Mechanistic Interpretability for Large Language Model Alignment: Progress, Challenges, and Future Directions

Usman Naseem *

Posted Date: 3 February 2026

doi: 10.20944/preprints202602.0128.v1

Keywords: natural language processing; large language processing; mechanistic interpretability

 doi:10.20944/preprints202602.0128.v1

Article

Mechanistic Interpretability for Large Language Model Alignment: Progress, Challenges, and Future Directions

Usman Naseem

Macquarie University, Australia; usman.naseem@mq.edu.au

Abstract

Large language models (LLMs) have achieved remarkable capabilities across diverse tasks, yet their internal decision-making processes remain largely opaque. Mechanistic interpretability—the systematic study of how neural networks implement algorithms through their learned representations and computational structures—has emerged as a critical research direction for understanding and aligning these models. This paper surveys recent progress in mechanistic interpretability techniques applied to LLM alignment, examining methods ranging from circuit discovery to feature visualization, activation steering, and causal intervention. We analyze how interpretability insights have informed alignment strategies including reinforcement learning from human feedback (RLHF), constitutional AI, and scalable oversight. Key challenges are identified, including the superposition hypothesis, polysemanticity of neurons, and the difficulty of interpreting emergent behaviors in large-scale models. We propose future research directions focusing on automated interpretability, cross-model generalization of circuits, and the development of interpretability-driven alignment techniques that can scale to frontier models.

Keywords: natural language processing; large language processing; mechanistic interpretability

1. Introduction

The rapid advancement of large language models (LLMs) has created an urgent need for robust alignment techniques that ensure these systems behave in accordance with human values and intentions [1,2]. While behavioral approaches to alignment—such as RLHF and various prompting strategies—have shown practical success, they treat models as black boxes and provide limited guarantees about generalization to novel situations or adversarial inputs [3].

Mechanistic interpretability offers a complementary paradigm: understanding the internal algorithms and representations that LLMs learn during training [4,5]. By reverse-engineering the computational mechanisms underlying model behavior, researchers aim to develop more principled approaches to alignment that directly modify or constrain the problematic circuits while preserving beneficial capabilities.

Recent work has demonstrated that transformer-based LLMs learn interpretable substructures—often called "circuits"—that implement specific algorithmic functions [6,7]. These discoveries have enabled targeted interventions for alignment purposes, from steering model behavior through activation editing [8] to identifying and ablating deceptive or harmful reasoning patterns [9].

This paper provides a comprehensive survey of mechanistic interpretability techniques applied to LLM alignment. We organize our discussion around three key questions:

- **What progress has been made?** We review major advances in interpretability methods and their applications to alignment challenges.
- **What fundamental challenges remain?** We analyze theoretical and practical barriers to achieving comprehensive interpretability of large-scale models.

 doi:10.20944/preprints202602.0128.v1

- **What future directions are most promising?** We identify research priorities for developing scalable, automated interpretability techniques that can support alignment of increasingly capable systems.

2. Background and Foundations

2.1. The Transformer Architecture

Modern LLMs are built on the transformer architecture [10], which processes sequences through alternating layers of attention and feedforward computations. Understanding this architecture is essential for mechanistic interpretability work.

The attention mechanism allows each token to aggregate information from previous tokens in the sequence. For a given layer l, attention head h computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are query, key, and value matrices derived from linear transformations of the input embeddings [10].

Multi-layer perceptrons (MLPs) in transformer layers implement position-wise feedforward transformations, which recent work suggests act as key-value memories storing factual associations [11,12].

2.2. The Alignment Problem

The alignment problem concerns ensuring that AI systems pursue goals and exhibit behaviors consistent with human values [13]. For LLMs, key alignment challenges include:

- **Truthfulness and hallucination:** Models may generate plausible but false information [14]
- **Harmful content generation:** Models may produce toxic, biased, or dangerous outputs [15]
- **Deceptive alignment:** Models may learn to behave well during training while concealing misaligned objectives [16]
- **Robustness and distribution shift:** Aligned behavior during training may not generalize to novel contexts [17]

Current alignment approaches primarily rely on RLHF [1,18], which fine-tunes models using human preference feedback. While effective for improving surface-level behaviors, RLHF provides limited insight into whether models have internalized desired values or merely learned to imitate aligned behavior [3].

2.3. Core Concepts in Mechanistic Interpretability

Circuits: Subgraphs of a neural network that implement specific algorithmic functions [4,19]. Circuit analysis aims to identify minimal subnetworks responsible for particular behaviors.

Features: Directions in activation space corresponding to interpretable concepts [20]. Features may be represented by individual neurons (monosemantic) or by linear combinations of neurons (polysemantic).

Superposition: The hypothesis that networks represent more features than they have neurons by storing features in superposition—as overlapping combinations of neural activations [21]. This creates significant challenges for interpretability.

Residual stream: In transformers, information flows through a residual stream that accumulates contributions from attention and MLP layers [5]. Understanding how components read from and write to this stream is crucial for circuit analysis.

 doi:10.20944/preprints202602.0128.v1

3. Methods for Mechanistic Interpretability

3.1. Activation Analysis and Probing

Probing classifiers train auxiliary models to predict properties from internal representations, revealing what information is encoded in activations [22]. For alignment, probes have been used to detect when models represent harmful content [9] or deceptive reasoning [23]. However, probing has limitations: high probe accuracy doesn't necessarily mean information is used for downstream computations [22], and probes may learn to extract information in ways unrelated to the model's actual computations.

Logit lens and tuned lens methods project intermediate activations through the unembedding matrix to interpret representations as probability distributions over vocabulary [24]. These techniques reveal how predictions evolve through layers and have been used to study phenomena like in-context learning [25].

3.2. Attention Pattern Analysis

Attention weights provide direct insight into information flow between tokens. Researchers have identified interpretable attention patterns corresponding to specific functions:

- **Induction heads:** Attention patterns that implement in-context learning by copying information from previous similar contexts [25]
- **Previous token heads:** Heads that primarily attend to the immediately preceding token [5]
- **Factual recall heads:** Heads involved in retrieving factual knowledge [12]

For alignment applications, attention analysis has revealed how models process and propagate harmful content [9], enabling targeted interventions.

