields

$$F(o_j + \Delta o_j) = F(o_j) + \nabla_{o_j} F(o_j)^\top \Delta o_j + \frac{1}{2} \Delta o_j^\top \mathbf{H}_{o_j} F(o_j + \xi \Delta o_j)\, \Delta o_j, \quad \text{for some } \xi \in (0,1), \tag{7}$$

where $\mathbf{H}_{o_j} F$ is the Hessian with respect to o_j. Therefore,

$$F(o_j + \Delta o_j) - F(o_j) = \nabla_{o_j} F(o_j)^\top \Delta o_j + \underbrace{\frac{1}{2} \Delta o_j^\top \mathbf{H}_{o_j} F(o_j + \xi \Delta o_j)\, \Delta o_j}_{\text{Higher-Order Remainder}}. \tag{8}$$

Gradient Detection corresponds to using the first-order term as a fast proxy and ignoring the remainder.

Define the 1D restriction along direction Δo_j: $g(\alpha) = F(o_j + \alpha \Delta o_j), \quad \alpha \in [0,1]$. By the chain rule, $g'(0) = \nabla_{o_j} F(o_j)^\top \Delta o_j$, hence this dot product is the directional derivative of F at o_j along Δo_j.

Grad×Input Score A common "local removal" surrogate uses $\Delta o_j = -o_j$, giving

$$F(o_j - o_j) - F(o_j) \approx -\nabla_{o_j} F(o_j)^\top o_j. \tag{9}$$

This motivates the signed score $s_j(x) = \nabla_{o_j} F(x)^\top o_j$, with $|s_j(x)|$ often used when only magnitude matters. Eq. (8) makes the approximation conditions explicit: curvature (Hessian) and perturbation size control the error.

Scores can be summarized over coordinates and/or over a dataset. For example, a component-wise decomposition is

$$\nabla_{o_j} F(o_j)^\top o_j = \sum_{k=1}^{d} \frac{\partial F}{\partial (o_j)_k}(o_j) \cdot (o_j)_k, \tag{10}$$

and dataset-level aggregation can use $\mathbb{E}_{x \sim \mathcal{D}}[s_j(x)]$, $\mathbb{E}[|s_j(x)|]$, or rank-based statistics.

E.4.2 Integrated Gradients: Definition and Intuition

Single-point gradients can be uninformative under saturation and do not, by themselves, specify a reference for "absence." IG addresses both by attributing $F(o_j) - F(o'_j)$ for a chosen baseline o'_j. Choose a baseline o'_j (same shape as o_j) and define the straight-line path $\gamma(\alpha) = o'_j + \alpha(o_j - o'_j), \qquad \alpha \in [0, 1]$.

Definition IG assigns an attribution to each coordinate k:

$$\mathrm{IG}_k(o_j; o'_j) = (o_j - o'_j)_k \int_0^1 \frac{\partial F(\gamma(\alpha))}{\partial \gamma_k}\, d\alpha. \tag{11}$$

Because $\frac{d}{d\alpha} F(\gamma(\alpha)) = \nabla_\gamma F(\gamma(\alpha))^\top (o_j - o'_j)$, we have

$$\sum_{k=1}^{d} \mathrm{IG}_k(o_j; o'_j) = \int_0^1 \frac{d}{d\alpha} F(\gamma(\alpha))\, d\alpha = F(o_j) - F(o'_j). \tag{12}$$

Thus IG decomposes the baseline-to-input change in F across coordinates of o_j.

Riemann Approximation With m steps,

$$\mathrm{IG}_k(o_j; o'_j) \approx (o_j - o'_j)_k \cdot \frac{1}{m} \sum_{t=1}^{m} \left. \frac{\partial F(\gamma(\alpha))}{\partial \gamma_k} \right|_{\alpha = t/m}. \tag{13}$$

If $o'_j = 0$ and $\nabla F(\gamma(\alpha))$ varies little along the path, the integral is close to the endpoint gradient, recovering a grad×input-like form as a locally linear special case. IG explains $F(o_j) - F(o'_j)$; hence the baseline is part of the explanation specification and should reflect an appropriate notion of "absence" for the object.

