

Article

Not peer-reviewed version

Mechanistic Interpretability of LLM Behavior: A Survey

[Priya Pitre](#)*, Jun Hou, Ziyu Yao, Naren Ramakrishnan, Xuan Wang

Posted Date: 26 October 2026

Keywords: mechanistic interpretability; LLM behaviors; cross-paper findings; survey



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Mechanistic Interpretability of LLM Behavior: A Survey

Priya Pitre^{1,*}, Jun Hou¹, Ziyu Yao², Naren Ramakrishnan¹ and Xuan Wang¹

¹ Virginia Tech, Blacksburg, VA, USA
² George Mason University, Fairfax, VA, USA
* Correspondence: priyapitre@vt.edu

Abstract

Mechanistic interpretability seeks to explain how internal computations produce large language model behaviors, yet findings remain fragmented across methods and behavioral domains. We survey 229 papers across five behavior families: epistemic, linguistic, reasoning, safety, and social behaviors. We introduce a framework distinguishing five levels of mechanistic evidence, from decodability to generalization. Our cross-behavior synthesis reveals recurring mechanisms, distinctions between representation and causal control, and limitations of existing explanations. We conclude by identifying gaps in understanding computational roles, behavioral interactions, and generalization to interactive and agentic settings.

Keywords: mechanistic interpretability; LLM behaviors; cross-paper findings; survey

1. Introduction

Large language models (LLMs) exhibit behaviors that are increasingly difficult to understand from their outputs alone. A model may know that a statement is false but still produce it, recognize that a request is harmful but fail to refuse it, or shift its answer toward a user’s stated belief despite internally representing contradictory information. Behavioral evaluation can reveal when these failures occur, but not why they occur inside the model.

Mechanistic interpretability (MI) aims to explain model behavior by studying its internal representations and computations. Recent work has identified internal signals associated with factual knowledge, truthfulness, reasoning, and refusal, and shown that intervening on these signals can change outputs [1–4]. These results suggest that high-level behaviors have identifiable internal structure.

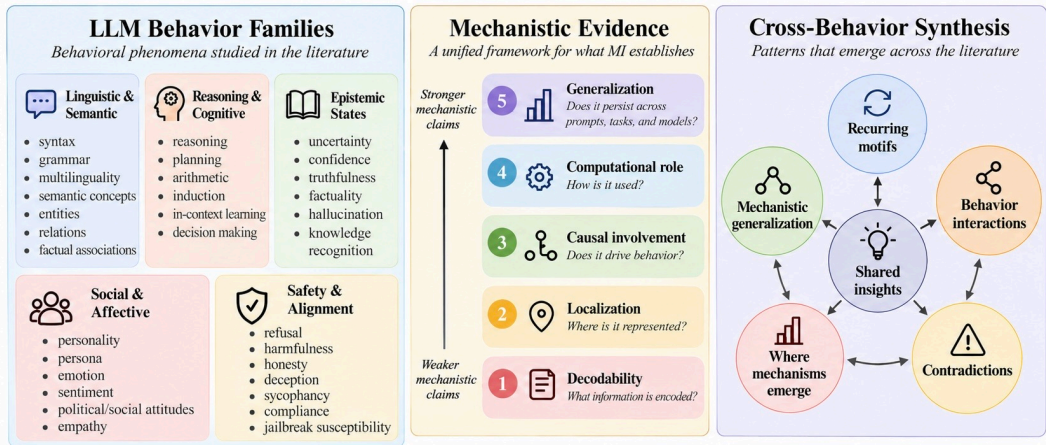


Figure 1. Our framework. LLM behavior families (left) are assessed by mechanistic evidence strength (center) and synthesized into cross-behavior patterns (right).

However, as the literature has grown, it has become less clear what counts as a mechanistic explanation. A linear probe may show that information is available, but not that the model uses it. Neurons or attention heads associated with a behavior may not explain the computation producing it. Likewise, activation steering may change a behavior without establishing whether the direction reflects the mechanism normally used by the model. Findings described as “mechanistic” therefore span different levels of evidence [5,6].

This creates a broader problem for understanding LLM behavior. Results about the same behavior are often distributed across different interpretability methods, while closely related behaviors are studied in separate literatures. Refusal is studied separately from harmfulness, hallucination from uncertainty and knowledge, and sycophancy from truthfulness and agreement. As a result, it is difficult to determine whether these behaviors rely on distinct mechanisms, shared representations, or interactions between internal signals. A behavior-centric view helps bring these findings together by organizing the literature around the behavior being explained rather than the method used.

We provide a behavior-centric survey of mechanistic interpretability in LLMs, reviewing 229 papers across five behavior families: linguistic and semantic, reasoning and cognitive, epistemic, social and affective, and safety and alignment behaviors. Literature selection methodology is described in Appendix A. We additionally discuss emerging work on multi-agent and interactive settings. We compare findings across methods, examine mechanistic evidence for each behavior, and identify connections between related behaviors.

Related Surveys.

Existing surveys organize mechanistic interpretability around methods, research tasks, and causal analysis. Some provide practical overviews of localization, steering, and actionable interventions [6–9], while others focus on sparse autoencoders, component descriptions, or specific capabilities [10–14]. Safety- and alignment-focused surveys examine mechanistic explanations and interventions for model alignment [15,16]. Related methodological work examines causal validity and activation-patching choices [17,18]. In contrast, we organize studies by the behaviors being explained and compare the strength of their mechanistic evidence across methods, allowing us to identify disagreements, shared computations, and gaps across behaviors.

2. From Behavior to Mechanistic Evidence

Mechanistic interpretability connects behavior to internal components, including neurons, sparse features, attention heads, MLPs, layers, activation directions, and circuits [19–21]. We organize evidence into five dimensions: *decodability*, *localization*, *causal involvement*, *computational role*, and *generalization*. Building on distinctions between representational and causal evidence [6,17,22], this framework distinguishes what studies establish, from available information to causal computations and their generalization.

Decodability.

Decodability asks whether behavior-related information can be recovered from activations using probes, classifiers, Patchscopes [23], LatentQA [24], or activation oracles [25]. Studies have decoded truth-related information from hidden states [2,26]. Decoding establishes information availability, not necessarily its use in producing behavior.

Localization.

Localization identifies where relevant information or computation resides, such as neurons, features, heads, or layers. Studies have localized factual knowledge to neurons [19], identified sparse features [20], examined attention heads involved in computations [21], and studied concept geometry [27]. Although localization can involve probing, it establishes where information is found rather than only whether it is recoverable.

Causal Involvement.

Causal involvement tests whether interventions on internal components change behavior, using methods such as causal mediation analysis [28] and interchange interventions [29]. Encoded information need not causally drive behavior [30], and successful intervention does not necessarily explain the underlying computation.

Computational Role.

Computational role explains how components interact to produce behavior. Circuit analyses, for example, have identified attention heads implementing indirect object identification in GPT-2 [21]. This goes beyond locating or manipulating components to explaining their computation.

Generalization.

Generalization tests whether explanations hold across prompts, tasks, datasets, or models, since interpretations may fail out of distribution [31]. Unlike the other dimensions, it measures the scope of an explanation rather than a separate internal mechanism. An additional question is whether mechanisms identified in language models transfer to multimodal models, where cross-modal representations and interactions may introduce different computations [32].

3. Findings Across LLM Behaviors

We review mechanistic findings across five behavior families, comparing their evidence, recurring results, and unresolved questions.

3.1. Epistemic Behaviors

Epistemic behaviors concern how models represent and use truth, correctness, and their own reliability. **Decodability is better established than full mechanistic explanation:** internal states reveal correctness, but how these signals produce behavior remains less clear.

For **truthfulness and factuality**, correctness is recoverable from hidden states, heads, FFN activations, and sparse features, and steering can change factual outputs [3,26,33–35]. However, signals can reflect polarity and transfer weakens across prompts or tasks requiring additional computation [36–38]. **This supports causal truth-related signals, but not a universal truth representation.**

For **uncertainty and confidence**, internal states predict correctness before generation and partly overlap with answer-producing computations [39–41]. Yet confidence can be steered to increase decisiveness without improving accuracy [42]. **Estimating correctness is therefore distinct from willingness to answer.**

For **hallucination**, internal states distinguish factual from hallucinated outputs [35,43,44]. Errors can arise from weak retrieval, failed selection, or competing contextual and parametric signals [45–47]; incorrect associations can even resemble factual recall [48]. **Hallucination therefore reflects different internal failures, not one mechanism.**

For **knowledge recognition**, models can predict whether they know an answer [49], yet correct information may be internally available despite incorrect outputs [50,51]. **Stored, retrieved, and expressed knowledge are distinct.**

Together, these findings separate representation, retrieval, evaluation, and expression, explaining why decodability does not imply a complete computational account.

3.2. Linguistic and Semantic Behaviors

Linguistic behaviors provide strong **causal and partial computational explanations**, although generalization remains limited.

For **syntax and grammatical structure**, information is distributed across layers but mediated by neurons, heads, and sparse features [21,52–54]. Circuits detect repeated tokens, transmit grammatical information, inhibit alternatives, and select outputs [21,55]. **The computation appears more stable than its components**, which differ across models and languages [41,56,57].

For **semantic and factual information**, MLPs and FFNs enrich subject representations, while attention retrieves and routes information to outputs [1,19,58,59]. **Storage, retrieval, and expression involve different components**, explaining differences in localization, although these components may not transfer across tasks [51,60].

For **entity and relational binding**, retrieval uses position, entity identity, and target pointers. Circuits separate position detection, transmission, retrieval, and suppression [61–64]. Some components recur across tasks [55], but **the mechanism varies with context**.

For **multilingual behavior**, shared conceptual representations coexist with language-specific features that control generation language [65–68]. Studies suggest language-specific inputs, shared intermediate states, and language-specific outputs [69,70]. Yet localization differs, and language control does not guarantee cross-lingual transfer [68,71,72]. **Controlling expression differs from representing shared content**.

3.3. Reasoning and Cognitive Behaviors

Mechanistic studies identify intermediate representations and causal information flow, but **reasoning does not reduce to one general circuit**.

For **in-context learning**, induction heads match patterns and copy continuations [73,74], while function-vector heads, label states, and sparse features represent tasks and causally affect performance [75–77]. The findings differ because induction heads mainly affect token prediction, while function-vector heads more strongly affect few-shot accuracy in larger tested models [78]. **ICL combines matching and copying with task recognition and execution** [79,80].

For **arithmetic and numerical reasoning**, attention routes operands, while MLPs and sparse neurons contribute answer information [81–84]. Arithmetic can rely on overlapping heuristics rather than a single algorithm [82,85]; counting tasks also reveal causally manipulable numerical states [86–88]. **These suggest specialized numerical mechanisms**.

For **multi-hop and chain-of-thought reasoning**, two-hop circuits resolve bridge entities, propagate information, and retrieve targets; delayed or persistent intermediate states can cause failures [89–91]. CoT activations causally affect answers [92,93], while latent reasoning states can be decoded and edited [94]. **This does not establish that written CoT faithfully reflects internal computation**.

For **planning and supporting cognitive processes**, future actions are decodable, but causal evidence for complete internal plans is limited [95,96]. Retrieval heads have stronger causal support, moving context into outputs when repeated access is needed [97]. The main gap is generalization beyond controlled tasks.

3.4. Safety and Alignment Behaviors

Safety studies show strong causal control, but **steering does not imply complete mechanistic explanation**.

For **refusal**, one activation direction can causally control refusal [98,99], while other work finds multidimensional representations and multiple effective directions [100,101]. **One strong control point does not imply a one-dimensional representation**. Over-refusal also uses partly distinct, task-dependent features [102,103].

Harm recognition and refusal policy are partly separate: harmfulness can be represented before alignment, while later states influence refusal [104,105]. Distinct signals for harmfulness and willingness to answer [106,107] help explain why models sometimes recognize harm but still comply.

For **jailbreaks**, attacks often suppress harmfulness or refusal signals, and steering changes attack success [108–111]. Yet some attacks work through alternative or distributed pathways without such suppression [112,113], while framing can move queries into ambiguous internal regions [114]. **No single mechanism explains all jailbreaks**.

For **truthfulness, deception, and sycophancy**, truth-related features are causally steerable [3,34,115], whereas deception is mainly decodable, with domain-sensitive signals that transfer poorly [116–118]. Sycophancy

phancy has stronger causal evidence, but agreement, belief-conditioned agreement, and praise can involve separate representations [119–121]. **Neither deception nor sycophancy has one established mechanism.**

For **instruction following**, directions causally control output constraints [122], while multi-layer interventions reveal interactions with later task processing [123]. Thus, simple control directions can participate in complex computation.

3.5. Social and Affective Behaviors

Social and affective behaviors are often decodable and steerable, but their representations are **distributed and interconnected**.

For **personality and persona**, directions predict and change traits [124–126]. Yet traits overlap, steering one affects others, and relevant neurons vary with context [127,128]. Opposing traits may reuse components but change their information flow [129]. **Simple directions therefore control distributed computations.**

For **emotion**, early work found a causal sentiment unit [130], while later studies find distributed pathways and representations organized by valence and arousal [131–133]. Processing appears staged, from detecting emotional salience to identifying specific emotions [134]. Attention and FFNs contribute causally [135], but components vary across models and datasets [136], explaining localization differences.

For **social cognition and attitudes**, beliefs, desires, and intentions are decodable and affected by intervention [137], while empathy directions transfer poorly across models [138]. Political features causally change ideological framing, with general and topic-specific components [139,140], but **changing framing need not change voting choices** [141].

For **values, culture, and bias**, features can shift moral values, cultural judgments, and stereotypes [142–145]. These representations do not always match human categories, and values remain coupled [146,147]. Culture neurons are localized differently across methods [148,149], making localization task- and method-dependent.

For **demographic bias**, gender information can be more localized than race information [150], while steerable race features need not explain bias in realistic decisions [151]. Gender features can support both occupational bias and pronoun prediction [152]. **Removing bias-related features can therefore disrupt useful behavior.**

Overall, these findings reach causal involvement more often than full computational explanation, especially across contexts and models.

3.6. Multi-Agent and Interactive Behaviors

Mechanistic work on multi-agent systems is still emerging. For **peer pressure and sycophancy**, activation patching localizes answer corruption to mid-layer components, suggesting that peer pressure suppresses reasoning features rather than activating a separate sycophancy circuit [153]. Other studies identify agent-specific activation subspaces that causally control internalized debate behavior [154], while learned representations can steer cooperative and competitive behavior [155]. Related work on **latent communication** shows that agents can exchange hidden representations directly, with studies examining hidden-state alignment and information transfer across communication channels [156–158]. Together, these findings provide early evidence for internal representations involved in multi-agent interactions [159].

4. Cross-Behavior Mechanistic Synthesis

Across behavior families, mechanistic studies reveal differences between what models represent and how they use it. Simple interventions can affect distributed computations, similar behaviors can rely on different mechanisms, and human-defined categories may not reflect internal organization [4,78,120,142]. These patterns help explain behavioral failures and differences across studies.

4.1. From Internal Information to Behavior

Available information does not always become correct behavior.

Across epistemic, reasoning, and safety behaviors, models can represent useful information without using it correctly. Hallucination can arise from weak retrieval, competition between contextual and parametric information, or incorrect selection [45,46,48,160]. Models may encode correctness signals before answering but fail to use them for abstention [39,42,161]. In multi-hop reasoning, intermediate results may form too late or be routed incorrectly [89,90,162]. Failures can therefore reflect how information is retrieved, selected, or used rather than its absence.

Models often detect the character of their output before producing it.

Behavior-relevant signals can emerge before the behavior itself. Hidden states can predict correctness or distinguish hallucinated outputs before generation completes; harmfulness can be separable before refusal and alignment; and future actions can sometimes be decoded before generation [39,43,95,104]. This reveals a gap between detecting information and acting on it.

 **Knowing is not the same as using.** Improving detection alone may be insufficient. Interventions may need to target how internal signals influence decisions, not only whether those signals are present.

Recognition and behavioral decisions can be separate.

Models can recognize harmfulness before refusal-related states emerge, while alignment changes how this information affects refusal or compliance [104,163]. Similarly, user beliefs can influence answers under sycophancy even when contradictory information remains available [119,164]. What controls a response can therefore differ from what the model represents.

Representing, executing, and expressing a behavior can involve different operations.

In-context learning separates induction-head pattern matching and copying from function-vector task recognition and execution [73,75]. These mechanisms affect next-token prediction and few-shot accuracy differently, explaining why studies may identify different components [78]. Similarly, multilingual models can share semantic representations while controlling output language separately, so language steering need not improve cross-lingual transfer [66,68,71]. Changing one stage need not affect the others.

4.2. How Behavioral Mechanisms Are Organized

Compact control can act on distributed mechanisms.

Low-dimensional directions can steer refusal, personality, truthfulness, and language choice despite their distributed or multidimensional representations [3,4,124,165]. A single direction can suppress refusal even when its representation spans multiple dimensions and varies across models [100,101]. Effective control therefore need not reveal the full underlying computation.

 **Steering success is not completeness.** A reliable intervention identifies a way to control behavior, but does not necessarily explain how the model normally produces it.

Different behaviors reuse operations, often in sequences.

Syntax, factual recall, binding, in-context learning, and reasoning reuse matching, copying, retrieval, routing, suppression, and selection [21,55,58]. These operations form sequences: factual recall enriches associations before answer routing, multi-hop reasoning passes intermediate results, and safety connects harm recognition to response policy [58,89]. Studies may identify different components because they examine different stages of the same behavior.

Desirable and undesirable behaviors can share components.

Gender features involved in occupational bias also support pronoun prediction [166]; safety and helpfulness neurons overlap [105]; and long-context retrieval heads also support chain-of-thought performance [97]. Removing a causal component may therefore affect capabilities beyond the targeted behavior.

 **Selective removal may involve a mechanistic tradeoff.** When desirable and undesirable behaviors share components, suppressing one can disrupt the other. Better localization alone may not resolve this overlap.

Computational roles can persist as components change.

Similar syntactic and circuit functions recur across languages, models, and training runs even when individual neurons or heads differ [141,167,168]. Generalization should therefore test causal roles, not only component identities. Representations and effects can also change across conversations and agent trajectories, limiting static localization [128,169].

Mechanistic depth depends on behavioral scope.

Detailed circuits exist for narrow tasks such as induction, indirect-object identification, and factual retrieval. Broader behaviors such as hallucination, personality, and safety involve more interacting processes and contexts, making complete explanations harder [45,112,170]. Mechanistic understanding therefore varies with behavioral complexity and scope.

4.3. *Why Mechanistic Findings Differ Across Studies*

Behavioral labels can group different mechanisms.

Hallucination can involve retrieval, grounding, or selection failures; sycophancy can involve agreement, belief-conditioned agreement, or praise; and in-context learning combines copying and task execution [75,120]. Appropriate refusal, over-refusal, and jailbreak susceptibility may also involve different safety processes [103,112]. A shared label does not guarantee a shared mechanism.