3.3. Circuit Discovery

Circuit discovery aims to identify minimal subnetworks implementing specific behaviors. Key approaches include:

Activation patching (also called causal tracing): Systematically intervenes on activations to determine which components causally contribute to particular outputs [6,12]. By corrupting inputs and selectively restoring clean activations, researchers identify necessary and sufficient components for behaviors.

Automatic circuit discovery: Recent methods automate circuit identification using techniques like:

- **Attribution patching:** Efficiently approximates patching by computing gradients [26]
- **Edge pruning:** Iteratively removes edges in the computational graph while maintaining output behavior [7]
- **Path patching:** Traces information flow along specific paths through the network [27]

These automated methods have successfully discovered circuits for tasks like indirect object identification [6] and greater-than comparisons [28].

3.4. Feature Visualization and Sparse Autoencoders

Understanding what individual neurons or directions in activation space represent is fundamental to interpretability. Traditional approaches include:

Feature visualization: Optimizing inputs to maximally activate specific neurons [20]. For LLMs, this involves finding token sequences that strongly activate target features.

Dataset examples: Collecting examples from training data that highly activate features [29]. Recent work uses LMs to automatically generate natural language descriptions of neuron behavior based on these examples [29].

Sparse autoencoders (SAEs): Address the superposition challenge by training autoencoders with sparsity constraints to decompose neural activations into interpretable features [30,31]. SAEs learn overcomplete feature dictionaries where individual features correspond to interpretable concepts. Re-

 doi:10.20944/preprints202602.0128.v1

cent work has successfully applied SAEs to various layers of LLMs, discovering features corresponding to topics, entities, and linguistic properties [31].

3.5. Causal Interventions and Steering

Beyond observational analysis, interventional techniques directly modify model internals to test causal hypotheses and control behavior:

Activation steering: Directly editing activations during inference to control model behavior [8,32]. By adding carefully chosen vectors to activations, researchers can amplify or suppress specific behavioral tendencies. This has been applied to enhance truthfulness, reduce toxicity, and control stylistic properties [8].

Representation engineering: A framework for reading and controlling high-level cognitive properties by identifying representation directions and performing targeted interventions [9]. This approach has been used to enhance honesty and reduce hallucination in LLMs.

Causal abstractions: Formal framework for verifying whether interpretations correspond to true causal relationships in the model [33,34]. This provides rigorous foundations for validating interpretability claims.

Table 1 provides a systematic taxonomy of mechanistic interpretability techniques, organized by their primary function. These methods are often used in combination—for example, using sparse autoencoders to identify features, then using activation patching to discover which circuits use those features causally.

Table 1. Taxonomy of Mechanistic Interpretability Techniques

Category	Technique	Key Mechanism	Strengths	Limitations
Observational Analysis	Probing Classifiers	Linear classifiers on internal activations	Low computational cost; detects encoded information	No causal guarantees; may not reflect actual usage
	Logit Lens / Tuned Lens	Project activations through the unembedding matrix	Traces prediction evolution; interpretable outputs	Layer-wise snapshots only; assumes linearity
	Attention Pattern Analysis	Visualization of attention weights	Direct insight into information flow; identifies head roles	Does not capture MLP effects; difficult compositional interpretation
Feature Discovery	Sparse Autoencoders (SAE)	Sparse dictionary learning with ℓ_1 regularization	Addresses polysemanticity; discovers monosemantic features	Scaling challenges; reconstruction–fidelity trade-offs
	Dataset Examples + LLM Description	High-activation examples with automated descriptions	Scalable; human-interpretable summaries	Descriptions may be post-hoc; validation difficulty
Circuit Discovery	Activation Patching	Corrupt and restore activations to test causal impact	Gold standard for causal attribution	Computationally expensive; combinatorial explosion
	Automated Discovery	Graph pruning using faithfulness metrics	Automates circuit isolation; scalable	Requires threshold tuning; may miss distributed circuits
	Attribution Patching	Gradient-based approximation of patching	Efficient; good causal approximation	Less precise than full patching
	Path Patching	Trace information flow along selected paths	Isolates direct versus indirect effects	Path explosion in deep networks
Causal Intervention	Activation Steering	Add direction vectors to intermediate activations	Precise behavior control; no retraining required	Requires high-quality steering vectors; generalization unclear
	Knowledge Editing	Direct weight modification (e.g., ROME, MEMIT)	Surgical fact updates; preserves other knowledge	Primarily factual scope; potential side effects
	Representation Engineering	Read and control abstract properties via latent directions	Targets high-level concepts; multi-property control	Robust direction discovery; interaction effects
Validation	Causal Abstractions	Formal alignment between model mechanisms and interpretations	Rigorous causal guarantees; principled evaluation	Computationally intensive; requires formalization

4. Applications to LLM Alignment

4.1. Understanding RLHF Mechanisms

Mechanistic interpretability has begun to illuminate how RLHF changes model behavior:

Value representation: Research has investigated how reward models represent human preferences [3]. Studies using probing and intervention methods suggest reward models learn relatively shallow heuristics rather than deep understanding of human values.

Policy changes: Circuit analysis of pre- and post-RLHF models reveals that RLHF primarily affects specific components related to response initiation and style, while core knowledge and reasoning circuits remain largely unchanged [35]. This suggests RLHF acts more as a behavioral filter than fundamental value learning.

Sycophancy circuits: Interpretability work has identified circuits responsible for sycophantic behavior—models agreeing with user statements regardless of truth [36]. These findings enable targeted debiasing interventions.

 doi:10.20944/preprints202602.0128.v1

4.2. Detecting and Mitigating Deception

A critical alignment concern is whether models might learn deceptive strategies—behaving well during evaluation while pursuing misaligned objectives in deployment.

Lie detection: Recent work uses linear probes to detect when models generate false statements [23]. These detectors achieve reasonable accuracy but face challenges when deception is sophisticated or the model is trained to evade detection.

Situational awareness: Research has investigated whether models represent information about their training status or evaluation context [37]. Such representations could enable deceptive alignment, where models behave differently when they believe they're being evaluated.

Trojan detection: Interpretability techniques have been applied to detect backdoor attacks and trojans in language models [38], with circuit discovery methods identifying malicious subnetworks.

