

Mechanistic Interpretability for Understanding Language Abilities in Large Language Models: A Survey

Xufeng Duan¹, Zhaoqian Yao¹, Xinyu Zhou^{2,3}, Zihao Fu^{1,*},
Beijia Kan¹, Yibo Wang¹, Bei Xiao¹, Hengyuan Zhang⁴,
Xuemei Tang⁵, Ercong Nie⁶, Zhenguang G. Cai^{1,*}

¹The Chinese University of Hong Kong ²Université Paris Cité

³Sorbonne Université ⁴The University of Hong Kong

⁵The Hong Kong Polytechnic University ⁶Shanghai Jiao Tong University

*Correspondence: zihao@cuhk.edu.hk; zhenguangcai@cuhk.edu.hk

Abstract

Large Language Models (LLMs) exhibit strong performance across language tasks, raising the question of whether this success reflects linguistic abstraction or reliance on surface-level statistical patterns. This survey reviews recent advances in **mechanistic interpretability (MI)** for analyzing candidate internal mechanisms underlying linguistic competence in LLMs. We trace the shift from model-agnostic analyses to transformer-specific methods and summarize core approaches including probing, neuron- and circuit-level analysis, causal abstraction, and targeted interventions. Synthesizing empirical findings, we argue that current evidence supports an intermediate view: LLMs appear to use reusable mechanisms for some grammatical behaviors, while these mechanisms remain frequency-conditioned, distributed, and only partially understood. We further review interpretability studies of multilingual models, highlighting evidence for both language-specific and shared representations. Finally, we outline open challenges in scalability, benchmarking, and ethics, and discuss directions toward a more cautious mechanistic account of language in LLMs. The curated list of papers for this work is available at <https://github.com/xufengduan/Awesome-Language-MI-Survey>.

1 Introduction

Large Language Models (LLMs) achieve strong performance across a wide range of language tasks, including translation and text generation. This success raises a fundamental question about the nature of their underlying linguistic knowledge: do LLMs internalize systematic grammatical and semantic abstractions akin to human-like *competence*, or do they primarily exploit surface-level statistical regularities in large-scale data to approximate linguistic *performance*? While their behavior suggests the emergence of nontrivial linguistic structure (Cai

Paper	Focus
Rai et al. (2025)	MI methods for Transformer.
Sharkey et al. (2025)	Roadmap and future directions.
Shu et al. (2025)	Feature disentanglement.
Mueller et al. (2026)	Causal Mediation Analysis.
Kendiukhov (2025)	Training dynamics.
Resck et al. (2025)	Multilingual.
Gao et al. (2026)	Intrinsic interpretability.
Graichen et al. (2026)	Syntactic Knowledge.
Zhang et al. (2026)	Locate, Steer, and Improve.

Table 1: Representative surveys in MI and their focus.

et al., 2024; Liu et al., 2025a), it remains plausible that such results arise from sophisticated pattern matching without the underlying symbolic or rule-based representations traditionally associated with human language (Chomsky, 2023; Bolhuis et al., 2024).

Mechanistic interpretability (MI) aims to clarify this distinction by decomposing a model’s black-box predictions into the internal computational structures, such as specialized attention heads or individual neurons for linguistic competence, that produce them. Recent studies have identified structured “circuits” or feature subspaces that appear to correlate with specific linguistic phenomena, offering partial analogies to human cognitive processes (Lindsey et al., 2025; Méloux et al., 2025). However, a critical challenge in this field is the conceptual gap between *mechanistic identification* (e.g., locating a circuit) and *linguistic attribution* (e.g., confirming that the circuit implements a generalizable rule). By scrutinizing these internal states, researchers seek to determine whether a model truly implements abstract generalizations or merely leverages high-dimensional heuristics that are sufficient for its training distribution.