Representation geometry need not follow human-defined categories.

Representations of moral foundations, values, and personality do not necessarily match their human-defined taxonomies [127,142,146]. Emotions can organize by valence and arousal rather than discrete categories, while truth directions vary across tasks [36,38,132,133]. Failed generalization may therefore reflect the behavioral label rather than the representation.

 **Behavioral labels may be the wrong unit** Human-defined categories may separate shared mechanisms or combine distinct ones. Mechanistic analysis may need to derive behavioral units from internal computations rather than assume them from labels.

Different evidence can reveal different parts of a mechanism.

Probing identifies recoverable information, steering tests causal influence, and circuit analysis examines component interactions [5,6]. Studies measuring next-token prediction versus few-shot accuracy, or correctness prediction versus willingness to answer, may identify different components because they test different processes [42,78]. Their findings need not conflict.

Training can change how existing representations are used.

Reasoning signals can predate reasoning fine-tuning, and harmfulness information can predate safety alignment. Training may change how these representations influence behavior or strengthen existing mechanisms rather than replace them [63]. Conversely, suppressed information may remain

recoverable, and removed capabilities can sometimes be restored [171]. Behavioral emergence need not imply new representations, and suppression need not imply deletion.

💡 **Suppression is not deletion.** Recoverable information may allow suppressed behaviors to reappear. Behavioral success after alignment does not establish that a capability has been removed or that suppression will remain robust.

Generalization depends on what is being tested.

Truth directions can fail across reasoning tasks, language steering may not improve cross-lingual transfer, social representations can shift with context, and circuit components may differ across models [36,38,68,128]. Representations, causal effects, and computational roles may generalize differently and must be tested separately.

Overall, behavior depends not only on represented information but on how internal operations select, transform, and use it. Mechanistic explanations must connect these operations to behavioral outcomes while distinguishing general patterns from setting-specific findings.

5. Open Problems and Research Agenda

Behavior-centered mechanistic interpretability must move beyond identifying individual components toward explaining how mechanisms interact, evolve, and generalize. Several broad problems remain.

From behavioral labels to mechanistic constructs.

Human-defined behavioral categories can group distinct internal mechanisms. Hallucination may involve failures of retrieval, selection, or source competition; sycophancy includes different forms of agreement; and refusal and over-refusal may involve different processes. Future work should identify **mechanistically meaningful subtypes of behavior** rather than assuming that behavioral labels correspond to distinct mechanisms.

From individual components to meaningful units of explanation.

Studies analyze neurons, attention heads, sparse features, and activation directions, but the right units for explaining behavior remain unclear. Sparse features, for example, may not provide a unique or complete decomposition. Future work should determine when mechanisms are best explained through individual features, distributed representations, circuits, or their interactions.

From internal signals to complete behavioral mechanisms.

Representations and control points are increasingly identifiable, but the computations connecting them to behavior remain less understood. This gap appears between harmfulness and refusal, uncertainty and abstention, factual knowledge and expression, and reasoning states and final decisions. Future work should trace how information is represented, routed, transformed, and selected, while distinguishing **causal control from the computation normally used by the model**.

From isolated behaviors to interacting behavioral systems.

Behaviors are usually studied separately even though their mechanisms can interact. Truth-related information may compete with agreement in sycophancy, uncertainty affects hallucination and abstention, and persona can influence social and safety behavior. Future work should determine whether behaviors use independent mechanisms, shared representations, or common decision processes, which may also explain unintended effects of steering.

From static mechanisms to multi-turn and evolving behavior.

Most studies focus on single prompts or forward passes, although persona, beliefs, uncertainty, and reasoning can change across interactions. Models also become less reliable when information is

distributed across turns [172]. Future work should track how representations and causal relationships evolve across reasoning stages, conversation turns, long contexts, and memory states.

From language models to agents and multi-agent systems.

Agents involve repeated observation, reasoning, action, and feedback, so failures may originate several steps before becoming visible. Existing MI methods provide only partial explanations of these trajectories, motivating methods that trace plans and internal states across layers and timesteps [173]. Multi-agent systems add influence, belief propagation, communication, and consensus, raising the question of how one agent's internal states affect others and collective behavior [159].

From final checkpoints to the formation of mechanisms.

Most findings describe finished models, leaving how mechanisms form during training less understood. Training may create representations, strengthen existing ones, change their routing, or connect them to new behavioral policies. Model diffing and training-time analysis can help distinguish these changes [174–176] and explain how training changes internal computations.

From local discoveries to general and validated mechanisms.

Mechanisms found in controlled settings may not generalize across prompts, tasks, languages, models, or training regimes, and similar functions may be implemented by different neurons or attention heads. Future work should test generalization at the levels of representation, causal effect, and computational function. **Explanations should also be evaluated for faithfulness and completeness**, including whether they make predictions beyond the discovery setting [177,178]. Scaling to larger models and naturalistic tasks will also require automated methods with validated causal effects and manageable computational cost.

Together, these problems point toward explanations that connect internal representations to behavior, capture interactions and changes over time, and remain valid across models and realistic settings.

6. Conclusions

We surveyed mechanistic interpretability across LLM behaviors, identifying recurring gaps between internal representations, causal control, and computational explanations. Our synthesis shows that behavioral labels can hide distinct mechanisms, while different behaviors may share computations. Future work should explain how these mechanisms interact, evolve, and generalize across realistic settings.

7. Limitations

First, although we review 229 papers, the rapidly growing mechanistic interpretability literature means that some recent or emerging work may not be included. Second, our organization into five behavior families is not exhaustive, and some studies span multiple categories. These boundaries are intended to support comparison rather than imply that behaviors have distinct underlying mechanisms. Third, our assessment of mechanistic evidence is qualitative. The five evidence dimensions provide a common framework for comparing studies, but differences in models, tasks, evaluation settings, and intervention methods make direct comparisons difficult. Our review focuses on language-model behaviors and does not systematically cover multimodal foundation models. Consequently, the extent to which the identified representations, causal effects, and computational roles generalize across modalities remains outside the scope of this survey. Finally, the surveyed literature is largely concentrated on controlled tasks and individual models, limiting the conclusions that can be drawn about mechanisms in realistic, multi-turn, and agentic settings.

8. Ethical Considerations and Reproducibility

Mechanistic interpretability can support model auditing, debugging, and safer deployment, but the same techniques can also be used to manipulate model behavior. Methods that identify or steer

internal representations may enable the suppression of safety mechanisms, stronger jailbreaks, or targeted changes to social and behavioral traits. Mechanistic findings may also create a false sense of assurance if localized or steerable components are treated as complete explanations of model behavior. This is particularly important for behaviors such as bias, personality, values, and safety, where interventions may affect related capabilities or encode assumptions from human-defined behavioral categories. We therefore treat mechanistic explanations as evidence about internal computation rather than guarantees of model safety or reliability.

Reproducibility remains challenging because results often depend on the model, checkpoint, prompt distribution, dataset, intervention method, and evaluation metric. Many studies also use proprietary models or release only partial code or activation data, making exact replication difficult. To make our own synthesis reproducible, we describe our literature search, inclusion criteria, study coding procedure, and mechanistic evidence definitions in Appendix A. However, because the literature is rapidly evolving, the reviewed corpus should be viewed as a snapshot of work available through September 2026 rather than an exhaustive or permanently fixed collection.

Appendix A. Literature Selection and Study Coding Methodology

Literature Search.

We collected research papers on mechanistic interpretability of language models from computational linguistics and machine learning conferences, including ACL, EMNLP, NAACL, NeurIPS, ICLR, and ICML, as well as relevant workshops and preprints. The search covered work available through September 2026, with particular emphasis on studies published since 2023, reflecting the rapid growth of mechanistic interpretability research during this period. Earlier foundational studies were also included where relevant. We used combinations of general search terms, including “mechanistic interpretability,” “activation patching,” “causal tracing,” “circuit analysis,” “knowledge neurons,” “sparse autoencoders,” “representation engineering,” and “activation steering.” These were combined with behavior-specific keywords such as “truthfulness,” “hallucination,” “uncertainty,” “reasoning,” “refusal,” “jailbreak,” “sycophancy,” “personality,” and “multilinguality.” We also examined references from relevant papers and existing surveys to identify additional studies. Technical reports and research published outside traditional conference proceedings, including research blogs such as LessWrong and Anthropic’s interpretability publications, were also considered. This process produced a final corpus of 229 papers.

Inclusion and Exclusion Criteria.

We included studies that investigate internal representations, components, or computations of language models and connect them to identifiable behaviors or capabilities. Eligible studies used methods such as probing, activation analysis, causal interventions, feature analysis, or circuit tracing to examine how behavior-related information is represented or processed. We excluded studies focusing exclusively on behavioral evaluation, prompting, or output-level explanations without analyzing model internals. General interpretability methods were included when they provided evidence relevant to a behavior covered by the survey. We retained studies with partial mechanistic evidence, since distinguishing incomplete explanations from stronger causal or computational accounts is a central goal of our framework.

Paper Screening.

Papers were screened for relevance to mechanistic interpretability and the behavioral scope of this survey. Titles and abstracts were examined to identify potentially relevant studies, followed by inspection of their methods, experiments, and findings. The screening considered whether a paper investigated an internal model quantity, connected it to a specific behavior, and provided evidence relevant to at least one dimension of our framework. Studies spanning multiple behavioral categories

were retained rather than restricted to a single category. Duplicate records and studies outside the survey scope were excluded.

Study Coding.

Each included study was examined along four aspects: (1) the behavior being investigated, (2) the internal unit of analysis, (3) the interpretability method employed, and (4) the mechanistic evidence established. We recorded the primary behavioral phenomenon and assigned it to one or more of five families: epistemic, linguistic and semantic, reasoning and cognitive, safety and alignment, and social and affective behaviors. Studies of emerging multi-agent and interactive behaviors were additionally identified. Units of analysis included neurons, sparse features, attention heads, MLPs, residual-stream states, activation directions or subspaces, and circuits or computational pathways. The interpretability method was recorded separately from the conclusions supported by its experiments. When a paper investigated multiple behaviors or reported distinct experiments, evidence was assessed separately for the relevant behavior-specific findings rather than assigning a single mechanistic evidence level to the entire paper.

Mechanistic Evidence Assignment.

We assessed each study using the five evidence dimensions introduced in Section 2. Decodability was assigned when behavior-related information could be recovered from internal states. Localization required identifying internal components or regions associated with the information or computation. Causal involvement required experimental interventions showing that changes to internal components affect the target behavior. Computational role required evidence explaining how components interact, transform, route, or select information to produce behavior. Generalization was recorded when a representation, causal effect, or proposed computation was evaluated across prompts, tasks, datasets, languages, models, or other settings beyond its original discovery conditions. These assignments were based on experimental findings rather than the terminology used by the authors. Studies could support multiple evidence dimensions, and partial or setting-specific evidence was distinguished from broader mechanistic claims.

Coding Consistency.

To maintain consistency, evidence assignments followed the criteria above rather than the methods used or the authors' descriptions of their findings. For example, successful activation steering was treated as evidence of causal involvement, but not necessarily an explanation of the computation normally producing the behavior. Similarly, identifying a behavior-related neuron or feature did not establish its causal role without intervention. Studies could satisfy multiple evidence dimensions, and ambiguous findings were assessed conservatively based on the experiments reported. For behavior-level comparisons, we recorded the strongest evidence established in the literature while distinguishing partial explanations from broadly supported mechanisms. Generalization was considered separately because it measures the scope of an explanation rather than a distinct internal computation.

LLM-Assisted Verification.

Study coding and evidence assessment were conducted by the researcher through examination of the original papers. LLM-generated summaries of research questions, methods, interventions, and findings were used as supplementary checks to identify potentially overlooked evidence or inconsistencies in interpretation. The summaries were compared against the original studies, with final classifications and interpretations determined by the researcher. LLM outputs were used only to support the review process, not as independent evidence for mechanistic claims. This use of LLMs as supporting tools is consistent with prior work on LLM-assisted annotation [179,180].

Cross-Study Synthesis.

After coding, findings were organized within each behavior family and compared across studies addressing related phenomena. We examined where different methods supported similar conclusions, where findings diverged, and whether differences could be explained by the behavioral definition, model, experimental setting, or strength of evidence. We then compared findings across behavior families to identify recurring computations, shared representations, behavioral interactions, and limitations in mechanistic understanding. The summary tables report the strongest evidence established for each behavior in the surveyed literature, rather than implying that every study reaches that level. Generalization was considered separately when assessing how broadly an explanation was supported.

Behavior Family Organization.

We organize studies into five families based on the primary behavioral phenomenon being explained, rather than assuming that each family corresponds to a distinct internal mechanism. Epistemic behaviors concern factual correctness, knowledge, and uncertainty; linguistic and semantic behaviors concern language structure, meaning, and expression; reasoning and cognitive behaviors concern intermediate computation, inference, and planning; safety and alignment behaviors concern compliance, refusal, and adherence to behavioral constraints; and social and affective behaviors concern interpersonal responses, attitudes, and traits. These boundaries are not mutually exclusive. For example, hallucination can involve reasoning failures, but is grouped under epistemic behavior when the main question concerns factual correctness or knowledge use. Similarly, instruction following can involve linguistic processing, but is considered under safety and alignment when the study examines compliance with instructions or constraints. Studies spanning several behaviors were assigned to multiple relevant categories. This organization supports comparison across related behavioral questions without implying that the underlying mechanisms are independent.

Evaluating Mechanistic Explanations.

Our evidence framework distinguishes the type of mechanistic claim established by a study, but does not treat reaching a higher evidence dimension as sufficient for a complete explanation. We additionally consider the faithfulness, completeness, and specificity of reported findings. Faithfulness concerns whether the proposed mechanism reflects computations actually used by the model, rather than an intervention that merely changes its output. Completeness concerns whether the identified components and interactions explain the relevant behavior or only part of it. Specificity concerns whether interventions affect the targeted behavior without broadly disrupting unrelated computations. These properties depend on experimental design, including intervention controls, evaluation metrics, and tests beyond the discovery setting [6,17,18]. We therefore distinguish successful behavioral control from explanations of the model's normal computation and avoid treating localized or causally influential components as complete mechanisms without further evidence.

Quantitative Analysis of Mechanistic Evidence

We analyzed the distribution of mechanistic evidence across 229 annotated study records. Figure A1 summarizes the evidence established across the reviewed literature. Decodability was fully supported in 132 records, localization in 26, and causal involvement in 135. In contrast, computational role was classified as partial in 214 records, with no records coded as establishing a complete computational explanation. These results suggest that identifying and causally manipulating behavior-related representations is more common than fully explaining the computations that produce behavior.

The distribution also differs across behavioral families (Figure A2). Causal involvement was most frequently supported in social and affective (33 records), safety and alignment (32), linguistic (31), and reasoning (27) behaviors, compared with epistemic behaviors (9). However, these counts also reflect differences in the number of studies covering each family. The analysis therefore describes the coverage and strength of evidence in the surveyed literature rather than the prevalence of mechanisms in language models.

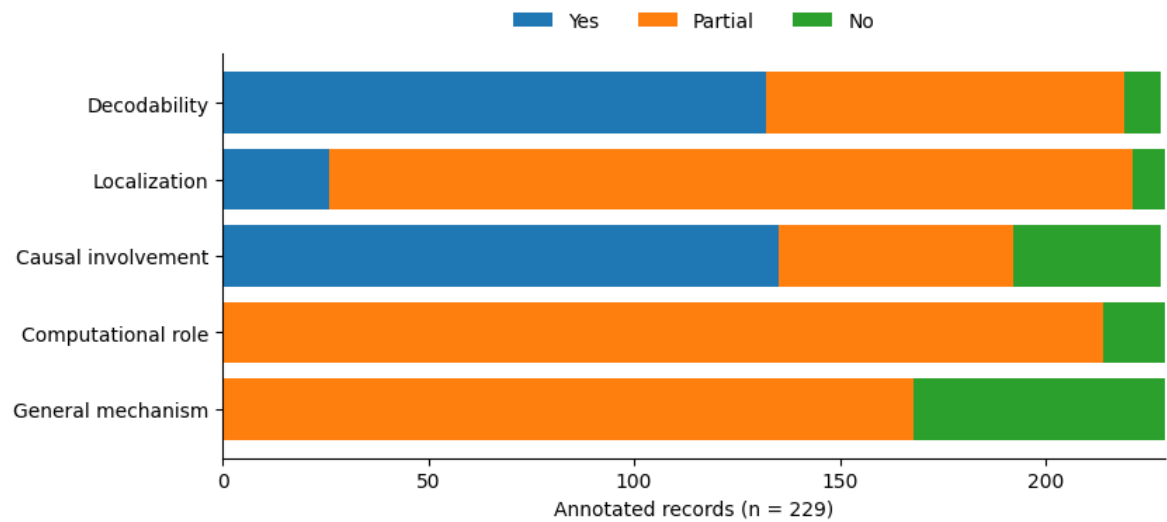


Figure A1. Distribution of mechanistic evidence across 229 annotated study records. Evidence is classified as fully supported (Yes), partially supported (Partial), or not established or unclear (No/unclear). The general mechanism category reflects whether a broader mechanism was established, rather than the framework’s generalization dimension.

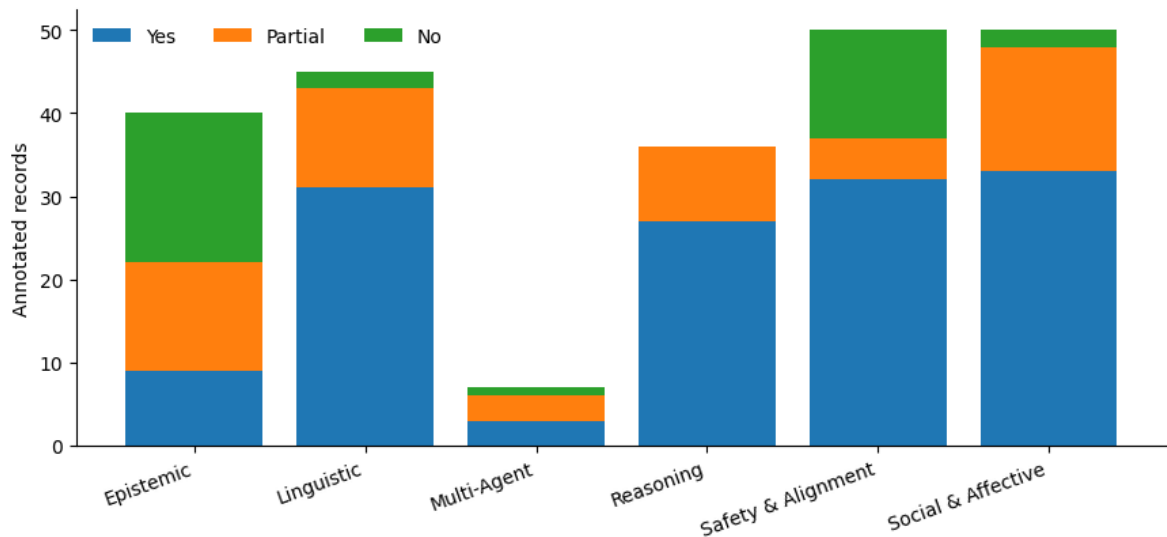


Figure A2. Causal involvement across behavior families. Bars show the number of annotated records with full, partial, or unsupported causal evidence. Differences in total records across families should be considered when comparing coverage.