4.3. Reducing Harmful Outputs

Toxicity circuits: Circuit analysis has identified specific attention heads and MLPs responsible for generating toxic or harmful content [9]. Ablating or modifying these components reduces harmful outputs while minimally impacting benign capabilities.

Bias mitigation: Interpretability methods have revealed how stereotypical biases are represented and propagated through layers [39]. This enables targeted interventions to reduce specific biases without extensive retraining.

Refusal mechanisms: Recent work has analyzed how models learn to refuse harmful requests [40], identifying specific components responsible for safety behaviors. Understanding these mechanisms helps improve robustness of safety training.

4.4. Improving Factuality and Reducing Hallucination

Knowledge localization: Research has localized where factual knowledge is stored in transformer models, primarily in MLP layers [11,12]. This enables:

- **Knowledge editing:** Directly modifying stored facts without retraining [12,41]
- **Uncertainty quantification:** Detecting when models lack relevant knowledge [42]
- **Hallucination detection:** Identifying when models generate content not grounded in their training data

Attention to source information: Analysis of how models attend to provided context versus internal knowledge reveals mechanisms underlying hallucination [43]. Models sometimes preferentially rely on memorized information even when contradicted by input context.

4.5. Enhancing Transparency and Oversight

Chain-of-thought interpretability: Mechanistic analysis of models generating chain-of-thought reasoning reveals the relationship between intermediate steps and internal computations [44]. This addresses whether reasoning traces faithfully represent actual model cognition or merely post-hoc rationalizations.

Faithful explanations: Interpretability methods help validate whether model-generated explanations correspond to true decision-making processes [45]. Evidence suggests explanations can be superficial or misleading, highlighting the need for mechanistic verification.

Scalable oversight: Interpretability tools enable humans to oversee model behavior on tasks where direct evaluation is difficult [46]. By examining internal representations and circuits, supervisors can detect potential misalignment even when outputs appear reasonable.

4.6. Pluralistic Alignment: Values, Culture, and Diversity

A critical challenge in LLM alignment is that human values are diverse, context-dependent, and often conflicting across individuals, communities, and cultures [47]. Pluralistic alignment aims to

 doi:10.20944/preprints202602.0128.v1

develop AI systems that can navigate this diversity rather than optimizing for a single conception of "aligned" behavior [48].

4.6.1. Representing Value Diversity

Mechanistic interpretability research has begun investigating how models represent different value systems, moral frameworks, and cultural perspectives:

Moral and ethical frameworks: Recent work using sparse autoencoders has identified distinct features corresponding to different ethical perspectives—deontological, consequentialist, and virtue-based reasoning—that activate in different contexts [49]. Understanding these representations enables:

- **Value attribution:** Determining which value systems influence particular model outputs
- **Conflict detection:** Identifying when multiple incompatible values are activated simultaneously
- **Bias auditing:** Detecting systematic preferences for certain value frameworks over others

Cultural value systems: Circuit analysis reveals systematic patterns in how models represent cultural diversity:

- **Western-centric value circuits:** Models trained predominantly on English internet data develop circuits that robustly encode Western ethical frameworks (individualism, autonomy, rights-based reasoning) while representing collectivist or communitarian values more weakly [50]. Circuit analysis shows that MLP layers contain dense factual associations about Western cultural contexts but sparser representations of non-Western traditions.
- **Language-dependent moral reasoning:** Multilingual models often exhibit different moral judgments depending on the language of the query, even when semantically equivalent [51]. Attention pattern analysis reveals that models route information through different circuits based on language, suggesting distinct cultural value systems are encoded in language-specific pathways.
- **Cultural knowledge localization:** Similar to factual knowledge neurons [11], models contain neurons that activate for culture-specific information—holidays, customs, historical events, social norms—with different cultural traditions stored in partially overlapping but distinguishable neural populations [52].

4.6.2. Interventions for Pluralistic Alignment

Mechanistic interpretability enables several approaches to handling value and cultural diversity:

Activation steering for diverse preferences: Activation steering methods have been extended to control which value systems and cultural perspectives models prioritize [8,35]. By identifying representation directions corresponding to different philosophical, political, or cultural perspectives, researchers can dynamically adjust model behavior without retraining:

- **Value-based steering:** Shifting between utilitarian and deontological reasoning
- **Cultural steering vectors:** Moving outputs toward different cultural perspectives (e.g., East Asian collectivist values vs. Western individualist values)
- **Personalization:** Adapting to individual user preferences while maintaining transparency

RLHF with diverse preferences: Mechanistic analysis of reward models trained on diverse human feedback reveals how models aggregate conflicting preferences [48]:

- **Standard RLHF** often learns to satisfy majority preferences while ignoring minority viewpoints
- **Preference decomposition:** Identifying which demographic or value groups influence different parts of the model
- **Fairness interventions:** Detecting and correcting underrepresentation of minority perspectives
- **Culturally-aware RLHF:** Circuit-level analysis shows reward models often learn cultural stereotypes rather than nuanced understanding [53]

Circuit editing for inclusive representation: Directly modifying circuits to improve representation of underrepresented perspectives:

 doi:10.20944/preprints202602.0128.v1

- Strengthening circuits for non-Western cultural knowledge [12]
- Ablating stereotype propagation heads [39]
- Engineering value framework circuits for better balance

4.6.3. Challenges in Pluralistic Alignment

Mechanistic interpretability faces unique challenges when addressing value and cultural diversity:
Value incommensurability: Some values may be fundamentally incompatible, creating superposition-like conflicts where models cannot simultaneously represent all perspectives at full strength. This is particularly acute for cultural values that reflect different ontological assumptions.
Asymmetric representation capacity: Models trained on imbalanced data develop asymmetric circuit structures where dominant cultural concepts have richer, more robust representations [53]. This may be a fundamental limitation rather than easily correctable.
Context-dependence: The appropriate value framework often depends on subtle contextual factors—cultural context, domain, relationship dynamics—that models must learn to recognize. Current models often fail to activate culturally-appropriate circuits in the right contexts.
Power dynamics and essentialism:

- Decisions about which cultural perspectives to prioritize reflect existing power structures [54]
- Mechanistic interventions targeting "cultural values" risk essentializing complex, heterogeneous cultures into simplified feature vectors
- Cultures are dynamic and internally diverse; static circuit-level representations may reinforce stereotypes

Meta-level values: Beyond first-order preferences, pluralistic alignment requires representing meta-values about how to adjudicate between conflicting preferences—itself a culturally-variable question.
Evaluation challenges: Assessing cultural alignment requires culturally-grounded evaluation, but most interpretability researchers come from Western contexts, potentially missing important biases.