E.4.3 Target Function Examples

Gradient Detection is modular in F. Beyond a single logit, common choices include:

- **Loss:** $F(x) = -\log p(y^\star|x)$ or cross-entropy objectives.
- **Logit Margin:** $F(x) = \mathrm{logit}_y(x) - \mathrm{logit}_{y^{\text{foil}}}(x)$.
- **Counterfactual Gap:** $F(x) = \left|\mathrm{logit}_y(x) - \mathrm{logit}_y(x^{cf})\right|$.
- **Distribution Shift:** $F(x) = D_{\mathrm{KL}}(p(\cdot|x) \,\|\, p(\cdot|x^{cf}))$ (or symmetric variants).

E.5 Details of Circuit Discovery

In the main text, we outlined methods for operationalizing *Circuit Discovery*. Here, we elaborate on the *Attribution-based* method, offering a detailed formulation of *Edge Attribution Patching (EAP)* (Syed et al., 2024) and its variant, *EAP with Integrated Gradients (EAP-IG)* (Hanna et al., 2024; Huang et al., 2025a).

Methodological Formulation Constructing a faithful circuit involves identifying a subset of edges $\mathcal{E}_{sub} \subset \mathcal{E}$ that mediate the model's performance on a specific task. Let u be a "sender" node (e.g., an attention head or MLP neuron) and v be a "receiver" node. An edge $e_{u\to v}$ transmits the activation $\mathbf{a}_u$ from u to the input $\mathbf{z}_v$ of v. The goal is to measure the *attribution score* of this edge, which quantifies the change in a task metric $\mathcal{R}$ (e.g., logit difference) when the edge is "patched" or ablated.

1. The Bottleneck of Activation Patching Exact evaluation requires performing an intervention $\mathrm{do}(e_{u\to v})$: replacing the activation along the edge with a counterfactual value (from corrupted input $\mathbf{x}_{corr}$) while keeping the rest of the model conditioned on the clean input $\mathbf{x}_{clean}$. Computing this for all edges requires $|\mathcal{E}|$ separate forward passes, which is computationally prohibitive for large models ($O(|\mathcal{E}|)$).

2. Linear Approximation (Standard EAP) EAP overcomes this bottleneck by approximating the patching effect using a first-order Taylor expansion. Instead of running a new forward pass for each edge, it estimates the effect using gradients computed from a single backward pass. The attribution score $S(u \to v)$ is approximated as the inner product of the *activation difference* at the sender and the *gradient* at the receiver:

$$S_{linear}(u \to v) \approx \underbrace{(\mathbf{a}_u(\mathbf{x}_{clean}) - \mathbf{a}_u(\mathbf{x}_{corr}))}_{\text{Sender Activation Difference}} \cdot \underbrace{\left.\frac{\partial \mathcal{R}}{\partial \mathbf{z}_v}\right|_{\mathbf{x}_{clean}}}_{\text{Receiver Input Gradient}} \tag{14}$$

This factorization allows EAP to compute scores for *all* edges simultaneously using just two forward passes (to get $\mathbf{a}_u$ for both inputs) and one backward pass (to get $\nabla \mathcal{R}$), achieving $O(1)$ efficiency.

3. Addressing Non-Linearity with Integrated Gradients (EAP-IG) While Linear EAP is efficient, it assumes the model behaves linearly between the corrupted and clean states. However, neural networks often exhibit *saturation* or highly non-linear behaviors, where the local gradient at $\mathbf{x}_{clean}$ may vanish or mislead the attribution (e.g., a neuron is already saturated). To address this, **EAP-IG** incorporates the method of *Integrated Gradients* (Sundararajan et al., 2017). Instead of relying on the gradient at a single point, EAP-IG averages the gradients along a linear interpolation path between the corrupted and clean inputs.

Let $\alpha \in [0, 1]$ be an interpolation coefficient. We define an interpolated input sequence:

$$\mathbf{x}_\alpha = \mathbf{x}_{corr} + \alpha \cdot (\mathbf{x}_{clean} - \mathbf{x}_{corr}) \tag{15}$$

The EAP-IG score is computed by integrating the gradients along this path:

$$S_{IG}(u \to v) = (\mathbf{a}_u(\mathbf{x}_{clean}) - \mathbf{a}_u(\mathbf{x}_{corr})) \cdot \int_{\alpha=0}^{1} \left.\frac{\partial \mathcal{R}}{\partial \mathbf{z}_v}\right|_{\mathbf{x}_\alpha} d\alpha \tag{16}$$