Appendix B. Detailed Explanations for Behaviors and Findings

Appendix B.1. Epistemic Behaviors

For **truthfulness and factuality**, the literature suggests that models often internally distinguish correct from incorrect content, but not through a single universal representation of truth. Truth-related information is recoverable from hidden states, attention heads, FFN activations, activation geometry, and sparse features, often most clearly in intermediate layers [3,26,33–35,181]. These representations can also influence generation: intervening on truth-associated activations can increase or decrease factual outputs [3,34,35,182,183]. However, what is encoded appears to depend on the task. Directions that separate simple true and false statements do not always transfer when correctness requires arithmetic, counting, logical transformation, or different prompts, and some early “truth” signals instead track properties such as statement polarity [36–38]. Strong but incorrect factual associations can also resemble correct recall internally, suggesting that some apparent truth signals may partly

reflect knowledge retrieval rather than correctness itself [48]. Overall, the evidence supports usable and sometimes causally effective information about correctness, but suggests that “truth” is better viewed as a family of task-dependent internal signals than as a single model-wide variable.

For **uncertainty and confidence**, models appear to encode information about whether an answer is likely to be correct, but confidence is not simply another name for correctness. Answer success can be predicted from internal states before generation and, in some settings, more accurately after an answer has been produced [39,161]. Causal analyses further suggest that these uncertainty signals partly overlap with the computation used to produce the answer: interventions that change answer correctness often produce corresponding changes in internally decoded confidence [41]. Other work finds useful uncertainty information in the stability of hidden states under perturbation and in task-relevant attention-head activations during in-context learning [184,185]. At the same time, confidence itself can be manipulated independently of correctness. Steering sparse middle-layer activations can make models more decisive and less likely to abstain without reliably making their answers more accurate [42], while adaptive methods use internal confidence to choose between answering and abstaining [183]. Behavioral measures such as semantic entropy and verbalized uncertainty also do not always align with factual accuracy [186–188]. Thus, the literature suggests that uncertainty and confidence reflect several related signals—including expected correctness, representational stability, and willingness to answer—rather than a single internal quantity.

For **hallucination**, internal states often contain enough information to distinguish factual from hallucinated generations, but the literature does not support a single mechanism that produces hallucinations. Hallucination-related signals can be decoded from final-token hidden states, claim-level representations, and layerwise hidden-state dynamics [35,37,43,44]. More detailed analyses suggest that errors can arise at different stages of information processing. In factual recall, relevant knowledge may fail to become sufficiently represented in MLP activations, or it may be represented but not selected by later attention mechanisms [45]; other work finds that correct answers tend to show sharper internal activation patterns over relevant context [189]. In retrieval-augmented settings, hallucinations are associated with weaker use of external context and stronger contribution from parametric knowledge [46], while structured and irrelevant contexts can introduce competition between different information sources or pathways [47,190,191]. Importantly, strong but incorrect parametric associations can look similar to correct factual recall internally, whereas hallucinations without such associations are easier to distinguish [48]. Failures can also occur when fact-associated neurons are not sufficiently activated during retrieval [51]. Together, these results suggest that hallucination is better understood as a collection of failures in retrieval, selection, grounding, and information competition than as a single internal “hallucination state.”

For **knowledge recognition**, the literature suggests that models can often internally distinguish questions they are likely to answer correctly from those outside their effective knowledge boundary. Answerability and future correctness can be decoded from hidden states before generation, particularly from intermediate-layer representations [39,161]. Similar structure also appears across languages: knowledge-boundary signals are partly shared across language-specific representation spaces and can transfer more effectively after simple geometric alignment [49]. However, knowing and answering are not equivalent. Correct answers can remain among highly ranked output candidates even when the model ultimately produces an incorrect or uninformative response [50], and factual recall depends on whether relevant knowledge-associated neurons are sufficiently activated [51]. Conversely, a strongly retrieved association can still be wrong, producing internal patterns similar to successful recall [48]. These findings support a distinction between knowledge that is stored, knowledge that is accessible in the current context, and knowledge that is ultimately expressed in the output [40]. Thus, incorrect generation should not be treated as direct evidence that the model lacks the relevant information.

Table A1. Current position of epistemic behaviors in our mechanistic-evidence hierarchy. The “strongest evidence reached” column summarizes the highest level supported by existing work; later stages may be supported only partially or in restricted settings.

Behavior	Strongest evidence reached	What is established	Where the literature gets stuck
Truthfulness	Causal involvement	Correctness is decodable from several internal representations, and steering truth-associated activations can change factual outputs.	The computational role remains unclear: successful steering does not show how truth is computed, and truth-related directions do not generalize consistently across tasks and prompts.
Uncertainty	Causal involvement	Correctness-related uncertainty is decodable before generation, and interventions show overlap with answer-producing computation. Confidence-related activations can also causally affect decisiveness and abstention.	The main gap is computational role : it is unclear how expected correctness, uncertainty, and expressed confidence interact, or whether they share one mechanism.
Hallucination	Partial computational role	Mechanistic studies identify several causal failure points, including weak retrieval, failed selection, poor grounding, and competition between information sources.	There is no general mechanism : different settings produce hallucinations through different internal processes, and some hallucinations remain difficult to distinguish from correct recall.
Knowledge recognition	Decodability / limited causal evidence	Answerability and future correctness are decodable before generation, and some interventions can recover or suppress available factual information.	The literature has not established a general computational role : it remains unclear how stored knowledge becomes accessible, retrieved, and ultimately expressed.

Appendix B.2. Linguistic and Semantic Behaviors

For **syntax and grammatical structure**, mechanistic work shows that grammatical information is distributed across the network but can depend on smaller sets of functionally important components. Subject number, agreement, animacy, and competing parses are recoverable from neurons, attention heads, residual-stream states, and sparse features, and interventions on these representations can change grammatical predictions [52–54,65,167,192,193]. Some tasks have been traced further. In indirect-object identification, different attention heads detect repeated tokens, transmit positional information, suppress incorrect candidates, and copy the selected name to the output [21]. Similar functional organization appears across languages and related models, even when the exact components differ [56,167]. This distinction is important: syntax appears to reuse *types of computation* more consistently than individual neurons or heads. Syntactic specialization increases during training, yet highly syntax-sensitive neuron identities can differ substantially across random seeds, while agreement-related units show only partial overlap across constructions and languages [54,54,57,194]. Thus, current evidence supports reusable grammatical computations without implying a single universal syntax circuit.

Mechanistic studies also give some insight into **how syntactic structure develops and is used**. Sparse-feature analyses of garden-path sentences find that models can simultaneously represent competing parses, while targeted feature interventions shift which continuation is preferred; however, the same features contribute little to later comprehension questions, suggesting that syntactic representations are not necessarily reused across tasks [53]. Training studies similarly suggest that grammatical structure becomes increasingly differentiated with exposure and scale [57]. Work on synthetic grammars finds parallel learning of grammar substructures and persistent difficulty with deeper recursion,

but does not localize or causally establish an internal grammar algorithm [195]. Formal-language pretraining can leave reusable attention-head subnetworks that later support natural-language grammaticality, although what those heads compute is not established [196]. Semantic information can also influence structure choice: animacy representations causally affect active–passive preferences through residual states and opposing attention-head effects [192]. Similar neuron interventions affect some psycholinguistic behaviors, such as sound–gender association and implicit causality, but not others [168]. Together, these findings place syntax beyond simple decodability and into **causal and partial computational explanation**, while the main unresolved question is how stable these computations are across tasks, models, and training runs.

For **semantic and factual information**, the clearest mechanistic pattern is a staged process of *representation, enrichment, routing, and selection*. Early work localized individual factual associations to FFN neurons and showed that suppressing, amplifying, or editing them changes factual predictions [19]. Causal tracing similarly identified middle-layer MLP activity at subject positions as important for factual recall and editing [1]. Later work refines this picture: early or middle MLP updates enrich the subject representation with candidate attributes, relation information reaches the prediction position, and upper-layer attention helps select and promote the requested attribute [58]. Neuron-level analyses likewise distinguish shallower “query” units from deeper units that directly promote answer tokens, although the proposed information flow is only partly causally established [197]. Other studies identify mover heads, MLP amplification or suppression, and later output control as additional stages of factual retrieval [59,198]. **These findings help reconcile an apparent disagreement in the literature: the component where a fact can be edited is not necessarily the complete mechanism by which that fact is retrieved and expressed.**

This staged account is also not universal. A memory-promoting attention head can strongly control country–capital recall yet fail to transfer to a broader factual dataset, showing that apparently interpretable mechanisms may be domain-specific [60]. In multi-hop reasoning, activating fact-associated neurons improves some incorrect answers and suppressing them harms many correct ones, while chain-of-thought prompting is associated with stronger activation of the relevant facts; nevertheless, correct answers can still occur through apparent shortcut pathways [51]. Factual neurons can also overlap across languages and provide useful targets for multilingual editing, but successful editing does not reliably transfer to reasoning over the edited fact [199]. Sparse circuit methods further decompose some MLP computations into interpretable features, but remain approximations rather than complete circuits [166]. A related pattern appears in conceptual definition tasks, where MLP activity is associated with content-level concept information while attention and later hidden states increasingly integrate information toward the final prediction [200]. Overall, factual and semantic work provides substantial **causal evidence and partial computational roles**, but not a single storage or retrieval mechanism that applies to all knowledge.

For **entity and relational binding**, mechanistic work suggests that retrieval relies on several complementary internal signals. Entity and relation indices can be decoded from low-dimensional middle-layer subspaces, and steering these indices changes which bound attribute is retrieved [61]. Other work finds that models combine at least three signals during in-context retrieval: a positional index of the relevant group, a lexical pointer to the queried entity, and a more direct pointer to the target entity [62]. Their relative importance changes with context position, which argues against a single positional addressing mechanism. A more component-level analysis of entity tracking identifies attention-head groups involved in detecting the correct position, transmitting that position, and fetching the associated value; fine-tuning improves performance largely by strengthening this existing mechanism rather than replacing it [63]. Selection can also involve suppression: copy-suppression heads detect an already-supported contextual token and write against its output direction when copying it would be harmful [64]. Across tasks, some of these motifs are reused: selection-and-copying components overlap strongly between indirect-object identification and colored-object retrieval, although other tasks rely on different circuits and exact head identities become less stable at larger

scales [55]. **Binding therefore has relatively strong causal and functional evidence, but no single retrieval algorithm explains all contexts.**

For **multilingual behavior**, the literature supports a combination of *shared representations and language-specific control*. Morphosyntactic properties can be encoded in overlapping dimensions or sparse features across languages, with greater sharing often found between linguistically similar languages [54,65]. Activation patching further shows that concept information can be transferred across languages while allowing the surrounding model computation to express that concept in another language [66]. Representational analyses similarly find increasing cross-language semantic alignment in parts of the network [69]. At the same time, language identity can be localized to selective FFN neurons, hidden-state dimensions, or SAE features, and manipulating these representations can steer the language of generation [67,71,72,165,201–203]. Thus, shared semantic information and language-specific representations coexist rather than forming competing explanations.

There is, however, disagreement about **where language-specific processing occurs**. Some neuron studies find strong specialization in both early and late layers [67,71], whereas others find language-sensitive neurons or sparse features concentrated mainly in middle-to-late or late layers [72,165,202]. Cross-layer feature tracing provides a possible reconciliation: early layers can retain language-specific input information, middle layers contain more shared multilingual features, and later layers again become language-specific as the model realizes the output language [204]. Language-confusion errors are consistent with the final part of this account: correct generations show a late shift toward target-language predictions, while confused generations often fail to make this transition [70]. However, the broader “shared semantic middle” account is still only partly causal. Parameter-ablation work identifies core multilingual and language-specific regions that are important for language modeling, but does not establish what computation those regions perform [205]. Moreover, manipulating language-specific neurons successfully controls output language but does *not* reliably improve cross-lingual task transfer [68]. This is an important qualification: **a representation that controls the language of expression is not necessarily the representation that supports cross-lingual reasoning or task transfer.**

Across linguistic and semantic behaviors, a common picture emerges: **distributed information is often acted on through more localized routing and control mechanisms**. Syntax circuits detect and propagate grammatical information; factual recall enriches and routes candidate knowledge; binding mechanisms select among contextual entities; and multilingual models separate, to varying degrees, shared content from language-specific realization. The literature therefore frequently reaches **causal involvement** and, for narrow tasks, **partial computational roles**. The main remaining bottleneck is **generalization**: detailed mechanisms are often established for controlled tasks and particular models, while the stability of the same computations across tasks, architectures, languages, and training runs remains much less certain.

Table A2. Current position of linguistic and semantic behaviors in our mechanistic-evidence framework. “Strongest evidence reached” denotes the highest level supported in at least some studies; it does not imply that all work within the behavior family reaches that level.

Behavior	Strongest evidence reached	What is established	Where the literature gets stuck
Syntax & grammar	Partial computational role	Causal circuits and features support operations such as agreement tracking, duplicate detection, inhibition, copying, and parse selection.	Similar functions recur across settings, but the exact neurons and heads vary across models, languages, constructions, and training runs. A general syntax algorithm is not established.
Semantic & factual recall	Partial computational role	Causal evidence supports a staged process involving MLP/FFN enrichment of factual information followed by routing and selection toward the output.	Studies localize different stages to different components. Storage, retrieval, editing, and expression are not yet unified into one general mechanism, and some identified components are strongly domain-specific.
Entity & relational binding	Partial computational role	Position, entity identity, relation information, copying, and suppression have identifiable causal roles in selecting bound information.	Different tasks and context positions use different mixtures of these signals. Lower-level circuits implementing the proposed representations remain incomplete.
Multilingual behavior	Causal involvement / partial computational role	Shared grammatical or conceptual information coexists with language-specific neurons, dimensions, and sparse features that can causally steer output language.	The proposed shared-middle / language-specific-output process is only partly causal, layer localization differs across studies, and language control does not reliably imply cross-lingual task transfer.

Appendix B.3. Reasoning and Cognitive Behaviors

For **in-context learning (ICL)**, the literature increasingly suggests that there is no single mechanism underlying all forms of adaptation from context. Early work identified induction heads that match the current input to earlier context and copy the associated continuation, with particularly strong causal evidence in small attention-only models [73,74]. Later work separates this copying operation from richer task learning. Task-conditioned information can be concentrated in function-vector heads or sparse task features, and injecting these representations can cause models to perform the corresponding task without the original demonstrations [75,77,206]. Classification studies further identify a progression in which demonstration information is accumulated at label or answer positions and later used to predict the query label [76,207–209]. More recent analyses separate this computation into distinct stages: early processing increases task-relevant separability, later heads align representations with output labels, and partially distinct head groups support *task recognition* and *task learning* [79,80].

These findings also resolve an important disagreement in the ICL literature. Induction heads were initially proposed as a major mechanism for ICL, but later ablations show that, for few-shot task accuracy in larger tested models, **function-vector heads can matter substantially more than “pure” induction heads**; induction heads remain more important for context-dependent next-token loss reduction [78]. Thus, different operational definitions of ICL can identify different mechanisms. False demonstrations provide another example: correct and incorrect prompts behave similarly early in processing, but late induction-like heads can retrieve and promote incorrect labels, causing the model to “overthink” an initially correct prediction [210]. Induction itself also appears to contain multiple routes: concept-induction heads support semantic transfer such as translation, while token-

induction heads support verbatim copying [211]. ICL therefore appears to involve several reusable operations—matching, aggregation, task recognition, routing, and execution—rather than one universal induction circuit. The literature reaches strong causal involvement and partial computational roles, but the complete mechanism and its generalization beyond curated ICL tasks remain unresolved.

For **arithmetic and numerical reasoning**, several studies support a staged computation in which input information is first represented or routed and later transformed into answer-related information. Causal mediation finds that early MLP states carry operand- and operator-related information, intermediate attention routes this information toward the prediction position, and later MLPs contribute result-related information [81]. Detailed circuit analysis of greater-than behavior similarly finds attention heads that relay the comparison value and later MLPs whose combined outputs favor larger numbers [82]. However, **the existence of structured numerical representations does not imply that models implement a general arithmetic algorithm**. In the greater-than circuit, removing linearly extractable numerical structure does not eliminate the behavior, and the computation could reflect learned lookup-like structure rather than abstract comparison [82]. Broader arithmetic work instead finds sparse middle- and late-layer neurons associated with operand ranges, modular patterns, digit properties, and other heuristics; targeted ablations show that these neurons causally contribute to the corresponding arithmetic examples [85]. Related analyses identify operation-sensitive attention heads and deeper answer-promoting FFN neurons, again supporting a distributed staged process rather than a single arithmetic module [83].

Other numerical behaviors show related but distinct mechanisms. Successor heads use structured numeric features, including mod-10-like representations, to promote the next element of ordinal sequences, although these features explain only part of the head's behavior [212]. Counting studies find transferable latent count information that accumulates across item positions, with separators and final-item states often carrying strong count signals [86]. Explicit decomposition can extend this computation: partition-level counts are represented at boundary tokens, transferred into generated intermediate counts, and then aggregated to produce a final answer [87]. Numerical-sequence extrapolation likewise shows decodable first-difference representations and causal evidence for induction-like retrieval of the next difference [88]. **Across numerical tasks, causal evidence is strong, but the computational picture favors multiple task-specific heuristics, counters, and retrieval mechanisms over one general numerical algorithm.**

For **multi-hop and compositional reasoning**, a recurring finding is that intermediate answers can exist internally before the final answer is produced. In factual two-hop questions, the first-hop or bridge entity becomes recoverable in intermediate states, and suppressing this signal can reduce the probability of the final answer [89,213]. More detailed tracing supports a sequential process: early layers resolve the bridge entity, middle-layer attention propagates it toward the final position, and later layers use it to retrieve the target answer [90]. Failures often occur when these stages are poorly timed or coordinated. Incorrect cases tend to resolve intermediate entities later, and moving later representations back to earlier layers can repair many failures [90]. Other work qualifies a simple “late first hop” explanation: the second-hop machinery can still activate while high-layer representations continue to over-promote the intermediate answer, motivating mechanisms that expose higher-layer information to earlier computation [91]. Thus, **multi-hop errors can arise from failure to form, preserve, route, or appropriately suppress an intermediate result**, rather than simply from absence of the required facts.

Logical reasoning provides one of the clearest narrow circuit accounts. In controlled propositional tasks, attention-head families causally support a sequence from query interpretation to relevant-rule selection, fact processing, and final decision [214]. The broad functional stages recur across several pretrained models, although exact circuitry changes with scale and a separately trained small model uses a different implementation. Other automated circuit analyses also recover compact task-relevant subgraphs and suggest progressive entity, relation, and decision processing, but the proposed reasoning stages are inferred from influence patterns rather than directly validated as computations [215].