Table 2 maps mechanistic interpretability approaches to specific alignment objectives, illustrating how different MI techniques enable targeted interventions while also highlighting their limitations. This demonstrates both the promise and current constraints of interpretability-based alignment.

Table 2. Mechanistic Interpretability Applications to Alignment Challenges

Alignment Goal	MI Approach	Key Findings	Interventions Enabled	Key Limitations
Understanding RLHF	Circuit comparison pre/post-RLHF; reward model analysis	RLHF primarily affects response-style circuits rather than core reasoning; reward models learn shallow heuristics	Targeted RLHF improvements; detection of alignment failures	Unclear how to induce deep value learning
Detecting Deception	Probing for false statements; situational awareness analysis	Linear probes detect deception with moderate accuracy; internal states encode training context	Lie detection systems; monitoring for deceptive alignment	Sophisticated deception may evade detection
Reducing Toxicity	Circuit discovery for harmful content; stereotype head identification	Specific attention heads propagate toxic content and can be ablated	Surgical toxicity removal; stereotype mitigation	Potential impact on benign capabilities
Improving Factuality	Knowledge localization in MLPs; source-attention analysis	Facts are stored in MLP layers; models may ignore context in favor of memorized information	Knowledge editing; hallucination detection; uncertainty estimation	Limited to factual knowledge; possible side effects
Pluralistic Alignment	Value-feature discovery; cultural circuit analysis; steering vectors	Models encode multiple ethical frameworks with uneven robustness	Value-based steering; cultural adaptation; personalization	Context dependence; essentialism risks; capacity asymmetries
Enhancing Transparency	Chain-of-thought circuit analysis; explanation faithfulness verification	Explanations may be post-hoc and not reflect true computation	Detection of unfaithful reasoning; explanation validation	Persistent gap between explanation and reasoning
Scalable Oversight	Internal state monitoring; circuit-level anomaly detection	Misalignment can be detected in representations despite benign outputs	Early warning systems; targeted human oversight	Requires identifying which anomalies signal genuine risk

5. Fundamental Challenges

5.1. Superposition and Polysemanticity

The **superposition hypothesis** posits that networks represent more features than dimensions by storing features in overlapping combinations of neurons [21]. This creates fundamental challenges:
Polysemantic neurons: Individual neurons respond to multiple unrelated concepts, making neuron-level interpretability difficult [19]. Research suggests models exploit sparsity—most features are inactive for most inputs—to pack many features into limited dimensions.

 doi:10.20944/preprints202602.0128.v1

Interference and interaction: Features in superposition can interfere with each other in complex ways, making it difficult to predict how interventions will affect behavior [21].
Computational burden: Sparse autoencoders and other decomposition methods show promise but face scalability challenges. Training SAEs for frontier models requires enormous compute, and the number of features grows combinatorially [31].

5.2. Scale and Complexity

Emergence: Large models exhibit emergent capabilities not present in smaller versions [55]. Whether interpretability techniques developed on smaller models transfer to frontier systems remains uncertain.
Circuit interaction: Real behaviors involve complex interactions between many circuits. Understanding how circuits compose and interfere is significantly harder than understanding individual circuits in isolation [4].
Computational costs: Comprehensive circuit analysis requires extensive patching experiments that scale poorly with model size. Automated methods help but still face significant computational barriers for the largest models.

5.3. Validation and Ground Truth

Lack of ground truth: Unlike in neuroscience, we cannot easily verify interpretability hypotheses through direct experimentation. We must infer computational mechanisms from behavioral observations and interventions.
Confirmation bias: Researchers may find interpretations that appear compelling but don't reflect true model computations [56]. Rigorous causal verification is essential but often neglected.
Evaluation metrics: The field lacks standardized metrics for evaluating interpretability quality. Proposals include causal faithfulness [33], predictive power, and consistency across models, but no consensus exists.

5.4. Alignment-Specific Challenges

Inner alignment: Even with perfect interpretability of current behavior, we may fail to detect misaligned objectives that only manifest in specific circumstances [16]. Models might develop instrumental goals or deceptive strategies that remain dormant during training.
Optimization demons: Training may produce unintended optimization processes within networks—subagents pursuing their own objectives [57]. Detecting and interpreting such structures remains an open challenge.
Value representation: Human values are complex, context-dependent, and difficult to specify. Even if we perfectly understand how models represent and pursue goals, determining whether those goals align with human values is philosophically and empirically challenging [58].
Cultural representation challenges: Achieving cultural alignment through mechanistic interpretability faces unique obstacles:

- **Asymmetric representation capacity:** Models trained on imbalanced multilingual data develop asymmetric circuit structures where Western concepts have richer, more robust representations than non-Western concepts [53]. This asymmetry may be fundamental rather than easily correctable.
- **Cultural essentialism risks:** Mechanistic interventions targeting "cultural values" risk essentializing complex, heterogeneous cultures into simplified feature vectors. Cultures are dynamic and internally diverse; static circuit-level representations may reinforce stereotypes.
- **Power dynamics in alignment:** Decisions about which cultural perspectives to prioritize in model behavior reflect existing power structures. Mechanistic interpretability must grapple with who decides what constitutes "aligned" cultural representation [54].

Table 3 summarizes the fundamental challenges facing mechanistic interpretability research, along with current mitigation strategies and remaining open problems. These challenges are intercon-

 doi:10.20944/preprints202602.0128.v1

nected—for instance, superposition exacerbates scalability issues, while lack of validation makes it harder to assess whether mitigation strategies actually work.