Practically, this integral is approximated using a Riemann sum over n discrete steps (e.g., $n = 50$):

$$S_{IG}(u \to v) \approx (\mathbf{a}_u(\mathbf{x}_{clean}) - \mathbf{a}_u(\mathbf{x}_{corr})) \cdot \frac{1}{n} \sum_{k=1}^{n} \left.\frac{\partial \mathcal{R}}{\partial \mathbf{z}_v}\right|_{\mathbf{x}_{\frac{k}{n}}} \tag{17}$$

Summary of the Discovery Workflow Using EAP-IG involves the following steps:

1. **Activation Collection**: Run forward passes on $\mathbf{x}_{clean}$ and $\mathbf{x}_{corr}$ to compute the activation difference $\Delta \mathbf{a}_u$ for every sender node.
2. **Integrated Gradient Computation**: Run n forward and backward passes on interpolated inputs $\mathbf{x}_{k/n}$ to compute the average gradient $\overline{\nabla \mathbf{z}_v}$ w.r.t the input of every receiver node.
3. **Edge Scoring**: Compute the score for every edge $u \to v$ via the Hadamard product: $S_{IG}(u \to v) = \Delta \mathbf{a}_u \odot \overline{\nabla \mathbf{z}_v}$.
4. **Pruning**: Rank all edges by $|S_{IG}|$ and apply a threshold to select the top edges, forming the final circuit.

Although EAP-IG requires n backward passes (making it slower than Linear EAP), it remains significantly faster than brute-force patching while providing much more faithful attributions in non-linear regimes.

F Summary of Surveyed Papers

Table 3: Summary of Surveyed Papers. We annotate each paper with tags for its Core Interpretable Objects (§2), Localizing Methods (§3), and Steering Methods (§4). For studies employing multiple objects or localizing/steering methods, we annotate the primary tag. The symbol "-" in the Steering Method column denotes works that apply localized mechanistic insights directly for analysis or monitoring, without employing active intervention techniques.