Multi-hop and logical reasoning therefore reach causal and partial computational explanations in controlled settings, while generalization to longer, open-ended reasoning remains the major gap.

For **chain-of-thought (CoT) and latent reasoning**, current mechanistic evidence supports a causal role for reasoning-conditioned internal states, but does not show that the visible rationale faithfully describes the underlying computation. In a controlled ontology task, relation information is recoverable in earlier and middle layers while later attention heads increasingly contribute to answer writing through multiple pathways [216]. SAE analyses similarly find CoT-associated features whose transfer into no-CoT runs can increase correct-answer likelihood in a larger tested model, but these interventions do not establish that individual features correspond to particular written reasoning steps [92]. Sequential patching provides stronger evidence that coordinated CoT-conditioned head activity contributes to final-answer generation, yet the same heads also support grounding, answer-slot binding, and arithmetic operations [93]. **The causal representation is therefore not necessarily a “reasoning-only” circuit.**

Continuous latent reasoning strengthens the distinction between intermediate computation and textual explanation. Latent reasoning states can be partially mapped to explicit CoT representations and carry recoverable lexical information, while truncating or editing early latent states changes final answers [94]. This supports genuine intermediate state representations, but still does not identify the component-level algorithm operating over them. Across these studies, **CoT changes internal computation and can provide causally useful states, but current evidence does not establish step-by-step faithfulness between generated rationales and the model’s internal reasoning process.**

Finally, work on **planning, retrieval, and capability reuse** shows that reasoning depends on supporting processes that should not themselves be mistaken for reasoning. In Blocksworld, future actions can be decoded several steps before they are generated, but interventions mainly establish that goal and history information affect later decisions rather than proving that the decoded future actions constitute a causally used plan [95]. Training-based experiments provide stronger evidence that future-token losses can teach models to precompute information for later positions, although natural-language results leave open whether this reflects deliberate pre-caching or features already useful for the current prediction [96]. Long-context retrieval is more directly localized: sparse retrieval heads causally redirect relevant context into generation, and ablating them harms retrieval and CoT performance [97]. This shows that reasoning can depend on specialized retrieval mechanisms without implying that those heads perform the reasoning themselves. Fine-tuning provides a related distinction between capability and expression: intermediate representations of pretrained procedural capabilities can remain decodable after the behavior is suppressed, and pruning small sets of components can sometimes restore the original capability [171]. **Together, these results separate having a latent capability, retrieving the information needed for it, planning future computation, and behaviorally expressing the result.**

Across reasoning behaviors, the strongest recurring pattern is **staged computation**. Information is encoded, routed between positions, transformed into intermediate task or answer representations, and finally converted into an output. Many narrow tasks now reach **causal involvement** and some reach **partial computational role**. The main bottleneck is **generalization**: detailed mechanisms are typically recovered for controlled tasks, while it remains unclear whether the same computations explain open-ended reasoning across models, prompts, task complexity, and scale.

Table A3. Current position of reasoning and cognitive behaviors in our mechanistic-evidence framework. “Strongest evidence reached” describes the strongest evidence available in the literature, not the evidence level reached by every study in the category.

Behavior	Strongest evidence reached	evidence	What is established	Where the literature gets stuck
In-context learning	Partial role	computational	Causal evidence supports matching/copying, task-vector transport, label-information aggregation, and partially separable task-recognition and task-learning operations.	There is no single ICL mechanism: induction-, function-vector-, and task-feature accounts depend on the behavior and metric being studied. The complete natural ICL circuit remains unresolved.
Arithmetic & numerical reasoning	Partial role	computational	Attention routes numerical information while MLPs, neurons, and latent counter-like states causally contribute to result formation, comparison, counting, and sequence continuation.	It remains unclear when these components implement general algorithms versus overlapping learned heuristics. Detailed mechanisms generalize poorly beyond restricted numerical formats and tasks.
Multi-hop & logical reasoning	Partial role	computational	Intermediate entities or results can be decoded and causally manipulated; controlled tasks reveal stages for retrieval, propagation, rule/fact selection, and decision.	The exact transformations within these stages remain incomplete, and causal circuits are largely restricted to short, controlled reasoning problems.
Chain-of-thought & latent reasoning	Causal involvement / partial computational role		CoT-conditioned and latent intermediate states can causally affect final answers, and latent states contain recoverable reasoning-related information.	Causal relevance does not establish faithfulness : current work does not show that textual CoT steps correspond one-to-one with the internal computations producing the answer.
Planning & retrieval	Causal involvement		Future information is sometimes decodable before generation, and specialized retrieval mechanisms causally supply context needed for later predictions and reasoning.	Planning representations themselves have weaker causal support. It remains unclear whether decoded future actions constitute an internally used plan rather than correlated predictive information.

Appendix B.4. Safety and Alignment Behaviors

For **refusal**, some of the strongest causal evidence in the safety literature comes from activation-space interventions. A single harmful-versus-harmless residual-stream direction can be sufficient to strongly suppress refusal when removed and induce refusal on harmless prompts when added, with similar effects across several model families [98]. Related work recovers steerable refusal directions even when behavioral outputs provide little contrast and finds that useful directions, layers, and intervention strengths vary across models [99]. However, later studies challenge the interpretation that refusal itself is one-dimensional. Refusal can instead be mediated by multiple independent directions or a higher-dimensional cone, and suppressing several directions can have stronger effects than removing the original one-dimensional direction [100,101]. Refusal-related information can also be distributed across many activation dimensions even when a single dense direction provides effective steering [217]. Training models while probabilistically removing a refusal direction can improve robustness to several jailbreaks, but the resulting alternative safety mechanism has not itself been mechanistically identified [218]. **Thus, a low-dimensional intervention can be an effective causal control point without establishing that the underlying refusal representation or computation is itself one-dimensional.**

This distinction becomes clearer for **over-refusal**. Appropriate harmful refusal can be comparatively well captured by a global direction, whereas inappropriate refusal of benign tasks is more strongly conditioned on task identity and occupies higher-dimensional, task-specific geometry [102]. Task-aware centroid directions can therefore reduce over-refusal more selectively than global refusal steering, although these representations are model-specific and do not transfer reliably [103]. Nonlinear energy-based steering similarly separates desirable from undesirable response states and improves benign compliance, but the learned energy surface is external to the model and does not establish its native refusal mechanism [219]. Another account separates safety judgment from willingness to answer: benign/toxic and answer/refuse labels are encoded along nearly orthogonal directions in several models, and weight edits that align them improve both jailbreak and over-refusal metrics [107]. **These findings explain why weakening “refusal” globally can improve over-refusal while simultane-**

ously damaging safety: harmful refusal and task-conditioned over-refusal only partially share a mechanism.

More broadly, **safety alignment appears to separate recognizing harmful content from deciding how to respond to it.** Representation-engineering studies show that harmfulness, honesty, morality, and other safety-related properties are linearly recoverable and steerable from distributed activations, although these directions are not complete explanations of the underlying computation [220]. Alignment training also modifies a sparse but distributed set of MLP neurons: patching activations from a safety-aligned model into a less-aligned model can recover much of the safety behavior, while substantial overlap between safety- and helpfulness-related neurons helps explain an alignment trade-off [221]. Layerwise analyses suggest that harmful-versus-normal inputs become separable early, including in base models, while alignment changes how these representations are associated with refusal- or compliance-related states in intermediate layers [104]. Consistent with this separation, safety information from an aligned source model can steer another model even when the target model itself already represents harmful intent but does not reliably refuse [222].

Other studies reach a similar conclusion from different levels of analysis. Preference-associated activation differences can be used to steer models toward broadly preferred outputs, but do not isolate separate mechanisms for helpfulness, truthfulness, or harmlessness [223]. Unsafe-response prediction improves when harmful-input recognition and likely compliance are modeled as separate internal signals rather than as one “unsafe” variable [106]. Category-specific safety directions can steer generation away from harmful content without always producing blanket refusals, although their effects and utility trade-offs vary considerably across categories and models [224]. Harmfulness itself is strongly decodable from middle-layer representations and sparse activation dimensions across several models, including during generated reasoning, but these predictive dimensions have not been causally linked to the model’s own safety decisions [225]. Together, these findings support a staged picture of **harm recognition → response policy → refusal or compliance**, with stronger causal evidence for controlling the later policy than for explaining how the full decision is computed.

For **jailbreak susceptibility**, several studies find that successful attacks alter the internal geometry associated with harmfulness and refusal. Harmful prompts that jailbreak successfully often move toward regions occupied by harmless prompts [226]; across many attack types, jailbreak-related directions share a common component that tends to oppose harmfulness representations, and removing this harmfulness component weakens—but does not eliminate—their effects [108]. Learned perturbations can similarly move malicious representations across a benign/malicious decision boundary [227]. Gradient-based attacks also appear to converge toward compliance-associated activation changes, although this account is correlational because the proposed compliance direction itself is not intervened on [228]. Activation patching provides stronger evidence for a middle-layer refusal “stance” concentrated at influential token positions: replacing these states changes refusal-token probabilities, and attacks targeting the same representation increase harmful compliance [109]. Learned low-rank interventions can instead move jailbreak representations toward representations of harmful inputs that the model normally rejects, substantially restoring refusal [110].

However, **harmfulness or refusal suppression is not a complete explanation of jailbreaks.** Neuron-level work finds causal rejection-versus-compliance signals in selected MLP components [229], and dynamically masking refusal-attributed hidden dimensions can increase harmful compliance [230]. Yet attribution-graph analysis finds that simple feature suppression does not distinguish successful from failed attacks; in a narrow set of semantic jailbreaks, success is instead associated with changes in computational pathways, and ablating the most emergent features fails to restore refusal [112]. Contextual reframing provides another observational mechanism: domain context can move harmful queries into an intermediate “gray zone” and reduce attention to sensitive tokens, but no causal intervention verifies that this produces the jailbreak [114]. Unlearning work similarly suggests that jailbreaks can exploit distributed parameter pathways that differ from those used by direct harmful prompts and converge only later toward harmful output semantics [113]. Steering-vector audits further

show that unrelated behavioral steering can weaken safety when its activation direction conflicts with refusal representations; removing a multidimensional refusal-aligned subspace mitigates these effects more than removing a single direction [111]. Behavioral defense analysis also distinguishes broad increases in refusal from better harmful-versus-benign discrimination, but does not localize either process internally [231]. **The emerging picture is therefore that many jailbreaks weaken safety-relevant representations, but others can exploit alternative routing, context, or response-policy pathways; there is no single jailbreak mechanism.**

For **truthfulness, honesty, and deception**, the evidence is more uneven. Truth-related information is decodable from selected attention heads and internal activations, and steering these signals can improve or reduce factual outputs [3,34]. Adaptive intervention further shows that the best steering strength varies across questions and that internal response states can help select between answering and abstaining [183]. Honesty can also be shifted using depth-distributed activation directions derived from honest-versus-dishonest examples [115]. These results establish causal control over truthful or honest reporting, but do not identify a complete truth- or honesty-producing computation.

Evidence for **deception mechanisms** is substantially weaker. Under explicit instructions to answer incorrectly, factual labels remain decodable while SAE features shift with the deceptive instruction, but there is no causal evidence that these features produce the false answer [232]. Prompted deceptive arguments are also linearly decodable from middle-layer states and appear distributed over many directions [116]. Trajectory-level analyses find different geometric signatures for sycophancy, confabulation, and strategic or instructed deception, but these signatures vary strongly across models and can disappear entirely for instructed lies [118]. Stress-testing deception probes further finds strong in-distribution decoding but poor cross-domain and stylistic robustness, with no evidence for a single statistically stable low-dimensional deception subspace [117]. Activation explainers can recover useful evidence for strategic deception from reasoning-model states, but again without showing that the decoded information causes the behavior [163]. Even near-perfect separability after fine-tuning on incorrect responses establishes primarily that the training conditions produce distinguishable representations, not that the model contains a general dishonesty mechanism [233]. **Thus, truthfulness has reached causal steering, whereas deception research remains dominated by decodability and detection; evidence for a causal deception mechanism is still limited.**

For **sycophancy**, mechanistic work increasingly rejects a single monolithic representation. User-provided incorrect opinions can alter the normal development of a correct-answer preference in later layers, and interventions on the identified states can shift the resulting answer [164]. A sparse set of middle-layer attention heads also carries particularly strong sycophancy signals, and steering those heads can change whether the model reverses a previously correct answer after user disagreement [119]. However, different forms of sycophancy separate internally. Generic agreement and sycophantic agreement can begin similarly and then diverge in middle layers, while excessive praise occupies a more distinct direction; targeted subspace interventions can selectively change one behavior while preserving the others [120]. Even factual and opinion sycophancy show model-dependent overlap: their directions are highly aligned in one tested model but substantially more distinct in another [121]. Reasoning adds another dynamic component: CoT can either reduce or amplify sycophantic answer preference over the course of generation, although these trajectory findings remain correlational [169]. **Sycophancy therefore appears to be a family of agreement- and user-conditioning mechanisms rather than one behavior with one internal direction.**

Finally, **instruction following** provides a broader alignment perspective. Contrastive instruction directions can causally steer format, length, and lexical constraints, even when the corresponding instruction is absent from the prompt [122]. More detailed patching finds that task identity is strongly linearly decodable from instruction-token representations and that patching multiple layers can have superadditive effects on downstream answers [123]. This creates another useful distinction: **a representation can be linearly simple while its causal effect is not additive.** In a reasoning model, harmfulness, refusal, and a separate prompt-CoT “scrutiny” signal peak at different depths; steering

the scrutiny signal changes deliberation and can interact with supplied reasoning to alter refusal [234]. This suggests that safety in reasoning models may depend not only on harmfulness detection or refusal states but also on how internal reasoning is coupled to the final policy.

Across safety and alignment behaviors, the literature repeatedly separates **recognition from response**: models may represent harmfulness, truth, user beliefs, or instructions without necessarily acting on that information in the aligned way. Many behaviors now reach **causal involvement** through activation steering, patching, ablation, or weight editing. Yet the main mechanistic gap is the next step: identifying the computation that converts these representations into context-appropriate behavior. Refusal directions, jailbreak vectors, safety neurons, deception probes, and sycophancy subspaces are often powerful control or detection variables, but **causal controllability should not be equated with a complete computational mechanism**.

Table A4. Current position of safety and alignment behaviors in our mechanistic-evidence framework. “Strongest evidence reached” describes the strongest evidence reported in at least some studies and does not imply that every paper within the behavior family reaches that level.

Behavior	Strongest evidence reached	What is established	Where the literature gets stuck
Refusal & over-refusal	Causal involvement / partial computational role	Refusal-associated directions and subspaces can causally induce or suppress refusal. Harmful refusal and over-refusal have partly different, task-conditioned geometry.	A strong one-dimensional control direction does not establish a one-dimensional mechanism. The computation combining harmfulness, task context, and refusal policy remains unresolved.
Harm recognition & safety alignment	Causal involvement	Harmfulness is strongly decodable, and alignment changes distributed neurons and representations whose manipulation alters safety behavior. Harm recognition and willingness to comply can be partly separable.	The field has not established how harmfulness representations are transformed into a context-appropriate response policy, or how this computation changes across alignment methods.
Jailbreak susceptibility	Causal involvement / partial computational role	Many attacks causally interact with harmfulness-, refusal-, or compliance-related states, and interventions on these states can alter jailbreak success.	No single account explains all attacks. Harmfulness suppression is important but insufficient; alternative routing, contextual ambiguity, and distributed paths also appear relevant.
Truthfulness & honesty	Causal involvement	Truth- and honesty-associated internal representations are decodable and can be steered to change reported answers or honesty-related behavior.	Steering does not identify the computation that determines when represented information is faithfully expressed, and generalization remains task- and model-dependent.
Deception	Decodability	Prompted or labeled deception is often detectable from residual-stream states, multiple linear directions, or activation trajectories.	Probe signals can reflect task regime, style, or answer correctness; cross-domain transfer is weak and causal interventions are largely absent. A deception-producing mechanism is not established.
Sycophancy	Causal involvement / partial computational role	User influence and answer reversal are decodable and causally steerable, with some attention heads and residual subspaces contributing to the behavior.	Sycophancy decomposes into distinct forms whose representations vary across behaviors and models. A general user-belief-to-answer computation has not been identified.
Instruction following	Partial computational role	Instruction information is decodable and causally affects answers; instruction states can steer constraints and multi-layer representations have superadditive downstream effects.	The proposed role of instruction representations as selectors of later circuits is only partly established, and current evidence comes mainly from controlled tasks and limited model families.

Appendix B.5. Social and Affective Behaviors

For **personality and persona**, a recurring result is that broad behavioral tendencies can be represented and controlled through relatively low-dimensional activation directions. Contrastive persona vectors predict subsequent trait expression and causally increase or suppress behaviors such as sycophancy, maliciousness, hallucination, and Big Five traits when added to residual-stream activations [124,126,170]. A broader “Assistant Axis” similarly separates the model’s default assistant-like behavior from alternative roles; moving away from this region increases role adoption and, in some settings, harmful behavior, while activation capping can stabilize the default persona [125].

Trait directions can also be composed and adapted to conversational context [126], and stable trait estimates can be obtained from their activation-space projections even without causal intervention [235,236]. These results provide strong evidence that persona-related behavior is **decodable and causally steerable**, but the vectors are better understood as control directions than as complete personality mechanisms.

Importantly, **linearly steerable does not mean independently represented**. Big Five directions share substantial low-rank structure [170], and steering one trait systematically changes others even after the directions are explicitly orthogonalized [127]. Situation-specific analyses further show that the neuron sets associated with the same trait change across conversational contexts [128]. SAE interventions identify smaller sets of trait-associated features, but their semantic specificity and causal roles remain incomplete [237]. A broader trait audit likewise finds that some behaviors are naturally expressed, some are readily steerable, and others are difficult to recover with a standard persona-vector procedure; simultaneous trait steering can also produce non-additive or destructive interactions [238]. More detailed circuit analysis provides a possible reconciliation: opposing levels of a personality trait can reuse many of the same nodes while differing more strongly in the edges connecting them, with a few early MLPs acting as causal bottlenecks [129]. **Thus, simple activation directions can control personality expression even when the underlying implementation is distributed, context-dependent, and interaction-heavy.**

This distinction is also visible for more specific social traits. Steering a small set of SAE features associated with Dark Triad behavior separately changes exploitation, callousness, aggression, and moral-harm endorsement, while leaving measured strategic deception largely unchanged [239]. Individual features have different behavioral profiles and combinations produce effects absent from single-feature interventions, suggesting partially separable pathways rather than one antisocial-persona mechanism. Personalization shows a similar pattern at a different level: sparse, user-specific attention-head sets causally affect reference likelihood and personalized generation, but the selected heads differ substantially across users [240]. These findings reinforce that **social dispositions often have useful low-dimensional control variables without sharing a single fixed circuit.**

For **sentiment and emotion**, the literature spans some of the earliest highly localized findings to recent distributed accounts. In a review-trained mLSTM, a single cell strongly tracks positive versus negative sentiment and clamping it changes generated review polarity [130]. Modern Transformers also contain approximately linear sentiment directions, but the computation is more distributed: sentiment can be written into intermediate summary positions such as punctuation and later read out by attention, and ablating the sentiment direction at these positions substantially reduces classification performance [131]. Contrast-derived activation vectors can similarly shift sentiment and toxicity during generation [241]. **The apparent difference between a single sentiment neuron and a distributed sentiment pathway is therefore largely architectural rather than a direct contradiction:** both identify causally useful low-dimensional signals, but Transformer implementations distribute their computation across representations and routing components.