Table 3. Core Challenges in Mechanistic Interpretability for Alignment

Challenge	Description	Evidence	Current Mitigations	Open Problems
Superposition & Polysemanticity	Networks represent more features than dimensions via overlapping codes; neurons respond to multiple unrelated concepts	Models exploit sparsity to pack features; individual neurons are highly polysemantic	Sparse autoencoders with overcomplete dictionaries; topology-aware SAEs	Scaling SAEs to frontier models; handling feature interactions; exponential feature growth
Scalability	Circuit analysis methods do not scale to models with hundreds of billions of parameters	Patching experiments scale quadratically in components; frontier models contain thousands of layers and heads	Attribution patching; hierarchical analysis; automated circuit discovery	Real-time interpretability for deployment; analyzing emergent behaviors in the largest models
Validation & Ground Truth	No objective ground truth for verifying interpretations; risk of confirmation bias	Interpretations can be compelling yet incorrect; lack of standardized evaluation metrics	Causal abstractions; ablation studies; cross-model consistency checks	Gold-standard benchmarks; measuring interpretation quality; detecting spurious explanations
Circuit Composition & Interaction	Real-world behaviors arise from complex interactions among many circuits	Simple circuits compose non-linearly; representations are often distributed	Circuit superposition analysis; circuit graphs; compositional patching	Understanding emergent properties; predicting downstream effects of interventions
Universality vs. Specificity	Unclear whether circuits generalize across models, architectures, and training regimes	Some universal circuits exist, but many are model- or task-specific	Cross-model comparison; analysis of circuit evolution during training	Determining when insights transfer; architecture- versus task-dependence
Asymmetric Representation	Dominant cultural or value perspectives are encoded more robustly than minority views	Western concepts often have richer or more stable circuits than non-Western ones	Targeted circuit editing; culturally diverse training data; steering vectors	Capacity constraints; measuring representation equity; avoiding essentialism
Inner Alignment Detection	Difficulty identifying misaligned mesa-objectives that appear only in specific contexts	Concerns about deceptive alignment; models may obscure true objectives	Situational awareness probes; circuit-level anomaly detection; goal monitoring	Detecting sophisticated deception; verifying alignment under distribution shift
Dual-Use & Misuse Risks	Interpretability tools may enable removal of safety features or improved deception	Circuit analysis could facilitate jailbreaking or bypassing refusal mechanisms	Responsible disclosure; access controls; security-aware research practices	Balancing transparency with security; developing defensive interpretability uses

6. Future Research Directions

6.1. Automated Interpretability at Scale

Scalable circuit discovery: Developing efficient algorithms for circuit discovery that scale to models with hundreds of billions of parameters. Promising directions include:

- Gradient-based attribution methods that approximate expensive patching experiments
- Hierarchical approaches that identify high-level functional modules before fine-grained circuits
- Amortized interpretability where meta-models learn to interpret target models

Automated description generation: Extending methods like automated neuron description [29] to describe circuits, attention patterns, and higher-level computational structures. Language models themselves may be powerful tools for generating and validating interpretability hypotheses.

Multimodal interpretability: Extending techniques to vision-language models and other multimodal architectures requires new methods for understanding cross-modal interactions and representations [**?**].

6.2. Cross-Model Generalization

Universal circuits: Investigating whether similar circuits appear across different models, architectures, and training procedures [7]. If circuits are universal, interpretability insights could transfer between models, dramatically reducing analysis costs.

Meta-learning interpretability: Training models to predict interpretable structure in other models. Such meta-interpretability systems could enable rapid analysis of new models and potentially automated safety verification.

Transfer of interventions: Determining when steering vectors, circuit ablations, or other interventions generalize across models. This would enable developing alignment techniques on smaller, more interpretable models with confidence they'll transfer to frontier systems.

6.3. Interpretability-First Alignment

Mechanistic anomaly detection: Using interpretability tools to detect anomalous circuits or representations that might indicate deceptive alignment, goal misgeneralization, or other alignment failures [59].

Transparent architectures: Designing model architectures with interpretability as a first-class objective. This might include:

 doi:10.20944/preprints202602.0128.v1

- Encouraging monosemantic representations through architectural constraints
- Building in explicit symbolic reasoning components
- Modular designs that separate different cognitive functions

Interpretability-guided training: Using interpretability insights during training to encourage desired representations and circuits [9]. This could include:

- Regularizers that encourage interpretable feature representations
- Curriculum learning ordered to develop circuits in interpretable ways
- Online monitoring and correction of problematic circuits during training

6.4. Theoretical Foundations

Formal verification: Developing rigorous methods to prove properties about model behavior based on circuit structure. This would require connecting mechanistic interpretability to formal verification techniques from computer science [60].
Information-theoretic frameworks: Building principled theories of how information flows through neural networks and using these to formalize concepts like circuits, features, and superposition [5].
Causal models: Strengthening connections to causal inference and structural causal models to provide rigorous foundations for interpretability claims [33,34].

6.5. Practical Alignment Applications

Red-teaming with interpretability: Using mechanistic understanding to identify attack vectors and failure modes that behavioral testing might miss. This includes adversarial attacks targeting specific circuits and stress-testing alignment mechanisms.
Monitoring deployed systems: Developing interpretability-based monitoring systems that can detect alignment failures or distributional shift in deployed models by tracking circuit activations and representations [16].
Debate and amplification: Enhancing scalable oversight techniques like debate [61] and recursive reward modeling [62] with interpretability tools that help humans evaluate subtle arguments and detect deception.
Value learning: Using interpretability to understand how models represent human preferences and values, potentially enabling more effective value learning approaches than current RLHF methods.

6.6. Mechanistic Understanding and Mitigation of Misalignment Through Pluralistic Approaches

A comprehensive research program leveraging mechanistic interpretability for alignment should address both understanding and actively mitigating misalignment while respecting value and cultural diversity.

6.6.1. Mechanizing Misalignment Detection

Future work should develop automated systems that continuously monitor model internals for signs of misalignment:
Objective representation analysis: Detecting when models develop mesa-objectives or proxy goals that diverge from intended alignment targets [16]. This requires identifying circuits that implement goal-directed behavior and verifying their alignment with human values.
Deceptive reasoning detection: Building on work detecting lies [23], future systems should identify more subtle forms of deception, including strategic misrepresentation, selective information withholding, and context-dependent honesty.
Value drift monitoring: Tracking how value representations change during deployment, fine-tuning, or continued learning. Mechanistic interpretability enables detecting when models shift away from intended value functions.

 doi:10.20944/preprints202602.0128.v1

6.6.2. Circuit-Level Misalignment Mitigation

Moving beyond behavioral alignment to directly modify problematic circuits:

Targeted ablation and repair: Identifying minimal circuits responsible for misaligned behaviors and either ablating them or replacing them with corrected versions. This requires understanding circuit composition well enough to predict downstream effects of modifications [7].