Scope and audience. In this survey, *linguistic competence* refers primarily to internal knowledge of morphosyntax (e.g., agreement, binding, word order) and syntax–semantics interface (e.g., argu-

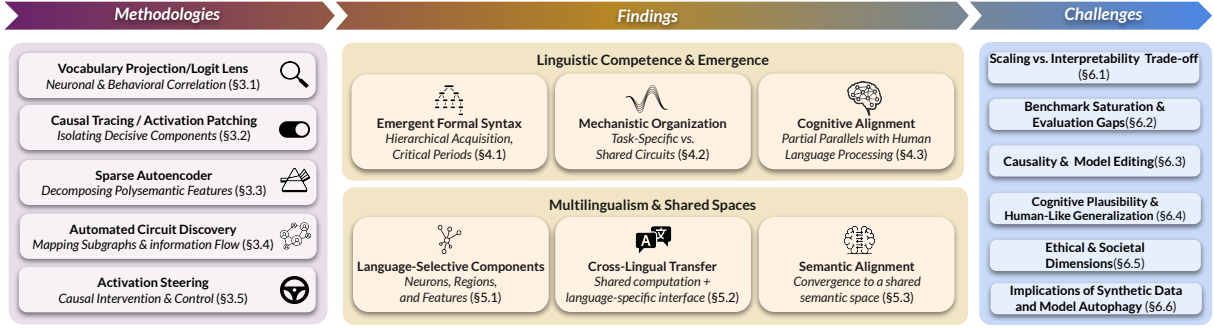


Figure 1: The overview of this survey.

ment structure and thematic roles). We discuss discourse, pragmatics, reasoning, and neural alignment only where mechanistic evidence directly connects them to linguistic representations. This focus is intentionally narrower than the full space of modern agentic, multimodal, or tool-using systems: our goal is to ask what MI reveals about structured language knowledge in Transformer-based LLMs. The survey is written for linguists seeking a map of what MI can and cannot establish, MI researchers seeking linguistic grounding, and NLP researchers interested in why models succeed or fail on linguistic evaluations.

Survey construction and coverage. This article is a structured conceptual survey rather than a preregistered systematic review. We began with recent surveys and public paper lists, and expanded the review through backward and forward citation tracing and targeted searches in Google Scholar, Semantic Scholar, ACL Anthology, and arXiv. Search queries combined interpretability methods (including probing, vocabulary projection, activation patching, causal abstraction, sparse autoencoders, circuit discovery, and activation steering) with linguistic phenomena (including agreement, binding, word order, filler-gap dependencies, argument structure, semantic inference, and multilingual representation). We included studies that analyze linguistic mechanisms in Transformer models and excluded general explainability work without a linguistic focus, non-Transformer studies except for historical context, and superseded versions. Our structured corpus comprises 210 papers selected according to the criteria above, with the complete paper list available in our GitHub repository.

Prior surveys primarily categorize interpretability methods (Table 1), leaving a critical gap in synthesizing what these techniques reveal about linguistic competence. We argue that mechanistic

evidence supports a **layered account** of linguistic implementation in LLMs: feature-level variables encode local grammatical and semantic distinctions; circuit-level mechanisms reuse computational subroutines across tasks; and multilingual models employ routing-level organization to link surface forms to shared semantic spaces. This thesis suggests that LLM behavior reflects structural organization rather than shallow memorization, without implying the existence of classical symbolic grammars.

This account occupies a middle position between two extremes (Boleda, 2025). Some mechanistic findings are difficult to reconcile with a purely local n-gram account, particularly when they identify explicit computational pathways associated with linguistic phenomena such as syntactic dependencies. However, such findings do not by themselves rule out broader forms of memorization or high-dimensional heuristic computation, nor do they justify claims of human-like symbolic competence. The discovered mechanisms remain distributed, frequency-conditioned, and model-specific. We therefore treat mechanistic evidence as a window into how specific models implement linguistic behavior, rather than as proof of a human-like cognitive architecture.

This survey presents a comprehensive overview of MI methods for studying linguistic competence in LLMs. Our contributions are as follows: (1) We review the evolution of interpretability techniques, from early probing approaches to recent advances, and clarify the linguistic claims each method can support. (2) We synthesize empirical evidence on how linguistic competence dynamically emerges in LLMs, uncover the underlying structural mechanisms, and evaluate its neurocognitive alignment with human language processing. (3) We extend the discussion to multilingual models, distinguish-

ing between language-specific mechanisms and shared, language-agnostic representational spaces that support cross-lingual transfer. (4) We identify open challenges in scaling interpretability to increasingly large models, limitations of existing linguistic benchmarks, and tensions between interpretability and performance, and outline directions for future research.