For richer **emotion inference**, several studies find that emotional information becomes strongly decodable around intermediate layers and that interventions there can change emotion predictions [135,242]. Mid-layer attention appears to gather emotionally relevant information from the prompt, followed by FFN processing, while appraisal-related directions can shift specific emotion judgments [135]. However, individual “emotion neurons” form overlapping, distributed populations and their ablation effects vary substantially by emotion and model [242]. Cross-dataset work similarly finds strong residual-stream emotion decodability but limited overlap among individual emotion neurons, suggesting that the representation is more stable than its exact unit-level realization [136].

Recent geometric analyses sharpen this picture. Emotion categories form structured internal spaces organized partly by valence and arousal, and these spaces can transfer across datasets, models, and sometimes languages [132,133,243]. Yet the dimensionality claims differ: some analyses recover a compact low-dimensional affective subspace [132], while others find broadly similar valence–arousal

geometry embedded in a diffuse, higher-rank representation [133]. These are not necessarily incompatible: a few dimensions may capture the dominant affective organization while the full representation remains distributed. SAE analyses similarly identify compact feature subsets that predict human affect ratings and causally steer emotional expression, without establishing a complete circuit [243]. Learned state-conditioned modulation can also steer sentiment and multiple emotion categories, but because the controller is externally trained, it demonstrates controllability rather than the model's native emotion mechanism [244].

Another important distinction concerns **emotion detection versus emotion categorization**. Emotional-versus-neutral content can become decodable very early, while identifying *which* emotion is present peaks later and is more sensitive to lexical cues and model scale [134]. This offers a useful explanation for apparently different layer-localization results across emotion studies: they may probe different stages of the process. Related work finds that explicit emotion instructions can determine how long emotion information remains visible through later layers [136]. In a frontier model, emotion-concept directions evolve from local affective associations toward more context-integrated, response-relevant representations and causally alter preferences, blackmail, reward hacking, sycophancy, and harshness [245]. **Emotion representations therefore appear to participate not only in emotion labeling, but also in later decision and response-policy computation**, although the mediating circuits remain unresolved.

For **empathy and theory of mind**, evidence is more mixed. Empathy-related examples are almost perfectly linearly separable within several tested models, and steering their model-specific directions changes empathy-related behavior [138]. However, the directions transfer poorly across models and may partly capture reduced task focus rather than empathy itself, so the underlying construct remains uncertain. For theory-of-mind-like false-belief tasks, perturbing a sparse set of task-sensitive parameters—particularly query/key parameters associated with positional processing—reduces performance relative to random perturbations [246]. Yet the same perturbation also affects contextual localization and general language performance, making it difficult to interpret these parameters as a dedicated belief-reasoning mechanism. CoSTOM instead finds belief, desire, and intention information decodable mainly in shallow or shallow-to-middle layers and uses supervised activation-mediated training to preserve this information and improve negotiation and persuasion behavior [137]. **Thus, mental-state information is clearly recoverable and can be made behaviorally useful, but a specific computation for inferring another agent's beliefs or intentions has not been isolated.**

For **political and social attitudes**, ideology is strongly linearly recoverable from selected attention-head activations. Probe directions trained on U.S. lawmakers generalize to news-media slant and causally alter political framing in generated text [139]. Related interventions shift political-bias classifications and neutral rewriting [141]. However, **the same ideological direction does not uniformly control all political behavior**: simulated voting preferences are substantially less consistent under intervention than discourse-level ideology judgments [141]. This suggests that ideological framing and candidate choice share only part of their internal representation. Multilingual political-bias steering similarly changes measured ideological outputs but varies substantially by model, language, and ideological axis [140]. Neuron-level fine-tuning analysis further separates general political neurons from topic-specific neurons; patching the former changes stance broadly across topics, while patching the latter has more localized effects [247]. These findings provide causal evidence for both shared and topic-specific political representations, but not a single general ideology mechanism.

For **values and morality**, many studies again find linear internal structure with causal steering effects. Activation interventions can shift the ethical framing of moral reasoning toward duty- or consequence-based arguments, although early work changes rationales more clearly than final decisions [248]. Classifier-derived value vectors causally increase responses consistent with Schwartz values across several model families [143], while SAE interventions reveal that features affecting one value often influence several others [146]. Moral Foundations Theory provides a more structured case: foundation-specific directions are decodable across many models, sparse features provide finer

semantic decomposition, and both dense and sparse interventions causally shift matched moral judgments [142]. However, the internal geometry frequently **does not reproduce the canonical human grouping of moral foundations**, indicating that successful value decoding should not be interpreted as human-like moral structure.

Other work reaches similar conclusions. Bilingual moral directions can steer virtue-versus-vice responses and safety behavior, but probe accuracy remains moderate and evidence is concentrated in one model family [249]. Abstract value probes transfer to concrete event interpretation and decision settings, and steering changes concrete value judgments, while abstract judgments themselves are comparatively resistant [250]. This suggests a distinction between representing an abstract value and using it in a decision. Cultural-value steering likewise reveals substantial coupling between supposedly separate cultural axes [147]. **Values therefore appear linearly accessible and causally influential, but their representations are coupled, context-dependent, and not necessarily organized according to the human taxonomies used to label them.**

For **culture**, there is a notable disagreement over localization. Entropy-based neuron selection finds culture-specific FFN neurons concentrated mainly in upper layers, with ablation increasing perplexity most strongly on the corresponding cultural corpus [148]. SAE analysis instead identifies sparse cultural features and task-sensitive intermediate-layer bottlenecks whose ablation and steering affect culturally conditioned question answering [251]. A separate gradient-attribution study finds sparse culture-general and culture-specific neurons concentrated mainly in shallow-to-middle MLP layers and reports weak effects when attempting to reproduce the upper-layer “pure culture neuron” account under its own setup [149]. **This is a genuine unresolved disagreement:** different neuron-selection methods, datasets, model variants, and behavioral endpoints identify different layers. The safest conclusion is that culturally relevant information is causally distributed across sparse components, while its precise localization remains method- and task-dependent.

For **stereotypes and demographic bias**, mechanistic evidence increasingly moves beyond output-level bias measurement. Warmth and competence information is highly decodable from a sparse set of attention heads, and steering those heads changes stereotype-related descriptions [145]. Other work finds sparse attribution-selected “bias neurons” whose pruning improves bias benchmarks and sometimes instruction robustness [252]. Gender-bias analyses identify lower-layer gender-associated neurons, attention pathways that propagate their information, and deeper prediction components; masking selected elements changes bias measures, although the detailed circuit is established mainly through case studies [162]. Observable-propagation analysis provides stronger evidence for reuse: the feature associated with gendered pronoun prediction is highly aligned with that used in occupational gender bias, and steering the latter also changes pronoun predictions [152]. **This reveals an important alignment trade-off: a feature implicated in undesirable bias can also support a legitimate linguistic behavior.**

Healthcare studies provide another example of behavior-specific structure. Gender information can be sharply localized to particular MLP activations and causally transferred into clinical generations, whereas race representations are more distributed and often require multi-layer interventions [150]. SAE analysis identifies race-associated features that can change assigned race and some downstream judgments, but ablating those features has little effect on several realistic clinical tasks [151]. Thus, a controllable demographic representation is not necessarily the primary cause of downstream bias. Finally, stereotype-pole directions derived from attention heads can qualitatively steer social-group descriptions [253], but the current evidence is too limited to establish that these directions modify underlying stereotypic beliefs rather than surface wording.

Across social and affective behaviors, a consistent distinction emerges between **representation and organization**. Traits, emotions, values, attitudes, and stereotypes are often surprisingly easy to decode and steer using simple directions, sparse features, or small component sets. Yet these same behaviors frequently show cross-trait coupling, context dependence, multiple processing stages, and model-specific implementations. The literature therefore often reaches **causal involvement** and, in

a few controlled settings, **partial computational role**. The major unresolved questions are whether these intervention targets correspond to the model’s naturally used computations and whether the same mechanisms remain stable across contexts, models, cultures, languages, and more naturalistic social interaction.

Table A5. Current position of social and affective behaviors in our mechanistic-evidence framework. “Strongest evidence reached” denotes the strongest evidence available in at least some studies and does not imply that every paper within the behavior family reaches that level.

Behavior	Strongest evidence reached	What is established	Where the literature gets stuck
Personality & persona	Partial computational role	Trait and persona information is decodable and causally steerable through residual directions, sparse features, neuron sets, and compact circuits.	Linear control directions do not imply independent traits. Traits interact, vary across situations, and can reuse components with different routing. A general computation producing persona-consistent behavior is not established.
Sentiment & emotion	Partial computational role	Sentiment and emotion are strongly decodable; causal interventions identify emotion-relevant directions, neurons, mid-layer attention/FFN processing, and information-routing through intermediate states.	Results range from single-unit to distributed-manifold accounts. Exact localization varies with architecture, prompt, task, and whether the study measures affect detection, emotion categorization, or generation.
Empathy & theory of mind	Causal involvement	Empathy- and BDI-related information can be decoded, and interventions can change empathy-related generation or strengthen mental-state information used in social dialogue.	Model-specific directions transfer poorly, interventions can affect broader contextual processing, and no specific algorithm for inferring another agent’s mental state has been identified.
Political & social attitudes	Causal involvement / partial computational role	Ideological information is decodable from attention heads and neurons; steering changes political framing, while general and topic-specific stance components can be causally separated.	Effects do not generalize uniformly across political behaviors: discourse framing transfers more reliably than voting or candidate choice, and results depend strongly on model and task.
Values & morality	Causal involvement	Moral-foundation and value information is linearly recoverable, and dense or sparse interventions causally change value-consistent judgments and choices.	Internal relationships among values often differ from human theoretical taxonomies, multiple values interact, and steering does not reveal the computation underlying moral or value-based decisions.
Culture	Causal involvement	Culture-associated neurons and SAE features can be localized, ablated, and steered to change culturally conditioned prediction or task performance.	Studies disagree on whether the strongest culture-related components lie in upper versus shallow-to-middle layers. Localization is sensitive to method, model, dataset, and behavioral endpoint.
Stereotypes & demographic bias	Causal involvement / partial computational role	Gender, race, warmth/competence, and other bias-related information can be localized to directions, sparse features, neurons, and attention pathways whose intervention changes outputs.	Controllable demographic features do not always explain downstream bias, mechanisms differ across demographics, and some bias-associated representations are reused for legitimate capabilities, complicating targeted removal.

References

1. Meng, K.; Bau, D.; Andonian, A.; Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems* **2022**, *35*, 17359–17372.
2. Burns, C.; Ye, H.; Klein, D.; Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827* **2022**.
3. Li, K.; Patel, O.; Viégas, F.; Pfister, H.; Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems* **2023**, *36*, 41451–41530.
4. Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; Nanda, N. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems* **2024**, *37*, 136037–136083.
5. Saphra, N.; Wiegrefe, S. Mechanistic? In Proceedings of the Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP; Belinkov, Y.; Kim, N.; Jumelet, J.; Mohebbi, H.; Mueller, A.; Chen, H., Eds., Miami, Florida, US, 2024; pp. 480–498. <https://doi.org/10.18653/v1/2024.blackboxnlp-1.30>.
6. Mueller, A.; Brinkmann, J.; Li, M.; Marks, S.; Pal, K.; Prakash, N.; Rager, C.; Sankaranarayanan, A.; Sen Sharma, A.; Sun, J.; et al. The Quest for the Right Mediator: Surveying Mechanistic Interpretability for NLP Through the Lens of Causal Mediation Analysis. *Computational Linguistics* **2026**, *52*, 331–378. <https://doi.org/10.1162/coli.a.572>.
7. Rai, D.; Zhou, Y.; Feng, S.; Saparov, A.; Yao, Z. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646* **2024**.
8. Zhang, H.; Zhang, Z.; Nie, E.; Wang, M.; Su, Z.; Wang, Y.; Wang, Q.; Yuan, S.; Duan, X.; Xue, Q.; et al. Locate, Steer, and Improve: A Practical Survey of Actionable Mechanistic Interpretability in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 10317–10362. <https://doi.org/10.18653/v1/2026.findings-acl.502>.
9. Cai, J.; Rahman, M.; Enouen, J.; Qian, L.; Liu, Y. A Survey on Actionable Interpretability in Large Language Models. In Proceedings of the Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26; Calvanese, D., Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2026, pp. 7787–7797. Survey Track, <https://doi.org/10.24963/ijcai.2026/865>.
10. Shu, D.; Wu, X.; Zhao, H.; Rai, D.; Yao, Z.; Liu, N.; Du, M. A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 1690–1712. <https://doi.org/10.18653/v1/2025.findings-emnlp.89>.
11. Feldhus, N.; Kopf, L. Interpreting Language Models Through Concept Descriptions: A Survey. In Proceedings of the Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP; Belinkov, Y.; Mueller, A.; Kim, N.; Mohebbi, H.; Chen, H.; Arad, D.; Sarti, G., Eds., Suzhou, China, 2025; pp. 149–162. <https://doi.org/10.18653/v1/2025.blackboxnlp-1.8>.
12. Zhou, Y.; Li, J.; Xiang, Y.; Yan, H.; Gui, L.; He, Y. The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 14365–14378. <https://doi.org/10.18653/v1/2024.emnlp-main.795>.
13. Resck, L.; Augenstein, I.; Korhonen, A. Explainability and Interpretability of Multilingual Large Language Models: A Survey. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 20454–20486. <https://doi.org/10.18653/v1/2025.emnlp-main.1033>.
14. Hu, Y.; Gu, J.; Wang, R.; Yao, Z.; Peng, H.; Wu, X.; Chen, J.; Zhang, M.; Pan, L. Towards a Mechanistic Understanding of Large Reasoning Models: A Survey of Training, Inference, and Failures. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 19449–19466. <https://doi.org/10.18653/v1/2026.acl-long.889>.
15. Bereska, L.; Gavves, E. Mechanistic Interpretability for AI Safety – A Review, 2024, [arXiv:cs.AI/2404.14082].
16. Naseem, U. Mechanistic Interpretability for Large Language Model Alignment: Progress, Challenges, and Future Directions, 2026, [arXiv:cs.CL/2602.11180].
17. Chan, L.; Garriga-Alonso, A.; Goldowsky-Dill, N.; Greenblatt, R.; Nitishinskaya, J.; Radhakrishnan, A.; Shlegeris, B.; Thomas, N. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In Proceedings of the AI Alignment Forum, 2022, Vol. 2, p. 19.

18. Zhang, F.; Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. In Proceedings of the International Conference on Learning Representations, 2024, Vol. 2024, pp. 1651–1678.
19. Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; Wei, F. Knowledge Neurons in Pretrained Transformers. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Muresan, S.; Nakov, P.; Villavicencio, A., Eds., Dublin, Ireland, 2022; pp. 8493–8502. <https://doi.org/10.18653/v1/2022.acl-long.581>.
20. Huben, R.; Cunningham, H.; Smith, L.; Ewart, A.; Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In Proceedings of the International Conference on Learning Representations, 2024, Vol. 2024, pp. 7827–7845.
21. Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; Steinhardt, J. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, 2022, [\[arXiv:cs.LG/2211.00593\]](https://arxiv.org/abs/2211.00593).
22. Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358* **2026**.
23. Ghandeharioun, A.; Caciularu, A.; Pearce, A.; Dixon, L.; Geva, M. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102* **2024**.
24. Pan, A.; Chen, L.; Steinhardt, J. Latentqa: Teaching llms to decode activations into natural language. In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 78726–78759.
25. Karvonen, A.; Chua, J.; Dumas, C.; Fraser-Taliente, K.; Kantamneni, S.; Minder, J.; Ong, E.; Sharma, A.S.; Wen, D.; Evans, O.; et al. Activation Oracles: Training and Evaluating LLMs as General-Purpose Activation Explainers, 2026, [\[arXiv:cs.CL/2512.15674\]](https://arxiv.org/abs/2512.15674).
26. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It's Lying. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 967–976. <https://doi.org/10.18653/v1/2023.findings-emnlp.68>.
27. Park, K.; Choe, Y.J.; Veitch, V. The Linear Representation Hypothesis and the Geometry of Large Language Models. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning; Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; Berkenkamp, F., Eds. PMLR, 21–27 Jul 2024, Vol. 235, *Proceedings of Machine Learning Research*, pp. 39643–39666.
28. Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; Shieber, S. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems* **2020**, 33, 12388–12401.
29. Geiger, A.; Wu, Z.; Potts, C.; Icard, T.; Goodman, N. Finding alignments between interpretable causal variables and distributed neural representations. In Proceedings of the Causal Learning and Reasoning. PMLR, 2024, pp. 160–187.
30. Yan, H.; Yao, Z. Interpreting and Steering for Safe and Correct Code Generation. *arXiv preprint arXiv:2608.30025* **2026**.
31. Friedman, D.; Lampinen, A.; Dixon, L.; Chen, D.; Ghandeharioun, A. Interpretability illusions in the generalization of simplified models. *arXiv preprint arXiv:2312.03656* **2023**.
32. Lin, Z.; Basu, S.; Beigi, M.; Manjunatha, V.; Rossi, R.A.; Wang, Z.; Zhou, Y.; Balasubramanian, S.; Zarei, A.; Rezaei, K.; et al. A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models, 2025, [\[arXiv:cs.LG/2502.17516\]](https://arxiv.org/abs/2502.17516).
33. Burns, C.; Ye, H.; Klein, D.; Steinhardt, J. Discovering Latent Knowledge in Language Models Without Supervision, 2024, [\[arXiv:cs.CL/2212.03827\]](https://arxiv.org/abs/2212.03827).
34. Zhang, S.; Yu, T.; Feng, Y. TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 8908–8949. <https://doi.org/10.18653/v1/2024.acl-long.483>.
35. Han, J.; Band, N.; Razzak, M.; Kossen, J.; Rudner, T.G.J.; Gal, Y. Simple Factuality Probes Detect Hallucinations in Long-Form Natural Language Generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 16209–16226. <https://doi.org/10.18653/v1/2025.findings-emnlp.880>.
36. Bao, Y.; Zhang, X.; Du, T.; Zhao, X.; Feng, Z.; Peng, H.; Yin, J. Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.;

- Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 682–700. <https://doi.org/10.18653/v1/2025.findings-acl.38>.
37. Wu, Y.; Wang, G.; Chen, Y.; Zhang, J.; Zhang, Y.; Chen, Y.; Shang, J.; Zhang, G.; Liu, Z. PRISM: Probing Reasoning, Instruction, and Source Memory in LLM Hallucinations. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 33619–33656. <https://doi.org/10.18653/v1/2026.acl-long.1551>.
 38. Poulis, A.; Crovella, M.; Terzi, E. Testing the Limits of Truth Directions in LLMs, 2026, [\[arXiv:cs.CL/2604.03754\]](https://arxiv.org/abs/2604.03754).
 39. Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. Language Models (Mostly) Know What They Know, 2022, [\[arXiv:cs.CL/2207.05221\]](https://arxiv.org/abs/2207.05221).
 40. Li, M.; Zhao, Y.; Zhang, W.; Li, S.; Xie, W.; Ng, S.K.; Chua, T.S.; Deng, Y. Knowledge Boundary of Large Language Models: A Survey. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 5131–5157. <https://doi.org/10.18653/v1/2025.acl-long.256>.
 41. Teplica, C.; Liu, Y.; Cohan, A.; Rudner, T.G.J. SCIURus: Shared Circuits for Interpretable Uncertainty Representations in Language Models. In Proceedings of the 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); Chiruzzo, L.; Ritter, A.; Wang, L., Eds., Albuquerque, New Mexico, 2025; pp. 12451–12469. <https://doi.org/10.18653/v1/2025.naacl-long.618>.
 42. Huang, P.; Hsu, C.H.; Huang, T.H.; Lin, S.D. Role-Sensitive Neurons: A Neuron-Level Gain Control Mechanism for Confidence Steering. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 5924–5944. <https://doi.org/10.18653/v1/2026.findings-acl.294>.
 43. Su, W.; Wang, C.; Ai, Q.; Hu, Y.; Wu, Z.; Zhou, Y.; Liu, Y. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 14379–14391. <https://doi.org/10.18653/v1/2024.findings-acl.854>.
 44. Zhang, Z.; Hu, X.; Zhang, H.; Zhang, J.; Wan, X. ICR Probe: Tracking Hidden State Dynamics for Reliable Hallucination Detection in LLMs. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 17986–18002. <https://doi.org/10.18653/v1/2025.acl-long.880>.
 45. Yu, L.; Cao, M.; Cheung, J.C.; Dong, Y. Mechanistic Understanding and Mitigation of Language Model Non-Factual Hallucinations. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 7943–7956. <https://doi.org/10.18653/v1/2024.findings-emnlp.466>.
 46. Sun, Z.; Zang, X.; Zheng, K.; Song, Y.; Xu, J.; Zhang, X.; Yu, W.; Song, Y.; Li, H. ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability, 2025, [\[arXiv:cs.CL/2410.11414\]](https://arxiv.org/abs/2410.11414).
 47. Cheng, Z.; Cao, M.; Rondeau, M.A.; Cheung, J.C. Stochastic Chameleons: Irrelevant Context Hallucinations Reveal Class-Based (Mis)Generalization in LLMs. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 30187–30214. <https://doi.org/10.18653/v1/2025.acl-long.1458>.
 48. Cheang, C.S.; Chan, H.P.; Zhang, W.; Deng, Y. Do LLMs Really Know What They Don't Know? Internal States Mainly Reflect Knowledge Recall Rather Than Truthfulness. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 713–730. <https://doi.org/10.18653/v1/2026.findings-acl.34>.
 49. Xiao, C.; Chan, H.P.; Zhang, H.; Aljunied, M.; Bing, L.; Al Moubayed, N.; Rong, Y. Analyzing LLMs' Knowledge Boundary Cognition Across Languages Through the Lens of Internal Representations. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 24099–24115. <https://doi.org/10.18653/v1/2025.acl-long.1174>.

50. Tao, X.; Wang, Y.; Cai, Y.; Yang, Z.; Tang, J. Are LLMs Really Not Knowledgeable? Mining the Submerged Knowledge in LLMs' Memory. In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 134151–134164.

51. Wang, Y.; Chen, Y.; Wen, W.; Sheng, Y.; Li, L.; Zeng, D.D. Unveiling Factual Recall Behaviors of Large Language Models through Knowledge Neurons. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 7388–7402. <https://doi.org/10.18653/v1/2024.emnlp-main.420>.

52. Finlayson, M.; Mueller, A.; Gehrmann, S.; Shieber, S.; Linzen, T.; Belinkov, Y. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In Proceedings of the 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); Zong, C.; Xia, F.; Li, W.; Navigli, R., Eds., Online, 2021; pp. 1828–1843. <https://doi.org/10.18653/v1/2021.acl-long.144>.

53. Hanna, M.; Mueller, A. Incremental Sentence Processing Mechanisms in Autoregressive Transformer Language Models. In Proceedings of the 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); Chiruzzo, L.; Ritter, A.; Wang, L., Eds., Albuquerque, New Mexico, 2025; pp. 3181–3203. <https://doi.org/10.18653/v1/2025.naacl-long.164>.

54. Kryvosheieva, D.; de Varda, A.G.; Fedorenko, E.; Tuckute, G. Different types of syntactic agreement recruit the same units within large language models. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 209–227. <https://doi.org/10.18653/v1/2026.acl-long.7>.

55. Merullo, J.; Eickhoff, C.; Pavlick, E. Circuit component reuse across tasks in transformer language models, 2024. URL <https://arxiv.org/abs/2310.08744> 2024.

56. Zhang, R.; Yu, Q.; Zang, M.; Eickhoff, C.; Pavlick, E. The same but different: Structural similarities and differences in multilingual language modeling. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 95372–95392.

57. Duan, X.; Yao, Z.; Zhang, Y.; Wang, S.; Cai, Z.G. How Syntax Specialization Emerges in Language Models, 2025, [[arXiv:cs.CL/2505.19548](https://arxiv.org/abs/2505.19548)].

58. Geva, M.; Bastings, J.; Filippova, K.; Globerson, A. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 12216–12235. <https://doi.org/10.18653/v1/2023.emnlp-main.751>.

59. Lv, A.; Chen, Y.; Zhang, K.; Wang, Y.; Liu, L.; Wen, J.R.; Xie, J.; Yan, R. Interpreting Key Mechanisms of Factual Recall in Transformer-Based Language Models, 2024, [[arXiv:cs.CL/2403.19521](https://arxiv.org/abs/2403.19521)].

60. Yu, Q.; Merullo, J.; Pavlick, E. Characterizing Mechanisms for Factual Recall in Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 9924–9959. <https://doi.org/10.18653/v1/2023.emnlp-main.615>.

61. Dai, Q.; Heinzerling, B.; Inui, K. Cell-Based Representation of Relational Binding in Language Models. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 47464–47524. <https://doi.org/10.18653/v1/2026.acl-long.2194>.

62. Gur-Arieh, Y.; Geva, M.; Geiger, A. Mixing Mechanisms: How Language Models Retrieve Bound Entities In-Context, 2026, [[arXiv:cs.CL/2510.06182](https://arxiv.org/abs/2510.06182)].

63. Prakash, N.; Shaham, T.; Haklay, T.; Belinkov, Y.; Bau, D. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In Proceedings of the International Conference on Learning Representations, 2024, Vol. 2024, pp. 8057–8082.

64. McDougall, C.S.; Conmy, A.; Rushing, C.; McGrath, T.; Nanda, N. Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads. In Proceedings of the Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP; Belinkov, Y.; Kim, N.; Jumelet, J.; Mohebbi, H.; Mueller, A.; Chen, H., Eds., Miami, Florida, US, 2024; pp. 337–363. <https://doi.org/10.18653/v1/2024.blackboxnlp-1.22>.

65. Brinkmann, J.; Wendler, C.; Bartelt, C.; Mueller, A. Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages. In Proceedings of the 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); Chiruzzo, L.; Ritter, A.; Wang, L., Eds., Albuquerque, New Mexico, 2025; pp. 6131–6150. <https://doi.org/10.18653/v1/2025.naacl-long.312>.
66. Dumas, C.; Wendler, C.; Veselovsky, V.; Monea, G.; West, R. Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations in Transformers. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 31822–31841. <https://doi.org/10.18653/v1/2025.acl-long.1536>.
 67. Tang, T.; Luo, W.; Huang, H.; Zhang, D.; Wang, X.; Zhao, W.X.; Wei, F.; Wen, J.R. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 5701–5715. <https://doi.org/10.18653/v1/2024.acl-long.309>.
 68. Mondal, S.K.; Sen, S.; Singhania, A.; Jyothi, P. Language-Specific Neurons Do Not Facilitate Cross-Lingual Transfer. In Proceedings of the The Sixth Workshop on Insights from Negative Results in NLP; Drozd, A.; Sedoc, J.; Tafreshi, S.; Akula, A.; Shu, R., Eds., Albuquerque, New Mexico, 2025; pp. 46–62. <https://doi.org/10.18653/v1/2025.insights-1.6>.
 69. Zeng, H.; Han, S.; Chen, L.; Yu, K. Converging to a Lingua Franca: Evolution of Linguistic Regions and Semantics Alignment in Multilingual Large Language Models. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics; Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B.D.; Schockaert, S., Eds., Abu Dhabi, UAE, 2025; pp. 10602–10617.
 70. Nie, E.; Schmid, H.; Schuetze, H. Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 690–706. <https://doi.org/10.18653/v1/2025.findings-emnlp.37>.
 71. Kojima, T.; Okimura, I.; Iwasawa, Y.; Yanaka, H.; Matsuo, Y. On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); Duh, K.; Gomez, H.; Bethard, S., Eds., Mexico City, Mexico, 2024; pp. 6919–6971. <https://doi.org/10.18653/v1/2024.naacl-long.384>.
 72. Gurgurov, D.; Trinley, K.; Ghussin, Y.A.; Baumele, T.; van Genabith, J.; Ostermann, S. Language Arithmetics: Towards Systematic Language Neuron Identification and Manipulation, 2025, [arXiv:cs.CL/2507.22608].
 73. Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895* 2022.
 74. Singh, A.K.; Moskovitz, T.; Hill, F.; Chan, S.C.Y.; Saxe, A.M. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024, ICML/24.
 75. Todd, E.; Li, M.; Sen Sharma, A.; Mueller, A.; Wallace, B.; Bau, D. Function vectors in large language models. In Proceedings of the International conference on learning representations, 2024, Vol. 2024, pp. 17282–17333.
 76. Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; Sun, X. Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 9840–9855. <https://doi.org/10.18653/v1/2023.emnlp-main.609>.
 77. Kharlapenko, D.; Shabalin, S.; Barez, F.; Conmy, A.; Nanda, N. Scaling sparse feature circuit finding for in-context learning. *arXiv preprint arXiv:2504.13756* 2025.
 78. Yin, K.; Steinhardt, J. Which Attention Heads Matter for In-Context Learning? In Proceedings of the Proceedings of the 42nd International Conference on Machine Learning; Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; Zhu, J., Eds. PMLR, 13–19 Jul 2025, Vol. 267, *Proceedings of Machine Learning Research*, pp. 72428–72461.
 79. Yang, H.; Cho, H.; Inoue, N. Localizing task recognition and task learning in in-context learning via attention head analysis. In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 93780–93839.
 80. Yang, H.; Cho, H.; Zhong, Y.; Inoue, N. Unifying attention heads and task vectors via hidden state geometry in in-context learning. *Advances in Neural Information Processing Systems* 2026, 38, 21701–21759.

81. Stolfo, A.; Belinkov, Y.; Sachan, M. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 7035–7052. <https://doi.org/10.18653/v1/2023.emnlp-main.435>.

82. Hanna, M.; Liu, O.; Variengien, A. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model, 2023, [arXiv:cs.CL/2305.00586].

83. Yu, Z.; Ananiadou, S. Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 3293–3306. <https://doi.org/10.18653/v1/2024.emnlp-main.193>.

84. Rai, D.; Yao, Z. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of llms. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 7174–7193.

85. Nikankin, Y.; Reusch, A.; Mueller, A.; Belinkov, Y. Arithmetic without algorithms: Language models solve math with a bag of heuristics, 2024. URL <https://arxiv.org/abs/2410.21272>.

86. Hasani, H.; Izadi, A.; Askari, F.; Bagherian, M.; Mohammadian, S.; Izadi, M.; Baghshah, M.S. Understanding Counting Mechanisms in Large Language and Vision-Language Models, 2026, [arXiv:cs.CV/2511.17699].

87. Hasani, H.; Banayeeanzade, M.; Nafisi, A.; Mohammadian, S.; Askari, F.; Bagherian, M.; Izadi, A.; Baghshah, M.S. Mechanistic Interpretability of Large-Scale Counting in LLMs through a System-2 Strategy. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 40885–40901. <https://doi.org/10.18653/v1/2026.findings-acl.2031>.

88. Chowdhury, R.; Rupprecht, T.A.; Cao, S.; Liu, J.; Camps, O.; Bau, D.; Zhao, P.; Wang, Y. Mechanistic Interpretability of Structure-Aware Numerical Reasoning in LLaMA 3.1 8B, 2026, [arXiv:cs.LG/2608.18419].

89. Li, Z.; Jiang, G.; Xie, H.; Song, L.; Lian, D.; Wei, Y. Understanding and Patching Compositional Reasoning in LLMs. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 9668–9688. <https://doi.org/10.18653/v1/2024.findings-acl.576>.

90. Biran, E.; Gottesman, D.; Yang, S.; Geva, M.; Globerson, A. Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 14113–14130. <https://doi.org/10.18653/v1/2024.emnlp-main.781>.

91. Yu, Z.; Belinkov, Y.; Ananiadou, S. Back Attention: Understanding and Enhancing Multi-Hop Reasoning in Large Language Models. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 11257–11272. <https://doi.org/10.18653/v1/2025.emnlp-main.567>.

92. Chen, X.; Plaat, A.; van Stein, N. How does chain of thought think? mechanistic interpretability of chain-of-thought reasoning with sparse autoencoding. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 30297–30305.

93. Dura, M.; Öztürk, S.; Tekir, S. Mechanistic Interpretability of Chain-of-Thought Reasoning via Sequential Activation Patching, 2026, [arXiv:cs.CL/2608.22332].

94. Chang, S.; Bai, T.; Zhang, X.; Ma, Q.; Liu, Q.; Liao, Z.; Miao, Y.; Niu, L. Unlocking the Black Box of Latent Reasoning: An Interpretability-Guided Approach to Intervention. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 34019–34032. <https://doi.org/10.18653/v1/2026.acl-long.1568>.

95. Men, T.; Cao, P.; Jin, Z.; Chen, Y.; Liu, K.; Zhao, J. Unlocking the Future: Exploring Look-Ahead Planning Mechanistic Interpretability in Large Language Models. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 7713–7724. <https://doi.org/10.18653/v1/2024.emnlp-main.440>.

96. Wu, W.; Morris, J.X.; Levine, L. Do language models plan ahead for future tokens?, 2024, [arXiv:cs.LG/2404.00859].

97. Wu, W.; Wang, Y.; Xiao, G.; Peng, H.; Fu, Y. Retrieval Head Mechanistically Explains Long-Context Factuality, 2024, [arXiv:cs.CL/2404.15574].

98. Ardit, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; Nanda, N. Refusal in language models is mediated by a single direction. In Proceedings of the Proceedings of the 38th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2024; NIPS '24.
99. Siu, V.; Crispino, N.; Yu, Z.; Pan, S.; Wang, Z.; Liu, Y.; Song, D.; Wang, C. COSMIC: Generalized Refusal Direction Identification in LLM Activations. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 25534–25553. <https://doi.org/10.18653/v1/2025.findings-acl.1310>.
100. Wollschläger, T.; Elstner, J.; Geisler, S.; Cohen-Addad, V.; Günnemann, S.; Gasteiger, J. The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. In Proceedings of the Proceedings of the 42nd International Conference on Machine Learning; Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; Zhu, J., Eds. PMLR, 13–19 Jul 2025, Vol. 267, *Proceedings of Machine Learning Research*, pp. 66945–66970.
101. Piras, G.; Mura, R.; Brau, F.; Oneto, L.; Roli, F.; Biggio, B. SOM directions are better than one: multi-directional refusal suppression in language models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 32728–32736.
102. Maskey, U.; Dras, M.; Naseem, U. Over-Refusal and Representation Subspaces: A Mechanistic Analysis of Task-Conditioned Refusal in Aligned LLMs. *arXiv preprint arXiv:2603.27518* 2026.
103. Maskey, U.; Yadav, S.; Dras, M.; Naseem, U. SafeConstellations: Mitigating Over-Refusals in LLMs Through Task-Aware Representation Steering. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 44424–44440. <https://doi.org/10.18653/v1/2026.acl-long.2056>.
104. Zhou, Z.; Yu, H.; Zhang, X.; Xu, R.; Huang, F.; Li, Y. How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 2461–2488. <https://doi.org/10.18653/v1/2024.findings-emnlp.139>.
105. Chen, J.; Wang, X.; Yao, Z.; Bai, Y.; Hou, L.; Li, J. Towards understanding safety alignment: A mechanistic perspective from safety neurons. *Advances in Neural Information Processing Systems* 2026, 38, 12665–12700.
106. Han, P.; Qian, C.; Chen, X.; Zhang, Y.; Ji, H.; Zhang, D. SafeSwitch: Steering Unsafe LLM Behavior via Internal Activation Signals. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 6936–6955. <https://doi.org/10.18653/v1/2025.findings-emnlp.366>.
107. Zhang, H.; Wang, D.; Liu, Y.; Chen, K.; Wang, W. LLM-VA: Resolving the Jailbreak-Overrefusal Trade-off via Vector Alignment. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 5760–5776. <https://doi.org/10.18653/v1/2026.acl-long.260>.
108. Ball, S.; Kreuter, F.; Panickssery, N. Understanding Jailbreak Success: A Study of Latent Space Dynamics in Large Language Models. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Demberg, V.; Inui, K.; Marquez, L., Eds., Rabat, Morocco, 2026; pp. 250–279. <https://doi.org/10.18653/v1/2026.eacl-long.12>.
109. Fu, S.; Su, D.; Huang, B.; Sun, F.; Wang, J.; Chen, W.; Shen, H.; Cheng, X. Jailbreak LLMs through Internal Stance Manipulation. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 15444–15459. <https://doi.org/10.18653/v1/2025.emnlp-main.780>.
110. Wu, J.; Chen, C.; Hou, C.; Yuan, X. SafeInt: Shielding Large Language Models from Jailbreak Attacks via Safety-Aware Representation Intervention. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 8473–8488. <https://doi.org/10.18653/v1/2025.findings-emnlp.450>.
111. Li, Y.; Fastowski, A.; Zaradoukas, E.; Prenkaj, B.; Kasneci, G. Analysing the Safety Pitfalls of Steering Vectors. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 11182–11204. <https://doi.org/10.18653/v1/2026.findings-acl.544>.
112. Wagle, A.; Uddin, I.I.; Zhang, C.; Wang, L. Mechanistic Interpretability of LLM Jailbreaks via Internal Attribution Graphs. *arXiv preprint arXiv:2607.07903* 2026.