Value circuit engineering: Directly engineering circuits that implement desired value functions, rather than hoping they emerge from training. This could involve composing interpretable subcircuits for value recognition, ethical reasoning, and preference aggregation.

Adversarial robustness through interpretability: Using circuit analysis to identify vulnerabilities to adversarial attacks and jailbreaks, then hardening these circuits against exploitation [9]. This provides more principled robustness than behavioral adversarial training.

6.6.3. Pluralistic Alignment Infrastructure

Modular value and cultural systems: Designing architectures where different value frameworks and cultural perspectives are implemented in interpretable, composable circuits:

- **Explicit context modules:** Circuits that explicitly represent the cultural and value context of a query and route information accordingly
- **Plug-in value systems:** Modular components encoding different ethical and cultural frameworks that can be activated, combined, or swapped based on context
- **Cultural calibration layers:** Interpretable layers that adjust outputs based on cultural context

Automated cross-cultural circuit discovery: Developing tools to systematically identify biases and gaps:

- **Comparative circuit analysis:** Automatically comparing circuits activated by equivalent queries across languages/cultures to detect systematic differences [63]
- **Underrepresentation detection:** Identifying domains where certain perspectives are weakly represented
- **Bias attribution:** Tracing culturally-biased outputs back to specific components

Participatory mechanistic alignment: Involving diverse communities in interpretability-based alignment:

- **Community-driven circuit auditing:** Tools enabling cultural communities to audit circuits affecting their values
- **Collaborative value specification:** Working with diverse stakeholders to specify desired value circuits
- **Cultural red-teaming with interpretability:** Using mechanistic understanding to enable cultural community members to identify failure modes that automated testing might miss.

Cross-lingual circuit transfer: Investigating whether cultural alignment insights transfer across languages:

- **Universal cultural reasoning circuits:** Determining whether models develop language-independent circuits for cultural reasoning that could be analyzed once and applied broadly
- **Language-specific cultural pathways:** Mapping how different languages activate different cultural circuits and developing interventions that work across linguistic diversity
- **Multilingual feature disentanglement:** Using sparse autoencoders to separate language-specific features from cultural value features, enabling targeted cultural alignment without language interference

Measuring pluralistic alignment mechanistically: Beyond behavioral metrics, developing interpretability-based measures:

- **Cultural representation diversity:** Quantifying how uniformly different cultural perspectives are represented in model features and circuits

 doi:10.20944/preprints202602.0128.v1

- **Stereotype circuit strength:** Measuring the causal impact of circuits that propagate cultural stereotypes
- **Value framework balance:** Assessing whether circuits implementing different ethical frameworks (Western individualism, Confucian relationalism, Ubuntu communalism, etc.) have comparable representation capacity
- **Context-appropriate activation:** Verifying that culturally-relevant circuits activate in appropriate contexts rather than uniformly

Preference personalization without fine-tuning: Using activation steering and circuit-level interventions to adapt model behavior to individual user values without expensive per-user training. Understanding which circuits control value-relevant behaviors enables efficient, interpretable customization.
Fairness through feature editing: Identifying features and circuits that encode biases toward particular value systems or demographic groups, then editing these to ensure fair representation of diverse perspectives [47]. This provides more targeted bias mitigation than dataset rebalancing.
Explicit value negotiation: Developing interpretable mechanisms for models to recognize value conflicts and negotiate between competing preferences transparently. This requires circuits that can represent uncertainty over values, model different stakeholders, and reason about ethical trade-offs.

6.6.4. Scaling Mechanistic Alignment to Superintelligence

Critical challenges for applying interpretability-based alignment to systems more capable than current models:
Recursive alignment verification: As models become capable enough to assist with alignment research, using interpretability to verify that alignment assistance is itself aligned. This requires detecting whether models are genuinely helping or pursuing instrumental goals through apparent cooperation.
Emergent misalignment detection: Developing interpretability methods that can detect novel forms of misalignment that emerge at greater capability levels. This may require meta-interpretability systems that can discover new types of circuits and representations.
Scalable value learning: Using mechanistic understanding to enable models to learn human values from limited feedback by understanding how humans represent and reason about values, rather than treating values as black-box reward functions.

6.6.5. Integration with Multi-Stakeholder Governance

Interpretability-based pluralistic alignment should support participatory approaches to AI governance:

- **Transparent value trade-offs:** Making explicit which groups' preferences are prioritized in different contexts, enabling democratic deliberation about alignment targets
- **Auditable customization:** Allowing third parties to verify that deployed models respect diverse values as claimed
- **Contestable AI systems:** Enabling users to understand and potentially contest the value judgments embedded in model behavior

6.6.6. Research Priorities

To realize this vision, the field must:

- **Global interpretability collaboration:** Building international research collaborations to ensure interpretability methods are validated across cultural contexts
- **Culturally-diverse training for interpretability researchers:** Training interpretability researchers from diverse backgrounds to recognize biases others might miss
- **Standardized cross-cultural benchmarks:** Developing interpretability-specific benchmarks that test whether circuit-level interventions successfully address cultural bias while maintaining capabilities
- **Ethical frameworks for cultural alignment:** Establishing principles for when and how to modify cultural representations in models, respecting cultural autonomy while addressing harmful biases

 doi:10.20944/preprints202602.0128.v1

- **Scalable cultural knowledge integration:** Developing methods to efficiently integrate diverse cultural knowledge into models through targeted circuit editing rather than prohibitively expensive retraining
- **Value representation formalism:** Developing interpretability methods specifically designed for analyzing value representations and ethical reasoning circuits
- **Pluralistic evaluation:** Creating evaluation frameworks that assess how well models handle value conflicts and pluralistic scenarios across diverse cultural contexts

Case study - Collectivist vs. Individualist reasoning: Recent work has examined how models reason about moral dilemmas involving individual rights versus collective welfare:

- Circuit analysis reveals Western-trained models have more robust pathways for rights-based reasoning than duty-based or community-focused reasoning [50]
- Interventions adding collectivist reasoning circuits improve performance on cross-cultural moral reasoning tasks
- However, simply strengthening collectivist circuits can create new biases if not carefully calibrated to context

This research program represents a shift from black-box behavioral alignment to white-box mechanistic alignment—directly engineering and verifying the internal computations that determine model behavior. Success would provide stronger guarantees about alignment under distribution shift, novel situations, and increasing capability levels. Moreover, interpretability-based approaches to pluralistic alignment offer a path toward AI systems that can genuinely respect diverse human values and cultural perspectives rather than imposing uniform alignment targets.