2 Interpretability Methodology

Running example. Throughout this section, consider subject–verb agreement in a sentence such as *The keys near the cabinet are/*is old*. A behavioral contrast tests whether the model prefers the grammatical continuation; vocabulary projection asks when the plural prediction becomes visible; probing or sparse features ask where number information is encoded; patching and causal abstraction test whether a component or subspace is functionally involved; circuit discovery traces how that information reaches the prediction; and steering asks whether intervening on it predictably changes the output. These methods therefore provide different, complementary levels of evidence rather than interchangeable explanations.

2.1 Vocabulary Projection

Vocabulary Projection (or the **Logit Lens**) (Nostalgebraist, 2020) interprets intermediate representations by projecting hidden states $\mathbf{h}_\ell$ directly onto the vocabulary via the unembedding matrix W_U , revealing real-time predictions (Chuang et al., 2024):

$$\text{Logits}_\ell = W_U \cdot \text{LayerNorm}(\mathbf{h}_\ell) \quad (1)$$

This method tracks layer-by-layer prediction evolution (Geva et al., 2022). Enhancements include the *Tuned Lens* (Belrose et al., 2025), which calibrates intermediate representations, and the *Future Lens* (Pal et al., 2023), which predicts subsequent states.

Linguistic leverage. Vocabulary projection helps localize when a model favors specific tokens (e.g., number-marked verbs). In plain terms, this converts an intermediate state into the token preferences it would support if it were read out at that layer. However, it does not confirm the causal necessity of these representations.

2.2 Activation Patching

To establish causality, **Activation Patching** systematically replaces a component’s activation in

a corrupted input (x_{corrupt}) with the corresponding activation from a clean input (x_{clean}) (Meng et al., 2023a; Wang et al., 2022; Zhang and Nanda, 2024). If the intervention restores the model’s performance on a target task, that component is considered causally relevant (Finlayson et al., 2021; Todd et al., 2024). **Attribution Patching** (Nanda, 2023) further refines this by using gradients to approximate these interventions linearly.

The importance of a component i is quantified by its *indirect effect* (IE), measuring the change in target token probability p_r after patching:

$$p_r(f(x_{\text{corrupt}}; a_i \leftarrow a_i(x_{\text{clean}}))) - p_r(f(x_{\text{corrupt}})) \quad (2)$$

While powerful, these methods require caution (Heimersheim and Nanda, 2024). Makelov et al. (2024) warn of an “interpretability illusion,” where interventions might activate dormant pathways rather than revealing the natural mechanism.

Linguistic leverage. Patching provides direct causal evidence for linguistic tasks. A positive indirect effect means that restoring the clean activation increases the target answer relative to the corrupted run. Its validity depends on demonstrating generalizability across diverse lexical items, templates, and distractors to avoid identifying prompt-specific pathways.

2.3 Causal Abstraction and Distributed Alignment

While activation patching asks whether a model component is causally involved in a behavior, **causal abstraction** asks a stronger question: whether a model representation implements a hypothesized high-level variable. The goal is to identify a neural representation whose causal structure corresponds to that of an interpretable abstraction, such as grammatical number, syntactic role, or a semantic property. In practice, methods such as Distributed Alignment Search identify candidate subspaces associated with the hypothesized variable, and *interchange interventions* replace representations across inputs to test whether the resulting model behavior matches the counterfactual prediction of the high-level abstraction (Rodriguez et al., 2025; Boguraev et al., 2025; Desai and Nair, 2026). In subject–verb agreement, patching may show that a representation causally affects predicting *are* rather than *is*. Causal abstraction goes further by testing whether interventions on a neural subspace systematically correspond to interventions on an

abstract variable such as grammatical number, providing stronger evidence for an interpretable causal representation.

Linguistic leverage. Causal abstraction tests whether a linguistic variable, such as grammatical number or syntactic role, is functionally implemented in a neural subspace rather than merely decodable from it. Successful interchange interventions provide evidence that manipulating this representation produces the counterfactual behavior predicted by the hypothesized variable. However, such abstractions must generalize across lexical and structural variation to support broader linguistic claims.