113. Wang, X.; Jian, S.; Li, S.; Li, X.; Li, Z.; Ji, B.; Wang, B.; Yu, J. JPU: Bridging Jailbreak Defense and Unlearning via On-Policy Path Rectification. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 7661–7674. <https://doi.org/10.18653/v1/2026.acl-long.348>.
114. Hung, K.S.; Yang, X.; Liu, C.; Li, H.; Chen, K.; Fan, C.; Kwok, T.O.; Zhang, W.; Li, X.; Song, Y. Into the Gray Zone: Domain Contexts Can Blur LLM Safety Boundaries. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 24830–24867. <https://doi.org/10.18653/v1/2026.acl-long.1139>.
115. Góral, G.; Winkels, M.; Basart, S. Depth-wise activation steering for honest language models. *arXiv preprint arXiv:2512.07667* **2025**.
116. Boxo, G.; Socha, R.; Yoo, D.; Raval, S. Caught in the Act: a mechanistic approach to detecting deception, 2025, [[arXiv:cs.AI/2508.19505](https://arxiv.org/abs/2508.19505)].
117. Kumar, S. Pressure-Testing Deception Probes in LLMs: Scaling, Robustness, and the Geometry of Deceptive Representations. In Proceedings of the Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM); Mille, S.; Gehrmann, S.; Schmidová, P.; Dušek, O.; Fadaee, M.; Lo, K.; Santus, E.; Stanovsky, G., Eds., San Diego, California, USA, 2026; pp. 472–489. <https://doi.org/10.18653/v1/2026.gem-main.43>.
118. Mothukuri, V.; Parizi, R.M. Trajectory Signatures of Deception in Large Language Models. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 34264–34276. <https://doi.org/10.18653/v1/2026.acl-long.1582>.
119. Genadi, R.A.; Nwadike, M.S.; Mukhituly, N.; Hiraoka, T.; AlQuabeh, H.; Inui, K. Sycophancy Hides Linearly in the Attention Heads. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Demberg, V.; Inui, K.; Marquez, L., Eds., Rabat, Morocco, 2026; pp. 6896–6912. <https://doi.org/10.18653/v1/2026.eacl-long.324>.
120. Vennemeyer, D.; Duong, P.A.; Zhan, T.; Jiang, T. Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs, 2026, [[arXiv:cs.CL/2509.21305](https://arxiv.org/abs/2509.21305)].
121. Baez, A.; Karny, S.; Pataranutaporn, P. Dissociating the Internal Representations of Sycophancy in LLMs. *arXiv preprint arXiv:2607.07003* **2026**.
122. Stolfo, A.; Balachandran, V.; Yousefi, S.; Horvitz, E.; Nushi, B. Improving Instruction-Following in Language Models through Activation Steering, 2025, [[arXiv:cs.CL/2410.12877](https://arxiv.org/abs/2410.12877)].
123. Bigoulaeva, I.; Rohweder, J.; Dutta, S.; Gurevych, I. Patches of Nonlinearity: Instruction Vectors in Large Language Models. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 12209–12262. <https://doi.org/10.18653/v1/2026.acl-long.559>.
124. Chen, R.; Ardit, A.; Sleight, H.; Evans, O.; Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509* **2025**.
125. Lu, C.; Gallagher, J.; Michala, J.; Fish, K.; Lindsey, J. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387* **2026**.
126. Feng, X.; Zhao, L.; Zhong, W.; Gu, Y.; Kong, L.; Feng, X.; Qin, B.; et al. PERSONA: dynamic and compositional inference-time personality control via activation vector algebra. In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 151839–151879.
127. Bhandari, P.; Naseem, U.; Nasim, M. Do Personality Traits Interfere? Geometric Limitations of Steering in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 29284–29293. <https://doi.org/10.18653/v1/2026.findings-acl.1463>.
128. Wei, Z.; Li, M.; Wang, Z.; Deng, Y. Beyond Static Personas: Situational Personality Steering for Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 19185–19210. <https://doi.org/10.18653/v1/2026.findings-acl.958>.
129. Dan, Y.; Yu, L.; Lin, J.; Chen, Q.; Zhou, J.; Tian, J.; Bai, Q.; He, L. A mechanistic interpretability perspective on personality in large language models. *Information Processing & Management* **2026**, 63, 104962.
130. Radford, A.; Jozefowicz, R.; Sutskever, I. Learning to Generate Reviews and Discovering Sentiment, 2017, [[arXiv:cs.LG/1704.01444](https://arxiv.org/abs/1704.01444)].

131. Tigges, C.; Hollinsworth, O.J.; Geiger, A.; Nanda, N. Language Models Linearly Represent Sentiment. In Proceedings of the Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP; Belinkov, Y.; Kim, N.; Jumelet, J.; Mohebbi, H.; Mueller, A.; Chen, H., Eds., Miami, Florida, US, 2024; pp. 58–87. <https://doi.org/10.18653/v1/2024.blackboxnlp-1.5>.
132. Reichman, B.; Avsian, A.; Heck, L. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 49924–49947.
133. Choi, B.J.; Weber, M. Latent Structure of Affective Representations in Large Language Models, 2026, [\[arXiv:cs.LG/2604.07382\]](https://arxiv.org/abs/2604.07382).
134. Keeman, M. Whether, Not Which: Mechanistic Interpretability Reveals Dissociable Affect Reception and Emotion Categorization in LLMs, 2026, [\[arXiv:cs.CL/2603.22295\]](https://arxiv.org/abs/2603.22295).
135. Tak, A.N.; Banayeeanzade, A.; Bolourani, A.; Kian, M.; Jia, R.; Gratch, J. Mechanistic Interpretability of Emotion Inference in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 13090–13120. <https://doi.org/10.18653/v1/2025.findings-acl.679>.
136. Maheswaran, A.; Desarkar, M.S. A Unified View on Emotion Representation in Large Language Models. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Demberg, V.; Inui, K.; Marquez, L., Eds., Rabat, Morocco, 2026; pp. 3589–3610. <https://doi.org/10.18653/v1/2026.eacl-long.165>.
137. Li, M.; Shi, X.; Deng, Y. CoSToM: Causal-oriented Steering for Intrinsic Theory-of-Mind Alignment in Large Language Models. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 9302–9317. <https://doi.org/10.18653/v1/2026.acl-long.421>.
138. Cadile, J.P. Detecting and Steering LLMs' Empathy in Action, 2025, [\[arXiv:cs.CL/2511.16699\]](https://arxiv.org/abs/2511.16699).
139. Kim, J.; Evans, J.; Schein, A. Linear Representations of Political Perspective Emerge in Large Language Models, 2025, [\[arXiv:cs.CL/2503.02080\]](https://arxiv.org/abs/2503.02080).
140. Nadeem, A.; Dras, M.; Naseem, U. Steering Towards Fairness: Mitigating Political Stance Bias in LLMs. In Proceedings of the Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts; Hürriyetoglu, A.; Tanev, H.; Thapa, S.; Adhikari, S., Eds., Varna, Bulgaria, 2025; pp. 52–61.
141. Zhang, T. Probing Political Ideology in Large Language Models: How Latent Political Representations Generalize Across Tasks. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 23349–23360. <https://doi.org/10.18653/v1/2025.findings-emnlp.1267>.
142. Yu, C.; Yi, B.; Karimi-Malekabadi, F.; Abdurahman, S.; Ye, J.; Narayanan, S.; Zhao, Y.; Dehghani, M. Tracing Moral Foundations in Large Language Models, 2026, [\[arXiv:cs.CL/2601.05437\]](https://arxiv.org/abs/2601.05437).
143. Jin, H.; Li, M.; Wang, X.; Xu, Z.; Huang, M.; Jia, Y.; Lian, D. Internal Value Alignment in Large Language Models through Controlled Value Vector Activation. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 27347–27371. <https://doi.org/10.18653/v1/2025.acl-long.1326>.
144. Khanuja, S.; Liu, H.; Zhang, S.; Lambert, J.; Chen, M.; Mathews, R.; Wang, L. Steering LLMs for Culturally Localized Generation, 2026, [\[arXiv:cs.CL/2603.23301\]](https://arxiv.org/abs/2603.23301).
145. Wu, J.; Li, B.; Xue, X. SCoUT: A Framework for Structured Stereotype Analysis in Language Models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 26931–26938.
146. Kang, Y.; Wang, J.; Li, Y.; Wang, M.; Tu, W.; Wang, Q.; Li, H.; Wu, T.; Feng, X.; Zhong, F.; et al. Are the Values of LLMs Structurally Aligned with Humans? A Causal Perspective. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 23147–23161. <https://doi.org/10.18653/v1/2025.findings-acl.1188>.
147. Dang, T.D.A.; Masud, S. Cultural Value Alignment Via Latent Activation Steering in Large Language Models, 2026, [\[arXiv:cs.CL/2605.26365\]](https://arxiv.org/abs/2605.26365).
148. Namazifard, D.; Galke Poeh, L. Isolating Culture Neurons in Multilingual Large Language Models. In Proceedings of the Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics; Inui,

- K.; Sakti, S.; Wang, H.; Wong, D.F.; Bhattacharyya, P.; Banerjee, B.; Ekbal, A.; Chakraborty, T.; Singh, D.P., Eds., Mumbai, India, 2025; pp. 768–785. <https://doi.org/10.18653/v1/2025.findings-ijcnlp.45>.
149. Yamamoto, T.; Kumon, R.; Bollegala, D.; Yanaka, H. Neuron-Level Analysis of Cultural Understanding in Large Language Models, 2026, [\[arXiv:cs.CL/2510.08284\]](https://arxiv.org/abs/2510.08284).
150. Ahsan, H.; Sen Sharma, A.; Amir, S.; Bau, D.; Wallace, B.C. Elucidating Mechanisms of Demographic Bias in LLMs for Healthcare. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 14614–14631. <https://doi.org/10.18653/v1/2025.findings-emnlp.789>.
151. Ahsan, H.; Wallace, B. Can SAEs reveal and mitigate racial biases of LLMs in healthcare? In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 4154–4172.
152. Dunefsky, J.; Cohan, A. Observable propagation: uncovering feature vectors in transformers. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024, ICML'24.
153. Kumarappan, A.; Mujoo, A. Not Just RLHF: Why Alignment Alone Won't Fix Multi-Agent Sycophancy, 2026, [\[arXiv:cs.LG/2605.12991\]](https://arxiv.org/abs/2605.12991).
154. Yi, J.S.K.; Mueller, A.; Lee, D. Latent Agents: A Post-Training Procedure for Internalized Multi-Agent Debate. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 15569–15602. <https://doi.org/10.18653/v1/2026.acl-long.709>.
155. Huh, D.; Mohapatra, P. Steering LLMs for Multi-agent Decision-making using Representation Learning.
156. Du, Z.; Wang, R.; Bai, H.; Cao, Z.; Zhu, X.; Cheng, Y.; Zheng, B.; Chen, W.; Ying, H. Enabling Agents to Communicate Entirely in Latent Space. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 27106–27129. <https://doi.org/10.18653/v1/2026.acl-long.1248>.
157. Peng, Y.; Zhang, D.C.; Wang, X.; Aletras, N. StateBridge: Training-free Hidden-state Alignment for Latent Communication in LLM Multi-Agent Systems, 2026, [\[arXiv:cs.AI/2608.13317\]](https://arxiv.org/abs/2608.13317).
158. Wenzel, M. Latent Communication Between Language Model Agents: Channels, Alignment, and the Limits of Text, 2026, [\[arXiv:cs.CL/2607.14103\]](https://arxiv.org/abs/2607.14103).
159. Lee, J.H.; Lauscher, A.; Albrecht, S.V. Towards Ethical Multi-Agent Systems of Large Language Models: A Mechanistic Interpretability Perspective, 2025, [\[arXiv:cs.AI/2512.04691\]](https://arxiv.org/abs/2512.04691).
160. Rai, D.; Miller, S.; Moran, K.; Yao, Z. Failure by Interference: Language Models Make Balanced Parentheses Errors When Faulty Mechanisms Overshadow Sound Ones. *Advances in Neural Information Processing Systems* **2026**, *38*, 132610–132639.
161. Ni, S.; Bi, K.; Guo, J.; Yu, L.; Bi, B.; Cheng, X. Towards Fully Exploiting LLM Internal States to Enhance Knowledge Boundary Perception. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 24315–24329. <https://doi.org/10.18653/v1/2025.acl-long.1184>.
162. Yu, Z.; Ananiadou, S. Understanding and Mitigating Gender Bias in LLMs via Interpretable Neuron Editing, 2025, [\[arXiv:cs.CL/2501.14457\]](https://arxiv.org/abs/2501.14457).
163. Chen, K.; Liu, Y.; Zhang, H.; Li, Y.; Deng, X.; Wang, D. Decoding Hidden Deception in Reasoning LLMs: Activation Explainers for Deception Auditing, 2026, [\[arXiv:cs.CL/2606.17478\]](https://arxiv.org/abs/2606.17478).
164. Wang, K.; Li, J.; Yang, S.; Zhang, Z.; Wang, D. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 33566–33574.
165. Chou, C.T.; Liu, G.; Sun, J.; Blondin, C.; Zhu, K.; Sharma, V.; O'Brien, S. Causal Language Control in Multilingual Transformers via Sparse Feature Steering, 2025, [\[arXiv:cs.CL/2507.13410\]](https://arxiv.org/abs/2507.13410).
166. Dunefsky, J.; Chlenski, P.; Nanda, N. Transcoders Find Interpretable LLM Feature Circuits, 2024, [\[arXiv:cs.LG/2406.11944\]](https://arxiv.org/abs/2406.11944).
167. Ferrando, J.; Costa-jussà, M.R. On the Similarity of Circuits across Languages: a Case Study on the Subject-verb Agreement Task. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 10115–10125. <https://doi.org/10.18653/v1/2024.findings-emnlp.591>.
168. Duan, X.; Zhou, X.; Xiao, B.; Cai, Z. Unveiling Language Competence Neurons: A Psycholinguistic Approach to Model Interpretability. In Proceedings of the Proceedings of the 31st International Conference

on Computational Linguistics; Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B.D.; Schockaert, S., Eds., Abu Dhabi, UAE, 2025; pp. 10148–10157.

169. Feng, Z.; Chen, Z.; Ma, J.; Po, Y.T.; Chersoni, E.; Li, B. Good Arguments Against the People Pleasers: How Reasoning Mitigates (Yet Masks) LLM Sycophancy. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 24536–24570. <https://doi.org/10.18653/v1/2026.acl-long.1126>.

170. Bhandari, P.; Fay, N.; Selvaganapathy, S.; Datta, A.; Naseem, U.; Nasim, M. Activation-Space Personality Steering: Hybrid Layer Selection for Stable Trait Control in LLMs. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Demberg, V.; Inui, K.; Marquez, L., Eds., Rabat, Morocco, 2026; pp. 6388–6403. <https://doi.org/10.18653/v1/2026.eacl-long.300>.

171. Jain, S.; Kirk, R.; Lubana, E.S.; Dick, R.; Tanaka, H.; Rocktäschel, T.; Grefenstette, E.; Krueger, D. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In Proceedings of the International Conference on Learning Representations, 2024, Vol. 2024, pp. 9944–10015.

172. Laban, P.; Hayashi, H.; Zhou, Y.; Neville, J. LLMs get lost in multi-turn conversation. In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 54738–54778.

173. Aghzal, M.; Geva, M.; Stein, G.J.; Yao, Z. Toward Mechanistic Interpretability of LLM Agents: Explaining Trajectories Demands New Methods 2026.

174. Minder, J.; Dumas, C.; Juang, C.; Chughtai, B.; Nanda, N. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. *Advances in Neural Information Processing Systems* 2026, 38, 106423–106474.

175. Ward, J.; Lin, C.; Venhoff, C.; Nanda, N. Reasoning-Finetuning Repurposes Latent Representations in Base Models, 2025, [arXiv:cs.LG/2507.12638].

176. Venhoff, C.; Arcuschin, I.; Torr, P.; Conmy, A.; Nanda, N. Base Models Know How to Reason, Thinking Models Learn When. In Proceedings of the Proceedings of the 43rd International Conference on Machine Learning; Zhang, T.; Dudik, M.; Jaggi, M.; Agarwal, A.; Li, S.; Schuurmans, D.; Zhu, J.; Berkenkamp, F.; Dong, H.; Bietti, A., Eds. PMLR, 06–11 Jul 2026, Vol. 306, *Proceedings of Machine Learning Research*, pp. 123688–123732.

177. Sharkey, L.; Chughtai, B.; Batson, J.; Lindsey, J.; Wu, J.; Bushnaq, L.; Goldowsky-Dill, N.; Heimersheim, S.; Ortega, A.; Bloom, J.; et al. Open Problems in Mechanistic Interpretability, 2025, [arXiv:cs.LG/2501.16496].

178. Nanda, N.; Engels, J.; Conmy, A.; Rajamanoharan, S.; Chughtai, B.; McDougall, C.; Kramár, J.; Smith, L. A Pragmatic Vision for Interpretability. AI Alignment Forum, 2025.

179. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 2023, 120, e2305016120.

180. Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; Liu, H. Large Language Models for Data Annotation and Synthesis: A Survey. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 930–957. <https://doi.org/10.18653/v1/2024.emnlp-main.54>.

181. Yin, F.; Srinivasa, J.; Chang, K.W. Characterizing truthfulness in large language model generations with local intrinsic dimension. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024, ICML/24.

182. Wang, T.; Jiao, X.; Zhu, Y.; Chen, Z.; He, Y.; Chu, X.; Gao, J.; Wang, Y.; Ma, L. Adaptive Activation Steering: A Tuning-Free LLM Truthfulness Improvement Method for Diverse Hallucinations Categories. In Proceedings of the Proceedings of the ACM on Web Conference 2025. ACM, 2025, WWW '25, p. 2562–2578. <https://doi.org/10.1145/3696410.3714640>.

183. Fatahi Bayat, F.; Liu, X.; Jagadish, H.; Wang, L. Enhanced Language Model Truthfulness with Learnable Intervention and Uncertainty Expression. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 12388–12400. <https://doi.org/10.18653/v1/2024.findings-acl.737>.

184. Khanmohammadi, R.; Miah, E.; Mardikoraem, M.; Kaur, S.; Brugere, I.; Smiley, C.; Thind, K.S.; Ghassemi, M.M. Calibrating LLM Confidence by Probing Perturbed Representation Stability. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 10448–10514. <https://doi.org/10.18653/v1/2025.emnlp-main.530>.

185. Chung, J.; Song, M.; Jung, H.; Lee, N. Quantifying Aleatoric Uncertainty of In-Context Learning for Robust Measure of LLM Prediction Confidence. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 2090–2108. <https://doi.org/10.18653/v1/2026.acl-long.95>.

186. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **2024**, *630*, 625–630.

187. Liu, G.K.M.; Yona, G.; Caciularu, A.; Szpektor, I.; Rudner, T.G.J.; Cohan, A. MetaFaith: Faithful Natural Language Uncertainty Expression in LLMs. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 29612–29656. <https://doi.org/10.18653/v1/2025.emnlp-main.1505>.

188. Amayuelas, A.; Wong, K.; Pan, L.; Chen, W.; Wang, W.Y. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 6416–6432. <https://doi.org/10.18653/v1/2024.findings-acl.383>.

189. Chen, S.; Xiong, M.; Liu, J.; Wu, Z.; Xiao, T.; Gao, S.; He, J. In-context sharpness as alerts: an inner representation perspective for hallucination mitigation. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024, ICML'24.

190. Li, S.; Han, J.; Wang, Y.; Zhu, Y.; Song, Z.; He, L.; Alghythee, K.K.A.; Yu, P.S. Why LLMs Hallucinate on Structured Knowledge: A Mechanistic Analysis of Reasoning over Linearized Representations. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 19943–19956. <https://doi.org/10.18653/v1/2026.acl-long.914>.

191. Luo, W.; Peng, G.; Li, W.; Wei, S.; Song, F.; Wang, L.; Yang, N.; Zhang, X.; Jin, J.; Wei, F.; et al. Two Pathways to Truthfulness: On the Intrinsic Encoding of LLM Hallucinations. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 25580–25626. <https://doi.org/10.18653/v1/2026.acl-long.1173>.

192. Li, Y.; Cong, Y.; Francis, E.J. Mechanistic Interpretability of Animacy Effects on Structure Choice in GPT-2. In Proceedings of the Proceedings of the 30th Conference on Computational Natural Language Learning; Bonial, C.; Berzak, Y., Eds., San Diego, California, USA, 2026; pp. 626–640. <https://doi.org/10.18653/v1/2026.conll-main.38>.

193. Jing, Y.; Yao, Z.; Guo, H.; Ran, L.; Wang, X.; Hou, L.; Li, J. LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 28232–28251. <https://doi.org/10.18653/v1/2025.emnlp-main.1433>.

194. Gurnee, W.; Horsley, T.; Guo, Z.C.; Kheirkhah, T.R.; Sun, Q.; Hathaway, W.; Nanda, N.; Bertsimas, D. Universal Neurons in GPT2 Language Models, 2024, [\[arXiv:cs.LG/2401.12181\]](https://arxiv.org/abs/2401.12181).

195. Schulz, L.Y.; Mitropolsky, D.; Poggio, T. Unraveling Syntax: Language Modeling and the Substructure of Grammars, 2026, [\[arXiv:cs.CL/2510.02524\]](https://arxiv.org/abs/2510.02524).

196. Hu, M.Y.; Petty, J.; Shi, C.; Merrill, W.; Linzen, T. Between Circuits and Chomsky: Pre-pretraining on Formal Languages Imparts Linguistic Biases. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 9691–9709. <https://doi.org/10.18653/v1/2025.acl-long.478>.

197. Yu, Z.; Ananiadou, S. Neuron-Level Knowledge Attribution in Large Language Models. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 3267–3280. <https://doi.org/10.18653/v1/2024.emnlp-main.191>.

198. Yao, Y.; Zhang, N.; Xi, Z.; Wang, M.; Xu, Z.; Deng, S.; Chen, H. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems* **2024**, *37*, 118571–118602.

199. Zhang, X.; Liang, Y.; Meng, F.; Zhang, S.; Chen, Y.; Xu, J.; Zhou, J. Multilingual Knowledge Editing with Language-Agnostic Factual Neurons. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics; Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B.D.; Schockaert, S., Eds., Abu Dhabi, UAE, 2025; pp. 5775–5788.

200. Aljaafari, N.; Carvalho, D.; Freitas, A. Bridging Linguistic Structure and Mechanistic Interpretability for Conceptual Interpretation in Language Models. In Proceedings of the Proceedings of the 30th Conference on Computational Natural Language Learning; Bonial, C.; Berzak, Y., Eds., San Diego, California, USA, 2026; pp. 722–741. <https://doi.org/10.18653/v1/2026.conll-main.44>.
201. Zhong, C.; Cheng, F.; Liu, Q.; Murawaki, Y.; Chu, C.; Kurohashi, S. Language Lives in Sparse Dimensions: Toward Interpretable and Efficient Multilingual Control for Large Language Models. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Demberg, V.; Inui, K.; Marquez, L., Eds., Rabat, Morocco, 2026; pp. 7958–7970. <https://doi.org/10.18653/v1/2026.eacl-long.372>.
202. Andrylie, L.M.; Rahmanisa, I.; Ihsani, M.K.; Wicaksono, A.F.; Wibowo, H.A.; Aji, A.F. Sparse Autoencoders Can Capture Language-Specific Concepts Across Diverse Languages, 2026, [\[arXiv:cs.CL/2507.11230\]](https://arxiv.org/abs/2507.11230).
203. Deng, B.; Wan, Y.; Yang, B.; Zhang, Y.; Feng, F. Unveiling Language-Specific Features in Large Language Models via Sparse Autoencoders. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 4563–4608. <https://doi.org/10.18653/v1/2025.acl-long.229>.
204. Harrasse, A.; Draye, F.; Pandey, P.S.; Jin, Z.; Schölkopf, B. Tracing Multilingual Representations in LLMs with Cross-Layer Transcoders, 2026, [\[arXiv:cs.CL/2511.10840\]](https://arxiv.org/abs/2511.10840).
205. Zhang, Z.; Zhao, J.; Zhang, Q.; Gui, T.; Huang, X. Unveiling Linguistic Regions in Large Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 6228–6247. <https://doi.org/10.18653/v1/2024.acl-long.338>.
206. Yang, H.; Cho, H.; Ding, K.; Inoue, N. Task Vectors, Learned Not Extracted: Performance Gains and Mechanistic Insights. In Proceedings of the International Conference on Learning Representations, 2026, Vol. 2026, pp. 74846–74903.
207. Cho, H.; Kato, M.; Sakai, Y.; Inoue, N. Revisiting in-context learning inference circuit in large language models. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 13081–13117.
208. Yu, Z.; Ananiadou, S. 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 Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 3281–3292. <https://doi.org/10.18653/v1/2024.emnlp-main.192>.
209. Zheng, B.; Ma, M.; Lin, Z.; Yang, T. Label Words as Local Task Vectors in In-Context Learning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 6319–6332. <https://doi.org/10.18653/v1/2026.findings-acl.314>.
210. Halawi, D.; Denain, J.S.; Steinhardt, J. Overthinking the Truth: Understanding how Language Models Process False Demonstrations, 2024, [\[arXiv:cs.LG/2307.09476\]](https://arxiv.org/abs/2307.09476).
211. Feucht, S.; Todd, E.; Wallace, B.; Bau, D. The Dual-Route Model of Induction, 2025, [\[arXiv:cs.CL/2504.03022\]](https://arxiv.org/abs/2504.03022).
212. Gould, R.; Ong, E.; Ogden, G.; Conmy, A. Successor heads: Recurring, interpretable attention heads in the wild. In Proceedings of the International Conference on Learning Representations, 2024, Vol. 2024, pp. 9236–9261.
213. Yang, S.; Gribovskaya, E.; Kassner, N.; Geva, M.; Riedel, S. Do Large Language Models Latently Perform Multi-Hop Reasoning? In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 10210–10229. <https://doi.org/10.18653/v1/2024.acl-long.550>.
214. Hong, G.Z.; Dikkala, N.; Luo, E.; Rashtchian, C.; Wang, X.; Panigrahy, R. A Implies B: Circuit Analysis in LLMs for Propositional Logical Reasoning, 2025, [\[arXiv:cs.LG/2411.04105\]](https://arxiv.org/abs/2411.04105).
215. Zhang, L.; Hu, L.; Wang, D. Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025; Chiruzzo, L.; Ritter, A.; Wang, L., Eds., Albuquerque, New Mexico, 2025; pp. 1387–1404. <https://doi.org/10.18653/v1/2025.findings-naacl.76>.
216. Dutta, S.; Singh, J.; Chakrabarti, S.; Chakraborty, T. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning, 2024, [\[arXiv:cs.CL/2402.18312\]](https://arxiv.org/abs/2402.18312).
217. García-Ferrero, I.; Montero, D.; Orus, R. Refusal Steering: Fine-grained Control over LLM Refusal Behaviour for Sensitive Topics. In Proceedings of the Proceedings of the Fifteenth Language Resources and Evaluation

Conference; Piperidis, S.; Bel, N.; van den Heuvel, H.; Ide, N.; Krek, S.; Toral, A., Eds., Palma de Mallorca, Spain, 2026; pp. 128–151. <https://doi.org/10.63317/2zypcp2r538o>.

218. Xie, Y.; Zhang, Y.; Liu, T.; Ma, D.; Liu, T. Beyond Surface Alignment: Rebuilding LLMs Safety Mechanism via Probabilistically Ablating Refusal Direction. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 17620–17632. <https://doi.org/10.18653/v1/2025.findings-emnlp.956>.
219. Jiang, E.H.; Ou, W.; Liu, R.; Pang, S.; Wan, G.; Duan, R.; Dong, W.; Chang, K.W.; Wang, X.; Wu, Y.N.; et al. Mitigating Over-Refusal in Aligned Large Language Models via Inference-Time Activation Energy. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 37930–37950. <https://doi.org/10.18653/v1/2026.acl-long.1759>.
220. Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.K.; et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* 2023.
221. Chen, J.; Wang, X.; Yao, Z.; Bai, Y.; Hou, L.; Li, J. Towards Understanding Safety Alignment: A Mechanistic Perspective from Safety Neurons, 2025, [arXiv:cs.CL/2406.14144].
222. Wang, P.; Zhang, D.; Li, L.; Tan, C.; Wang, X.; Zhang, M.; Ren, K.; Jiang, B.; Qiu, X. InferAligner: Inference-Time Alignment for Harmlessness through Cross-Model Guidance. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 10460–10479. <https://doi.org/10.18653/v1/2024.emnlp-main.585>.
223. Liu, W.; Wang, X.; Wu, M.; Li, T.; Lv, C.; Ling, Z.; JianHao, Z.; Zhang, C.; Zheng, X.; Huang, X. Aligning Large Language Models with Human Preferences through Representation Engineering. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 10619–10638. <https://doi.org/10.18653/v1/2024.acl-long.572>.
224. Ghosh, S.; Bhattacharjee, A.; Ziser, Y.; Parisien, C. A Simple Yet Effective Method for Non-Refusing Context Relevant Fine-grained Safety Steering in LLMs. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 35128–35148. <https://doi.org/10.18653/v1/2025.emnlp-main.1781>.
225. Jiao, D.; Liu, Y.; Yuan, Y.; Tang, Z.; Du, L.; Wu, H.; Anderson, A. LLM Safety From Within: Detecting Harmful Content with Internal Representations. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 39711–39727. <https://doi.org/10.18653/v1/2026.acl-long.1844>.
226. Lin, Y.; He, P.; Xu, H.; Xing, Y.; Yamada, M.; Liu, H.; Tang, J. Towards Understanding Jailbreak Attacks in LLMs: A Representation Space Analysis. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., Miami, Florida, USA, 2024; pp. 7067–7085. <https://doi.org/10.18653/v1/2024.emnlp-main.401>.
227. Li, X.; Ning, Y.; Bao, Z.; Xu, M.; Chen, J.; Qian, T. CAVGAN: Unifying Jailbreak and Defense of LLMs via Generative Adversarial Attacks on their Internal Representations. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 6664–6678. <https://doi.org/10.18653/v1/2025.findings-acl.346>.
228. LeVi, A.; Himelstein, R.; Nemcovsky, Y.; Mendelson, A.; Baskin, C. Jailbreak Attack Initializations as Extractors of Compliance Directions. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 6672–6705. <https://doi.org/10.18653/v1/2025.findings-emnlp.354>.
229. Zhao, C.; Ke, Y.; Huang, K. Unraveling LLM Jailbreaks Through Safety Knowledge Neurons. In Proceedings of the Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Demberg, V.; Inui, K.; Marquez, L., Eds., Rabat, Morocco, 2026; pp. 1889–1906. <https://doi.org/10.18653/v1/2026.eacl-long.83>.
230. Xing, W.; Fang, M.; Wang, G.; Lin, C.; Han, M. Silencing the Guardrails: Inference-Time Jailbreaking via Dynamic Contextual Representation Ablation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 4321–4334. <https://doi.org/10.18653/v1/2026.findings-acl.211>.

231. Long, Z.; Wang, S.; Liu, S.; Lai, Y. How Jailbreak Defenses Work and Ensemble? A Mechanistic Investigation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 21263–21290. <https://doi.org/10.18653/v1/2025.findings-emnlp.1160>.
232. Long, X.; Fu, Y.; Li, R.; Sheng, M.; Yu, H.; Han, X.; Li, P. When Truthful Representations Flip Under Deceptive Instructions? In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 16315–16335. <https://doi.org/10.18653/v1/2025.emnlp-main.826>.
233. Zolfaghari, V. Linear Separability of Activation Representations after Supervised Fine-Tuning on Incorrect Responses: A Study of Synthetic Dishonesty in Large Language Models, 2026, [\[arXiv:cs.LG/2605.30381\]](https://arxiv.org/abs/2605.30381).
234. Henniger, S.; Marin-Llobet, A.; Karty, R.E.; Amin, N. A Case Study of How Intuition and Reasoning Interact in Language Models. In Proceedings of the Mechanistic Interpretability Workshop at ICML 2026, 2026.
235. Ma, X.; Zhang, X.; Weng, Z. Stable and Explainable Personality Trait Evaluation in Large Language Models with Internal Activations. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 16322–16340. <https://doi.org/10.18653/v1/2026.findings-acl.803>.
236. Karny, S.; Baez, A.; Pataranutaporn, P. Neural Transparency: Mechanistic Interpretability Interfaces for Anticipating Model Behaviors for Personalized AI. In Proceedings of the Proceedings of the 31st International Conference on Intelligent User Interfaces, New York, NY, USA, 2026; IUI '26, p. 868–884. <https://doi.org/10.1145/3742413.3789120>.
237. Courtis, D.; Hu, T. Mechanistic Personality Analysis of LLMs: Steering Personality via Latent Feature Interventions, 2026, [\[arXiv:cs.AI/2606.28770\]](https://arxiv.org/abs/2606.28770).
238. Zeng, W.; Emami, A.; Choi, J.D. What Models Express, Suppress, and Resist: Auditing Open-Weight LLMs with Persona Vectors, 2026, [\[arXiv:cs.CL/2607.13162\]](https://arxiv.org/abs/2607.13162).
239. Berg, C.; Lulla, R. Exploitation Without Deception: Dark Triad Feature Steering Reveals Separable Antisocial Circuits in Language Models, 2026, [\[arXiv:cs.CL/2605.09773\]](https://arxiv.org/abs/2605.09773).
240. Zhang, W.; Yuan, Y.; Han, C.; Tian, Y.; Sun, Z.; Du, L.; Kang, J.; Kang, H.; Liu, X.; Wu, H. Preference Heads in Large Language Models: A Mechanistic Framework for Interpretable Personalization. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 47742–47754. <https://doi.org/10.18653/v1/2026.acl-long.2205>.
241. Turner, A.M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J.J.; Mini, U.; MacDiarmid, M. Steering Language Models With Activation Engineering, 2024, [\[arXiv:cs.CL/2308.10248\]](https://arxiv.org/abs/2308.10248).
242. Lee, J.; Lee, W.; Kwon, O.W.; Kim, H. Do Large Language Models Have “Emotion Neurons”? Investigating the Existence and Role. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 15617–15639. <https://doi.org/10.18653/v1/2025.findings-acl.806>.
243. Wu, X.; Wang, H.; Yan, Z.; Tang, X.; Xu, P.; Siok, W.T.; Li, P.; Gao, J.H.; Lyu, B.; Qin, L. Decoding and controlling emotion in LLMs through human-aligned representational geometry with enhanced interpretability. *Computers in Human Behavior* **2026**, p. 109051.
244. Hiremath, B.S.; Polignano, M.; Levantesi, M.; Semeraro, G.; De Luca, E.W.; Poznanski, A. State-wise Linear Modulation (SLiM): A Novel Approach for Steering Large Language Models. *Neural Networks* **2026**, p. 108708.
245. Sofroniew, N.; Kauvar, I.; Saunders, W.; Chen, R.; Henighan, T.; Hydrie, S.; Citro, C.; Pearce, A.; Tarng, J.; Gurnee, W.; et al. Emotion Concepts and their Function in a Large Language Model, 2026, [\[arXiv:cs.AI/2604.07729\]](https://arxiv.org/abs/2604.07729).
246. Wu, Y.; Guo, W.; Liu, Z.; Ji, H.; Xu, Z.; Zhang, D. How large language models encode theory-of-mind: a study on sparse parameter patterns. *npj Artificial Intelligence* **2025**, *1*, 20.
247. Zhang, J.; Yang, S.; Wu, J.; Wong, D.F.; Wang, D. Understanding and Mitigating Political Stance Cross-topic Generalization in Large Language Models. In Proceedings of the Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 29775–29797. <https://doi.org/10.18653/v1/2026.acl-long.1374>.
248. Tlaie, A. Exploring and steering the moral compass of Large Language Models, 2024, [\[arXiv:cs.AI/2405.17345\]](https://arxiv.org/abs/2405.17345).

249. Hu, L.; Zeng, J.; Yang, L.; Lin, H. The Straight and Narrow: Do LLMs Possess an Internal Moral Path?, 2026, [\[arXiv:cs.CL/2601.10307\]](#).
250. Zhang, J.; Kang, Y.; Guo, J.; Zhan, J.; Wang, J. BACH-V: Bridging Abstract and Concrete Human-Values in Large Language Models, 2026, [\[arXiv:cs.CL/2601.14007\]](#).
251. Zou, C.; Jiao, D.; Hu, L. Deciphering Cultural Representations in Large Language Models via Sparse Autoencoders. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026; Liakata, M.; Moreira, V.P.; Zhang, J.; Jurgens, D., Eds., San Diego, California, United States, 2026; pp. 5656–5677. <https://doi.org/10.18653/v1/2026.findings-acl.278>.
252. Yang, N.; Kang, T.; Choi, S.J.; Lee, H.; Jung, K. Mitigating Biases for Instruction-following Language Models via Bias Neurons Elimination. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 9061–9073. <https://doi.org/10.18653/v1/2024.acl-long.490>.
253. Jalali Farahani, F.; Staab, S. Steering Stereotypes: Inference-Time Intervention on Social Group Descriptions Generated by Large Language Models. In Proceedings of the Companion Publication of the 2026 18th ACM Web Science Conference, New York, NY, USA, 2026; WebSci Companion '26, p. 44–47. <https://doi.org/10.1145/3795513.3807434>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.