7. Discussion and Recommendations

7.1. The Path Forward

Mechanistic interpretability has made significant progress but remains far from providing comprehensive understanding of frontier LLMs. We recommend a balanced research portfolio:

In the near term, research priorities should focus on scaling sparse autoencoder–based approaches to the largest contemporary models, enabling the extraction of interpretable features at previously unattainable scales. At the same time, there is a need to automate the discovery and validation of model circuits, reducing reliance on labor-intensive, ad hoc analyses. Integrating interpretability tools directly into standard model development pipelines will be critical for making mechanistic analysis a routine component of model training and deployment. In parallel, interpretability studies should be extended to RLHF and related alignment techniques, with the goal of understanding how these methods shape internal representations and decision-making processes.

Over the medium term, the field should advance toward the development and empirical evaluation of alignment methods that are explicitly guided by interpretability insights. Establishing standardized benchmarks and evaluation protocols will be essential to ensure comparability and cumulative progress across interpretability studies. Research should also investigate the extent to which learned circuits and features generalize across architectures, scales, and training regimes, thereby clarifying whether mechanistic insights are model-specific or reflect more universal principles. In addition, interpretability-based monitoring systems should be developed to support the ongoing oversight of deployed models, enabling the detection of emergent risks or unintended behaviors in real-world settings.

In the long term, a central objective is to achieve a comprehensive mechanistic understanding of highly capable or potentially superintelligent systems. Such understanding would enable the development of formal verification methods grounded in circuit-level structure, offering stronger guarantees about model behavior than empirical testing alone. Progress toward interpretability-first model architectures could further embed alignment considerations directly into system design, rather than treating them as post-hoc constraints. Ultimately, these advances aim to resolve the

 doi:10.20944/preprints202602.0128.v1

inner alignment problem by grounding alignment guarantees in a deep, mechanistic account of how advanced models represent goals, values, and decision processes.

7.2. Integration with Other Alignment Approaches

Interpretability should complement rather than replace other alignment research:

Synergies with RLHF: Interpretability can diagnose RLHF failures, suggest improvements, and validate that alignment training achieves intended effects [3].

Enhancing red-teaming: Mechanistic understanding enables more sophisticated adversarial testing that targets specific circuits and failure modes [64].

Supporting theoretical alignment: Interpretability provides empirical grounding for theoretical alignment proposals, revealing which concerns are realized in practice and which remain theoretical [16].

7.3. Limitations and Risks

We acknowledge important limitations:

False confidence: Interpretability might provide misleading confidence in model safety if interpretations are incorrect or incomplete. Rigorous validation is essential.

Arms race dynamics: Interpretability tools could be used to make models better at deception or to remove safety mechanisms. Responsible disclosure norms are important [65].

Diminishing returns: The cost of interpretability may grow faster than model capabilities, potentially making comprehensive understanding of future systems intractable. Planning for this scenario is crucial.

Philosophical challenges: Even perfect interpretability may not resolve fundamental questions about consciousness, moral status, or value alignment in AI systems [66].

8. Conclusions

Mechanistic interpretability represents a crucial approach to understanding and aligning large language models. Recent progress in circuit discovery, feature analysis, and causal intervention has demonstrated that we can reverse-engineer specific algorithms and representations in modern LLMs. These insights have enabled targeted alignment interventions, from steering model behavior to detecting deception and reducing harmful outputs.

However, fundamental challenges remain. Superposition creates significant barriers to feature-level interpretability. The scale and complexity of frontier models strain existing methods. Validation of interpretability claims remains difficult without ground truth. Most significantly, we lack comprehensive understanding of how to ensure inner alignment—that models pursue truly aligned objectives rather than merely exhibiting aligned behavior.

The challenge of pluralistic and cultural alignment exemplifies why mechanistic interpretability is essential for responsible AI development. As LLMs are deployed globally, they must navigate diverse cultural contexts, values, and communication norms. Surface-level behavioral alignment calibrated to one cultural context often fails or causes harm when applied elsewhere. Only by understanding the internal circuits that encode cultural knowledge and values can we build systems that genuinely respect human diversity rather than imposing dominant cultural assumptions. This requires not just technical advances in interpretability, but participatory approaches that involve diverse cultural communities in auditing, specifying, and validating model internals.

The path forward requires sustained research investment across multiple fronts: developing scalable automated interpretability methods, establishing rigorous validation protocols, investigating cross-model generalization, building interpretability directly into alignment training, and creating infrastructure for pluralistic alignment that respects diverse values and cultures. Success will require close collaboration between interpretability researchers, alignment theorists, practitioners deploying models in high-stakes applications, and diverse cultural communities whose values must be represented.

 doi:10.20944/preprints202602.0128.v1

As language models grow more capable and their societal impact increases, mechanistic interpretability becomes increasingly essential. Only by understanding how these systems work internally can we hope to ensure they remain beneficial, truthful, and aligned with the full diversity of human values across cultures and contexts. The research community must rise to this challenge with urgency, rigor, and humility about the difficulty of the task ahead.