2.4 Sparse Autoencoders

A fundamental barrier to interpreting individual units (e.g., neurons) is **superposition**, where models compress more features than they have dimensions, resulting in *polysemantic neurons* (Elhage et al., 2022). For example, a single neuron might fire for both “plural nouns” and “Python code comments,” making it impossible to assign it a unique linguistic label. **Sparse Autoencoders (SAEs)** address this by decomposing dense, polysemantic activations into an overcomplete, sparse basis where each direction represents a single concept (Bricken et al., 2023; Huben et al., 2024; He et al., 2025; Shu et al., 2025; Tang et al., 2026c). SAEs reconstruct an activation $\mathbf{x}$ as a linear combination of features:

$$\mathbf{x} \approx \hat{\mathbf{x}} = \mathbf{b}_{dec} + \sum_i f_i(\mathbf{x}) \mathbf{d}_i \quad (3)$$

where $f_i(\mathbf{x})$ is the sparse feature activation (via L_1 regularization) and $\mathbf{d}_i$ is the decoder direction.

Researchers have used this to identify monosemantic units for specific linguistic phenomena (e.g., “German plural endings”) (Jing et al., 2025). However, Kantamneni et al. (2025) caution that SAEs require rigorous evaluation, as they do not consistently outperform baselines in downstream tasks.

Linguistic leverage. SAEs improve inspectability by disentangling dense activations. The primary evidential challenge is demonstrating that these extracted features are causally active: does ablating, steering, or tracing a specific feature reliably alter the corresponding linguistic behavior in a model?

2.5 Circuit Discovery

While SAEs disentangle representations into interpretable features, these units can also be com-

posed into higher-level computational mechanisms. Building on activation patching, **Circuit Discovery** (Zhang and Nanda, 2024; Mondorf et al., 2025) aims to identify the complete subgraph of components, such as attention heads and neurons, that are responsible for a specific task. While early work relied on manual hypothesis testing (Olah et al., 2020), recent approaches employ automated search algorithms. Automated Circuit Discovery (Conmy et al., 2023) automates the process of pruning the computational graph. It iteratively ablates edges in the network, keeping only those whose removal would significantly degrade performance on a specific task. To address the difficulty of analyzing dense MLP layers in circuits, Dunefsky et al. (2024) introduce *Transcoders*, which approximate MLPs with sparse, interpretable components. More recently, researchers have shifted focus from “component circuits” (where nodes are heads/MLPs) to “SAE feature circuits”. Marks et al. (2025) and Kharlapenko et al. (2025) utilize sparse autoencoders to define circuits over fine-grained semantic features rather than coarse model components. This allows for the discovery of circuits that are more interpretable, as edges represent causal dependencies between specific concepts (e.g., a “plural subject” feature causing a “plural verb” prediction) rather than opaque vectors.

Linguistic leverage. Circuit discovery is most informative when it identifies a reusable algorithmic pathway rather than a single correlated unit. For language, the central test is whether the circuit implements an abstraction across lexical and structural variation, or whether it only solves a narrow benchmark template.

2.6 Activation Steering

Methodology has expanded from observation to control, enabling direct intervention in a model’s linguistic behavior. **Activation Steering** involves injecting vectors (Todd et al., 2024; Wu et al., 2025a; Poterì et al., 2025; Karny et al., 2025) into the residual stream to elicit specific behaviors. Turner et al. (2024) introduced Activation Addition (ActAdd), utilizing mean activation differences between positive and negative examples. To improve robustness, Contrastive Activation Addition (CAA) (Rimsky et al., 2024) refines this by averaging differences across contrastive pairs (e.g., “Helpful” vs. “Unhelpful”). Building on this, Steering Target Atoms (STA) (Wang et al., 2025a) leverages

SAEs to decompose steering vectors into sparse, interpretable features, allowing for more precise control with fewer side effects.

Linguistic leverage. Steering links explanation to intervention: if linguistic properties are predictably shifted, the underlying representation is confirmed as functional rather than merely descriptive. However, these results require extra validation to ensure that steering does not inadvertently degrade unrelated fluency, style, or factual accuracy.

3 MI on Linguistic Competence

This section reviews MI evidence on linguistic competence, focusing on how grammatical knowledge emerges during training, how it is implemented in model internals, and how far these representations align with human language processing. Together, these perspectives shift the analysis from whether models perform well on linguistic tasks to how such competence is formed, organized, and constrained.