Paper	Object	Localizing Method	Steering Method	Venue	Year	Link
Safety and Reliability (Improve Alignment)						
Du et al.	Token Embedding	Gradient Detection	*Vector Arithmetic*	ArXiv	2025	Link
Jiang et al.	Attention	Causal Attribution	*Targeted Optimization*	ArXiv	2024	Link
Zhou et al.	Attention	Causal Attribution	*Amplitude Manipulation*	ICLR	2025	Link
Huang et al.	Attention	Circuit Discovery	*Targeted Optimization*	EMNLP	2025	Link
Chen et al.	Neuron	Causal Attribution	*Amplitude Manipulation*	ArXiv	2025	Link
Suau et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	ICML	2024	Link
Gao et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Zhao et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	ICLR	2025	Link
Li et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	ArXiv	2025	Link
Templeton et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	Blog	2024	Link
Goyal et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2025	Link
Yeo et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2025	Link
Weng et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Wu et al.	SAE Feature	Magnitude Analysis	*Vector Arithmetic*	ICML	2025	Link
Li et al.	SAE Feature	Magnitude Analysis	*Vector Arithmetic*	ArXiv	2025	Link
He et al.	SAE Feature	Magnitude Analysis	*Vector Arithmetic*	ArXiv	2025	Link
Li et al.	Residual Stream	Causal Attribution	*Targeted Optimization*	ICLR	2025	Link
Lee et al.	Residual Stream	Probing	*Targeted Optimization*	ICML	2024	Link
Arditi et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	NeurIPS	2024	Link
Zhao et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	NeurIPS	2025	Link
Yin et al.	Residual Stream	Probing	*Vector Arithmetic*	ArXiv	2025	Link
Ball et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ArXiv	2024	Link
Wang et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICLR	2025	Link
Wang et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	NeurIPS	2025	Link
Ferreira et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICML	2025	Link
Huang et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICML	2025	Link
Pan et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICML	2025	Link
Chuang et al.	Residual Stream	Vocab Projection	*Vector Arithmetic*	ICLR	2024	Link
Chen et al.	Residual Stream	Vocab Projection	*Vector Arithmetic*	ICML	2024	Link
Zhang et al.	Residual Stream	Probing	*Vector Arithmetic*	ACL	2024	Link
Orgad et al.	Residual Stream	Probing	*Vector Arithmetic*	ICLR	2025	Link
Stolfo et al.	Residual Stream	Gradient Detection	*Vector Arithmetic*	ICLR	2025	Link
Fairness and Bias (Improve Alignment)						
Cai et al.	FFN	Causal Attribution	*Targeted Optimization*	ICIC	2024	Link
Ahsan et al.	FFN	Causal Attribution	*Amplitude Manipulation*	EMNLP	2025	Link
Li and Gao	FFN	Vocab Projection	*Targeted Optimization*	ACL	2025	Link
Vig et al.	Attention	Causal Attribution	*Amplitude Manipulation*	NeurIPS	2020	Link
Chintam et al.	Attention	Causal Attribution	*Targeted Optimization*	ACLWS	2023	Link
Wang et al.	Attention	Magnitude Analysis	*Amplitude Manipulation*	ICLR	2025	Link
Chandna et al.	Attention	Magnitude Analysis	*Amplitude Manipulation*	TMLR	2025	Link
Dimino et al.	Attention	Magnitude Analysis	-	ICAIF	2025	Link
Kim et al.	Attention	Probing	*Vector Arithmetic*	ICLR	2025	Link
Liu et al.	Neuron	Gradient Detection	*Amplitude Manipulation*	ICLR	2024	Link
Yu and Ananiadou	Neuron	Circuit Discovery	*Targeted Optimization*	ArXiv	2025	Link
Yu et al.	Residual Stream	Causal Attribution	-	ArXiv	2025	Link
Guan et al.	Residual Stream	-	*Amplitude Manipulation*	ICML	2025	Link
Yu et al.	Residual Stream	Magnitude Analysis	*Amplitude Manipulation*	ACL	2025	Link
Raimondi et al.	Residual Stream	Causal Attribution	*Amplitude Manipulation*	ArXiv	2025	Link
Persona and Role (Improve Alignment)						
Su et al.	Neuron	Causal Attribution	*Amplitude Manipulation*	EMNLP	2025	Link
Deng et al.	Neuron	Causal Attribution	*Amplitude Manipulation*	ICLR	2025	Link
Chen et al.	Neuron	Causal Attribution	*Targeted Optimization*	ICML	2024	Link
Rimsky et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ACL	2024	Link
Potertì et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	EMNLP	2025	Link
Chen et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ArXiv	2025	Link
Handa et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	NeurIPS	2025	Link
Tak et al.	Residual Stream	Probing	*Vector Arithmetic*	ACL	2025	Link
Yuan et al.	Residual Stream	Probing	-	ArXiv	2025	Link