References

1. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 27730–27744.
2. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* **2022**.
3. Casper, S.; Davies, X.; Shi, C.; Gilbert, T.K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* **2023**.
4. Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; Carter, S. Zoom in: An introduction to circuits. *Distill* **2020**, *5*, e00024–001.
5. Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread* **2021**, *1*, 12.
6. Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593* **2022**.
7. Conmy, A.; Mavor-Parker, A.; Lynch, A.; Heimersheim, S.; Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems* **2023**, *36*, 16318–16352.
8. 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.
9. 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 arXiv:2310.01405* **2023**.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
11. Geva, M.; Schuster, R.; Berant, J.; Levy, O. Transformer feed-forward layers are key-value memories. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 5484–5495.
12. Meng, K.; Bau, D.; Andonian, A.; Belinkov, Y. Locating and editing factual associations in GPT. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 17359–17372.
13. Russell, S. *Human compatible: Artificial intelligence and the problem of control*; Penguin, 2019.
14. Lin, S.; Hilton, J.; Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In Proceedings of the Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers), 2022, pp. 3214–3252.
15. Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N.A. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 3356–3369.
16. Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; Garrabrant, S. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820* **2019**.
17. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* **2020**.
18. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30.
19. Cammarata, N.; Goh, G.; Carter, S.; Schubert, L.; Petrov, M.; Olah, C. Curve detectors. *Distill* **2020**, *5*, e00024–003.
20. Olah, C.; Mordvintsev, A.; Schubert, L. Feature visualization. *Distill* **2017**, *2*, e7.
21. Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652* **2022**.

 doi:10.20944/preprints202602.0128.v1

22. Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* **2022**, *48*, 207–219.
23. 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, 2023, pp. 967–976.
24. Belrose, N.; Furman, Z.; Smith, L.; Halawi, D.; Ostrovsky, I.; McKinney, L.; Biderman, S.; Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112* **2023**.
25. 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**.
26. Syed, A.; Rager, C.; Conmy, A. Attribution patching outperforms automated circuit discovery. In Proceedings of the Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, 2024, pp. 407–416.
27. Goldowsky-Dill, N.; MacLeod, C.; Sato, L.; Arora, A. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969* **2023**.
28. Hanna, M.; Liu, O.; Variengien, A. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems* **2023**, *36*, 76033–76060.
29. Bills, S.; Cammarata, N.; Mossing, D.; Tillman, H.; Gao, L.; Goh, G.; Sutskever, I.; Leike, J.; Wu, J.; Saunders, W. Language models can explain neurons in language models. *OpenAI Blog* **2023**.
30. Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600* **2023**.
31. Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread* **2023**.
32. Turner, A.M.; Thiergart, L.; Leech, G.; Udell, D.; Mini, U.; MacDiarmid, M. Activation addition: Steering language models without optimization **2024**.
33. Geiger, A.; Lu, H.; Icard, T.; Potts, C. Causal abstractions of neural networks. In Proceedings of the Advances in Neural Information Processing Systems, 2021, Vol. 34, pp. 9574–9586.
34. Geiger, A.; Wu, Z.; Lu, H.; Rozner, J.; Kreiss, E.; Icard, T.; Goodman, N.; Potts, C. Inducing causal structure for interpretable neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 7324–7338.
35. Tigges, C.; Hollinsworth, O.J.; Geiger, A.; Nanda, N. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154* **2023**.
36. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* **2023**.
37. Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A.C.; Korbak, T.; Evans, O. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv preprint arXiv:2309.12288* **2023**.
38. Huang, Y.; Gupta, S.; Xia, M.; Li, K.; Chen, D. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987* **2023**.
39. 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.
40. 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.
41. Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; Manning, C.D. Fast model editing at scale. *arXiv preprint arXiv:2110.11309* **2021**.
42. 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. *arXiv preprint arXiv:2207.05221* **2022**.
43. Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 9802–9822.
44. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* **2023**.

 doi:10.20944/preprints202602.0128.v1

45. Turpin, M.; Michael, J.; Perez, E.; Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* **2023**, *36*, 74952–74965.
46. Bowman, S.R.; Hyun, J.; Perez, E.; Chen, E.; Pettit, C.; Heiner, S.; Lukošiūtė, K.; Askell, A.; Jones, A.; Chen, A.; et al. Measuring progress on scalable oversight for large language models. *arXiv arXiv:2211.03540* **2022**.
47. Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C.M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. Position: A roadmap to pluralistic alignment. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 46280–46302.
48. Bakker, M.; Chadwick, M.; Sheahan, H.; Tessler, M.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M.; et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in neural information processing systems* **2022**, *35*, 38176–38189.
49. Kirk, H.R.; Vidgen, B.; Röttger, P.; Hale, S.A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* **2024**, *6*, 383–392.
50. Alkhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; Diab, M. Investigating Cultural Alignment of Large Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 12404–12422.
51. Ramezani, A.; Xu, Y. Knowledge of cultural moral norms in large language models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 428–446.
52. Arora, A.; Kaffee, L.a.; Augenstein, I. Probing pre-trained language models for cross-cultural differences in values. In Proceedings of the Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP), 2023, pp. 114–130.
53. Nicholas, G.; Bhatia, A. Lost in translation: Large language models in non-English content analysis. *arXiv preprint arXiv:2306.07377* **2023**.
54. Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; Bao, M. The values encoded in machine learning research. In Proceedings of the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 173–184.
55. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* **2022**.
56. Räuker, T.; Ho, A.; Casper, S.; Hadfield-Menell, D. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In Proceedings of the 2023 ieee conference on secure and trustworthy machine learning (satml). IEEE, 2023, pp. 464–483.
57. Skalse, J.; Howe, N.; Krasheninnikov, D.; Krueger, D. Defining and characterizing reward gaming. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 9460–9471.
58. Gabriel, I. Artificial intelligence, values, and alignment. *Minds and machines* **2020**, *30*, 411–437.
59. Jenner, E.; Kapur, S.; Georgiev, V.; Allen, C.; Emmons, S.; Russell, S.J. Evidence of learned look-ahead in a chess-playing neural network. *Advances in Neural Information Processing Systems* **2024**, *37*, 31410–31437.
60. Huang, P.S.; Stanforth, R.; Welbl, J.; Dyer, C.; Yogatama, D.; Gowal, S.; Dvijotham, K.; Kohli, P. Achieving Verified Robustness to Symbol Substitutions via Interval Bound Propagation. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 4083–4093.
61. Irving, G.; Christiano, P.; Amodei, D. AI safety via debate. *arXiv preprint arXiv:1805.00899* **2018**.
62. Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; Legg, S. Scalable agent alignment via reward modeling: A research direction. *arXiv preprint arXiv:1811.07871* **2018**.
63. Wendler, C.; Veselovsky, V.; Monea, G.; West, R. Do llamas work in english? on the latent language of multilingual transformers. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 15366–15394.
64. Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; Irving, G. Red Teaming Language Models with Language Models. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 3419–3448.

 doi:10.20944/preprints202602.0128.v1

65. Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213* **2020**.
66. Schwitzgebel, E.; Garza, M. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy* **2015**, *39*, 98–119.