3.1 Dynamic View: How It Emerges

Behavioral benchmarks often report strong performance on core grammatical phenomena (Hu et al., 2020). However, analyses of internal model states suggest that this competence is not a monolithic body of knowledge acquired instantaneously, but rather the outcome of a complex, frequency-dependent developmental process.

Critical Periods and Structural Evolution Evidence suggests that the timing of language acquisition may be structurally important. Duan et al. (2025a) identify “critical periods” in training, specific windows where rapid internal syntactic specialization occurs. Ou et al. (2025) and Bayazit et al. (2026) trace this evolution using circuit analysis and crosscoders, observing that linguistic features are not merely added but actively consolidated during pre-training. This temporal evolution is complemented by cross-scale findings from Liu et al. (2025b), who note that syntactic acquisition also saturates at specific model scales.

From Memorization to Abstraction The acquisition of language competence during training appears to follow a hierarchical trajectory shaped by data distribution. In early stages, models preferentially learn high-frequency tokens and local n-grams, while rarer and long-distance dependencies emerge only through gradual refinement (Wei et al., 2021; Chang et al., 2024b). This suggests that linguistic abstraction is not purely symbolic

but remains anchored to the statistical density of the training data. At the mechanistic level, this temporal progression has been linked to shifts in internal circuitry, including reported transitions from transient in-context strategies to more stable in-weights representations (Singh et al., 2023). For a comprehensive review, see (Kendiukhov, 2025).

What we learn. A recurring pattern across these studies is developmental: formal linguistic competence appears to emerge through staged, frequency-dependent trajectories rather than all at once. Mechanistic evidence points to increasingly stable grammatical variables during training, but claims about abstract rules are best supported when they survive lexical, frequency, and structural controls rather than only benchmark-level accuracy.

3.2 Mechanistic View: Specialized vs. Shared Circuits

Following the gradual emergence of linguistic capabilities during training, it remains unclear whether neural networks rely on isolated modules or reconfigure overlapping subsystems to support diverse tasks. While numerous studies have identified language-specific neurons and features, recent circuit-level analyses increasingly challenge the assumption of strict functional independence (Kryvosheieva et al., 2026). Merullo et al. (2024) report that circuits identified for Indirect Object Identification (IOI) are reused in tasks sharing similar algorithmic structure, with substantial overlap in attention heads. These findings suggest that models rely on reusable circuits rather than strictly task-specific mechanisms (e.g., Conmy et al., 2023). Beyond circuit overlaps, researchers have identified *functional neurons* whose targeted perturbation can degrade or redirect performance. The works (Gurgurov et al., 2025; Huo et al., 2024; Wu et al., 2024; Tang et al., 2024) report that ablating these neurons yields large drops in downstream accuracy.

However, the degree of localization varies: simple local agreement checks may be handled by a compact set of heads (Olah et al., 2020; Gurnee et al., 2023), whereas complex reasoning requires interactions among numerous polysemantic units (Marks et al., 2025).

Causal abstractions across linguistic constructions. Causal-abstraction studies provide a complementary test of whether apparently different behaviors rely on a shared internal variable. For semantic property inference, Rodriguez et al. (2025) analyze

how representational similarity supports property inheritance. For syntax, [Boguraev et al. \(2025\)](#) report a transferable causal representation across English filler–gap constructions, while [Desai and Nair \(2026\)](#) find shared but item-sensitive mechanisms in models trained under developmental data constraints. Related analyses examine gradient blocking in syntactic islands ([Boguraev and Mahowald, 2026](#)), shared syntactic mechanisms ([Kumon and Yanaka, 2026](#)), and pronoun fidelity ([Trinley et al., 2026](#)). Together, these studies move beyond localization by testing explicit counterfactual mappings between linguistic variables and neural representations. Their current scope nevertheless remains concentrated in controlled English-language settings and a limited set of model families.

What we learn. Current evidence favors reusable subroutines and, in several controlled settings, transferable causal abstractions over fully isolated task-specific modules. However, the literature does not yet establish that a single mechanism for agreement, binding, filler–gap dependencies, or semantic inference reliably recurs across architectures, corpora, random seeds, and intervention protocols.