Paper	Object	Localizing Method	Steering Method	Venue	Year	Link
Ju et al.	Residual Stream	Probing	*Targeted Optimization*	COLM	2025	Link
Karny et al.	Residual Stream	Causal Attribution	-	ArXiv	2025	Link
Banayeeanzade et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ArXiv	2025	Link
Bas and Novak	Residual Stream	Causal Attribution	*Vector Arithmetic*	ArXiv	2025	Link
Sun et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	EMNLP	2025	Link
Pai et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ArXiv	2025	Link
Joshi et al.	Residual Stream	Probing	-	EMNLP	2024	Link
Ghandeharioun et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	NeurIPS	2024	Link
Logic and Reasoning (Improve Capability)						
Wu et al.	Token Embedding	Gradient Detection	-	ICML	2023	Link
You et al.	Token Embedding	Magnitude Analysis	-	EMNLP	2025	Link
Cywiński et al.	Token Embedding	Causal Attribution	*Amplitude Manipulation*	Blog	2025	Link
Wang et al.	FFN	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Yu and Ananiadou	Attention	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2024	Link
Zhang et al.	Attention	Causal Attribution	*Targeted Optimization*	ICML	2024	Link
Yu and Ananiadou	Attention	Causal Attribution	*Amplitude Manipulation*	EMNLP	2024	Link
Yu et al.	Attention	Causal Attribution	-	EMNLP	2025	Link
Stolfo et al.	FFN& Attention	Causal Attribution	-	EMNLP	2023	Link
Akter et al.	FFN& Attention	Causal Attribution	-	COMPSAC	2024	Link
Yang et al.	FFN& Attention	Magnitude Analysis	-	ArXiv	2024	Link
Quirke and Barez	FFN& Attention	Causal Attribution	*Amplitude Manipulation*	ICLR	2024	Link
Chen et al.	FFN& Attention	Gradient Detection	*Targeted Optimization*	ACL	2025	Link
Hanna et al.	FFN& Attention	Circuit Discovery	-	NeurIPS	2023	Link
Nikankin et al.	FFN& Attention	Circuit Discovery	-	ICLR	2025	Link
Galichin et al.	SAE Feature	Magnitude Analysis	*Vector Arithmetic*	ArXiv	2025	Link
Pach et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Troitskii et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2025	Link
Venhoff et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICLR	2025	Link
Højer et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICLR	2025	Link
Tang et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ACL	2025	Link
Hong et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ACL	2025	Link
Zhang and Viteri	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICLR	2025	Link
Liu et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ArXiv	2025	Link
Sinii et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	EMNLP	2025	Link
Li et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	EMNLP	2025	Link
Ward et al.	Residual Stream	Causal Attribution	*Vector Arithmetic*	ICML	2025	Link
Biran et al.	Residual Stream	Probing	-	EMNLP	2024	Link
Ye et al.	Residual Stream	Probing	-	ICLR	2025	Link
Sun et al.	Residual Stream	Probing	-	EMNLP	2025	Link
Wang et al.	Residual Stream	Probing	*Vector Arithmetic*	AAAI	2026	Link
Tan et al.	Residual Stream	Vocab Projection	*Targeted Optimization*	ArXiv	2025	Link
Multilingualism (Improve Capability)						
Xie et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	ACL	2021	Link
Kojima et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	NAACL	2024	Link
Tang et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	ACL	2024	Link
Zhao et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	NeurIPS	2024	Link
Gurgurov et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Liu et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2025	Link
Jing et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2025	Link
Andrylie et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Brinkmann et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	NAACL	2025	Link
Libovický et al.	Residual Stream	Probing	-	EMNLP	2020	Link
Chi et al.	Residual Stream	-	*Vector Arithmetic*	ACL	2023	Link
Philippy et al.	Residual Stream	Magnitude Analysis	*Vector Arithmetic*	ACL	2023	Link
Wendler et al.	Residual Stream	Vocab Projection	*Vector Arithmetic*	ACL	2024	Link
Mousi et al.	Residual Stream	Magnitude Analysis	*Vector Arithmetic*	ACL	2024	Link
Hinck et al.	Residual Stream	Probing	*Vector Arithmetic*	EMNLP	2024	Link
Zhang et al.	Residual Stream	Magnitude Analysis	*Vector Arithmetic*	ACL	2025	Link
Wang et al.	Residual Stream	Vocab Projection	*Vector Arithmetic*	ACL	2025	Link
Wu et al.	Residual Stream	Vocab Projection	-	ICLR	2025	Link
Wang et al.	Residual Stream	Vocab Projection	*Vector Arithmetic*	EMNLP	2025	Link
Nie et al.	Residual Stream	Vocab Projection	*Vector Arithmetic*	EMNLP	2025	Link
Liu et al.	Residual Stream	Vocab Projection	*Vector Arithmetic*	EMNLP	2025	Link
Knowledge Management (Improve Capability)						
Meng et al.	FFN	Causal Attribution	*Targeted Optimization*	NeurIPS	2022	Link
Meng et al.	FFN	Causal Attribution	*Targeted Optimization*	ICLR	2023	Link