3.3 Cognitive View: Partial Alignment with Human Language Processing

Do model representations predict human neural responses? Hidden states can predict neural activity during naturalistic language comprehension, and the strongest correspondence can vary with model depth and neural response timescale ([Hosseini et al., 2024](#); [Gao et al., 2025](#); [Goldstein et al., 2025](#)). Predictive correspondence, however, does not establish a shared algorithm.

Do language-selective model components align with human language regions? Language-selective units identified in LLMs show stronger correspondence with human language-selective regions than control components in several analyses ([AlKhamissi et al., 2025a](#); [Kumar et al., 2024](#)). Theory-driven interventions can also identify small sets of units associated with particular psycholinguistic behaviors ([Duan et al., 2025b](#)). These results support selective alignment, not a one-to-one mapping between neurons or cognitive mechanisms.

How does alignment change during training? Brain alignment develops non-uniformly during training and is reported to correlate more

strongly with formal linguistic competence than with broader functional abilities ([AlKhamissi et al., 2025b](#)). This developmental relationship remains sensitive to the neural dataset, model family, and alignment metric.

What we learn. Current findings establish selective representational correspondence under particular analyses. They do not demonstrate shared algorithms, common cognitive mechanisms, or broadly human-like language processing.

4 Multilingualism

This section reviews MI evidence on multilingualism, focusing on how multilingual LLMs represent language-specific information, enable cross-lingual transfer, and align semantic content across languages. Together, these findings suggest that multilingual behavior is supported by a hybrid organization: specialized components mediate language identity and surface-form control, while shared representations enable transfer and semantic convergence.

4.1 Language-Selective Components

A growing body of work reports language-selective components in multilingual decoder-style LLMs at multiple granularities, from coarse region-level localization to fine-grained, steerable features.

At the neuron level, language competence exhibits partial spatial localization. [Kojima et al. \(2024\)](#) find these language-specific neurons are concentrated in early and late layers, and that perturbing fewer than 1% of neurons can substantially alter output language probabilities. [Tang et al. \(2024\)](#) further study this localized sensitivity by introducing Language Activation Probability Entropy to pinpoint these neurons, reporting that deactivating a small subset selectively impairs comprehension or generation in a target language. Beyond individual neurons, region-level analyses show similar patterns. [Zhang et al. \(2024\)](#) identify distributed *monolingual regions* associated with the syntactic and lexical properties of particular languages.

Beyond static neuron parameters, dynamic routing mechanisms further isolate language processing. On the attention side, [Liu et al. \(2026\)](#) identify both language-specific and language-general attention heads. Circuit-level analyses further support this view: Anthropic’s circuit tracing work on Claude 3.5 Haiku shows that multilingual behavior recruits language-dominant clusters along-

side abstract, shared pathways, with the proportion of shared circuitry increasing with model scale (Ameisen et al., 2025; Lindsey et al., 2025).

Recent SAE-based studies provide additional perspective and fine-grained steering. Deng et al. (2025) and Andrylie et al. (2026) identify monosemantic language-specific features whose ablation selectively degrades individual languages. Building on this, Zhong et al. (2026) propose that language control relies on a sparse, cross-layer set of dimensions, and Chou et al. (2025) report zero-shot language control through sparse feature steering.

Overall, these findings suggest that multilingual LLMs, despite extensive parameter sharing, preserve some specialized internal structures that support language-specific processing.

What we learn. Language-selective components offer a useful entry point to study how multilingual LLMs encode language identity, typological variation, and language-specific form. Current evidence suggests that parameter sharing does not eliminate specialized structure; rather, multilingual models combine localized language-sensitive units with broader shared pathways.

4.2 Cross-Lingual Transfer

We define cross-lingual transfer as a model’s ability to reuse knowledge or task competence learned in one language when processing inputs in another. Mechanistically, recent evidence supports a two-component view: transfer is primarily enabled by language-agnostic computation in intermediate layers, while language-specific units help route information between language-dependent surface forms and that shared computation.