Paper	Object	Localizing Method	Steering Method	Venue	Year	Link
Lai et al.	Attention	Magnitude Analysis	*Targeted Optimization*	ICML	2025	Link
Li et al.	Attention	Magnitude Analysis	*Amplitude Manipulation*	ICML	2025	Link
Jin et al.	Attention	Magnitude Analysis	*Amplitude Manipulation*	ICML	2025	Link
Jin et al.	Attention	Circuit Discovery	*Amplitude Manipulation*	ACL	2024	Link
Lv et al.	Attention	Causal Attribution	*Amplitude Manipulation*	ArXiv	2024	Link
Niu et al.	Attention	Causal Attribution	*Amplitude Manipulation*	ACL	2025	Link
Zhao et al.	Attention	Probing	*Targeted Optimization*	EMNLP	2025	Link
Yadav et al.	FFN& Attention	Magnitude Analysis	*Vector Arithmetic*	NeurIPS	2023	Link
Yu and Ananiadou	FFN& Attention	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2024	Link
Zhang et al.	FFN& Attention	Magnitude Analysis	*Targeted Optimization*	ACL	2024	Link
Chen et al.	FFN& Attention	Magnitude Analysis	*Amplitude Manipulation*	ICLR	2025	Link
Li et al.	FFN& Attention	Magnitude Analysis	*Targeted Optimization*	AAAI	2025	Link
Muhamed and Smith	FFN& Attention	Magnitude Analysis	-	ICML	2025	Link
Yao et al.	FFN& Attention	Circuit Discovery	*Amplitude Manipulation*	NeurIPS	2024	Link
Du et al.	FFN& Attention	Probing	*Targeted Optimization*	ArXiv	2024	Link
Zhang et al.	FFN& Attention	Gradient Detection	*Targeted Optimization*	ACL	2024	Link
Liu et al.	FFN& Attention	Gradient Detection	*Vector Arithmetic*	ACL	2025	Link
Geva et al.	FFN& Attention	Causal Attribution	-	EMNLP	2023	Link
Zhang et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	COLING	2025	Link
Chen et al.	Neuron	Gradient Detection	*Amplitude Manipulation*	AAAI	2024	Link
Shi et al.	Neuron	Gradient Detection	*Amplitude Manipulation*	NeurIPS	2024	Link
Chen et al.	Neuron	Gradient Detection	*Amplitude Manipulation*	AAAI	2025	Link
Kassem et al.	Neuron	-	*Amplitude Manipulation*	EMNLP	2025	Link
Muhamed et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	ICML	2025	Link
Goyal et al.	SAE Feature	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2025	Link
Marks et al.	SAE Feature	Circuit Discovery	*Amplitude Manipulation*	ICLR	2025	Link
Kang and Choi	Residual Stream	Probing	-	EMNLP	2023	Link
Katz et al.	Residual Stream	Vocab Projection	*Targeted Optimization*	EMNLP	2024	Link
Wu et al.	Residual Stream	Causal Attribution	*Targeted Optimization*	NeurIPS	2024	Link
Zhao et al.	Residual Stream	Probing	-	ArXiv	2024	Link
Ju et al.	Residual Stream	Probing	-	COLING	2024	Link
Jin et al.	Residual Stream	Probing	-	COLING	2025	Link
Chen et al.	Residual Stream	Probing	*Vector Arithmetic*	NeurIPS	2025	Link
Efficient Training (Improve Efficiency)						
Sergeev and Kotelnikov	Attention	Magnitude Analysis	*Targeted Optimization*	ICAI	2025	Link
Olsson et al.	Attention	Magnitude Analysis	-	ArXiv	2022	Link
Wang et al.	Attention	Magnitude Analysis	-	ArXiv	2024	Link
Singh et al.	Attention	Magnitude Analysis	-	ICML	2024	Link
Hoogland et al.	Attention	Magnitude Analysis	-	TLMR	2025	Link
Minegishi et al.	Attention	Magnitude Analysis	-	ICLR	2025	Link
Lai et al.	Attention	Magnitude Analysis	*Vector Arithmetic*	ICML	2025	Link
Thilak et al.	Attention& FFN	Magnitude Analysis	-	NeurIPS	2022	Link
Varma et al.	Attention& FFN	Magnitude Analysis	-	ArXiv	2023	Link
Furuta et al.	Attention& FFN	Magnitude Analysis	-	TMLR	2024	Link
Nanda et al.	Attention& FFN	Magnitude Analysis	-	ICLR	2023	Link
Notsawo Jr et al.	Attention& FFN	Magnitude Analysis	-	ArXiv	2023	Link
Qiye et al.	Attention& FFN	Magnitude Analysis	-	ArXiv	2024	Link
Liu et al.	Attention& FFN	Magnitude Analysis	-	ICLR	2023	Link
Wang et al.	Attention& FFN	Magnitude Analysis	-	NeurIPS	2024	Link
Huang et al.	Attention& FFN	Magnitude Analysis	-	COLM	2024	Link
Li et al.	Attention& FFN	Circuit Discovery	*Targeted Optimization*	ArXiv	2025	Link
Panigrahi et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	ICML	2023	Link
Zhu et al.	Neuron	Gradient Detection	*Targeted Optimization*	ACL	2024	Link
Song et al.	Neuron	Gradient Detection	*Targeted Optimization*	ICML	2024	Link
Zhang et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	ACL	2023	Link
Xu et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	COLING	2025	Link
Mondal et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	ACL	2025	Link
Gurgurov et al.	Neuron	Magnitude Analysis	*Targeted Optimization*	AACL	2025	Link
Zhao et al.	Neuron	Causal Attribution	*Targeted Optimization*	NeurIPS	2024	Link
Li et al.	Neuron	Magnitude Analysis	-	ArXiv	2025	Link
Efficient Inference (Improve Efficiency)						
Xia et al.	Token Embedding	Gradient Detection	*Amplitude Manipulation*	EMNLP	2025	Link
Lei et al.	Token Embedding	Gradient Detection	*Amplitude Manipulation*	ArXiv	2025	Link
Guo et al.	Token Embedding	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2024	Link
Ye et al.	Token Embedding	Magnitude Analysis	*Amplitude Manipulation*	AAAI	2025	Link
He et al.	Token Embedding	Magnitude Analysis	*Amplitude Manipulation*	NeurIPS	2024	Link
Cai et al.	Token Embedding	Magnitude Analysis	*Amplitude Manipulation*	COLM	2025	Link

Paper	Object	Localizing Method	Steering Method	Venue	Year	Link
Lu et al.	FFN	Magnitude Analysis	*Amplitude Manipulation*	ACL	2024	Link
Yu et al.	FFN	Magnitude Analysis	*Amplitude Manipulation*	Arxiv	2024	Link
Su et al.	FFN	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Xiao et al.	Attention	Magnitude Analysis	*Amplitude Manipulation*	ICLR	2024	Link
Su et al.	Attention	Magnitude Analysis	*Amplitude Manipulation*	IJCAI	2025	Link
Bi et al.	Attention	Magnitude Analysis	-	CVPR	2025	Link
Tang et al.	Attention	Circuit Discovery	*Amplitude Manipulation*	ICLR	2025	Link
Xiao et al.	Attention	Circuit Discovery	*Amplitude Manipulation*	ICLR	2025	Link
Xiao et al.	FFN& Attention	Magnitude Analysis	*Amplitude Manipulation*	NeurIPS	2022	Link
Sun et al.	FFN& Attention	Magnitude Analysis	-	NeurIPS	2024	Link
Lin et al.	FFN& Attention	Magnitude Analysis	*Amplitude Manipulation*	MLSyS	2024	Link
Ashkboos et al.	FFN& Attention	Magnitude Analysis	*Amplitude Manipulation*	NeurIPS	2025	Link
Su and Yuan	FFN& Attention	Circuit Discovery	-	COLM	2025	Link
An et al.	FFN& Attention	Circuit Discovery	-	ICLR	2025	Link
Bondarenko et al.	FFN& Attention	Circuit Discovery	-	NeurIPS	2023	Link
Liu et al.	Neuron	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2024	Link
Tan et al.	Neuron	Magnitude Analysis	-	EMNLP	2024	Link
Laitenberger et al.	Residual Stream	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Wang et al.	Residual Stream	Magnitude Analysis	*Amplitude Manipulation*	EMNLP	2023	Link
Shelke et al.	Residual Stream	Magnitude Analysis	*Amplitude Manipulation*	ACL	2024	Link
Lawson and Aitchison	Residual Stream	Magnitude Analysis	*Amplitude Manipulation*	ArXiv	2025	Link
Men et al.	Residual Stream	Magnitude Analysis	*Amplitude Manipulation*	ACL	2025	Link
Dumitru et al.	Residual Stream	Magnitude Analysis	-	ArXiv	2024	Link
Zhang et al.	Residual Stream	Magnitude Analysis	-	ArXiv	2025	Link
Xiao et al.	Residual Stream	Magnitude Analysis	-	ArXiv	2025	Link
Valade	Residual Stream	Probing	*Amplitude Manipulation*	ArXiv	2024	Link
Elhoushi et al.	Residual Stream	Probing	*Amplitude Manipulation*	ACL	2024	Link
Ranjan and Savakis	Residual Stream	Gradient Detection	-	ArXiv	2025	Link
Zeng et al.	Residual Stream	Vocab Projection	-	ArXiv	2024	Link