The Multilingual Workflow and Shared Latent Spaces. Zhao et al. (2024) formalize this as a multilingual workflow, where non-English inputs are progressively mapped into an English-like internal workspace for reasoning, and later projected back to the query language during generation. This remains a model-specific hypothesis rather than a settled account of multilingual computation. The mechanistic explanations for this transfer are becoming more precise: Tezuka and Inoue (2025) propose the *Transfer Neuron Hypothesis*, identifying specific MLP neurons associated with transformations between language-specific and shared latent spaces. As evidence for partial functional independence, Zhao et al. (2024) report that targeted tuning of those neurons can selectively im-

prove a language without broadly affecting others. To map this flow to the entire network, Harrasse et al. (2026) utilize cross-layer transcoders, finding that while internal representations are largely shared, language-specific decoding emerges in the final layers driven by high-frequency features. Furthermore, Brinkmann et al. (2025) reports shared feature directions for grammatical concepts even in models trained primarily on English.

Developmental Dynamics of Transfer. Chen et al. (2025) posit that during training, an LLM initially allocates its resources to a “primary” or dominant language, only later partitioning its internal representation to accommodate new languages. Their empirical analysis tracks *language transferring neurons* that mark the point at which the model begins establishing dedicated circuits for additional languages. Riemenschneider and Frank (2025) observe a compression process during training, where models initially form language-specific representations that gradually converge into cross-lingual abstractions.

While the localization of cross-lingual mechanisms is becoming precise, the causal utility of these components remains contested. Mondal et al. (2025) systematically test neuron-specific interventions and find that manipulating language-specific neurons does not reliably improve downstream cross-lingual performance on benchmarks such as XNLI and XQuAD for low-resource languages.

What we learn. Cross-lingual transfer appears to depend on an interaction between language-specific routing and shared task-solving computation. Localizing language-specific neurons is therefore not sufficient: the open question is how these components coordinate with shared intermediate representations, and why targeted interventions sometimes change language identity without improving downstream transfer.

4.3 Mechanisms of Semantic Alignment

While multilingual LLMs develop components that exhibit strong language selectivity, a growing body of work shows that these models converge on a shared latent space for semantic content. Zhong et al. (2024) and Wendler et al. (2024) find that representations for inputs in different languages often pass through English-like intermediate spaces. Generalizing this observation, Zeng et al. (2025) propose a hidden-state *lingua franca*: a shared semantic manifold to which inputs are

mapped irrespective of language, with larger models and increased training further strengthening this language-agnostic alignment.

This shared semantic structure is supported by partially overlapping subspaces and token representations. Zhang et al. (2025a) report that multilingual embeddings cluster near-synonymous tokens across languages in overlapping regions, while Liu et al. (2024) find that domain knowledge is integrated into these spaces in a way that balances shared semantic representations with language- or domain-specific structure. Together, these results point to a dual organization where localized mechanisms coexist with a common semantic hub.

The shared space itself may be English-centric in some models. Rather than showing that multilingual LLMs literally “think in English,” Schut et al. (2025) provide evidence that intermediate states can become English-aligned before generation in the target language. Related evidence extends beyond translation: Brinkmann et al. (2025) report that grammatical concepts such as number and gender can be encoded in shared feature directions across typologically diverse languages, and Son et al. (2025) find semantic alignment across models of varying scales.

Overall, these findings suggest that multilingual LLMs implement a hybrid representational strategy: language-specific circuits capture idiosyncratic form, while higher-level semantic representations converge into a shared latent space. Understanding this balance between specialization and convergence has practical implications for model design, enabling targeted control of language-specific behavior while preserving cross-lingual generalization.

What we learn. The multilingual evidence extends the layered account: local language-specific features and circuits handle surface-form control, while higher-level representations increasingly converge into a shared semantic hub. The English-centric nature of this hub is plausible but not settled; it should be tested across more low-resource, non-Indo-European, and typologically distant languages before being treated as a universal property of multilingual LLMs.

5 Challenges and Open Questions

Despite progress in MI, model scaling complicates the isolation and manipulation of linguistic mechanisms. Simultaneously, existing evaluation

paradigms face increasing scrutiny regarding their scope and inclusivity.

MI is also most useful when triangulated with complementary approaches. Behavioral evaluation constrains what a model can do; probing and representational geometry characterize what information is encoded; cognitive modeling tests whether model behavior and representations align with human processing; and MI contributes causal internal evidence about how a behavior is produced. A full account of linguistic competence requires these sources of evidence to be interpreted together.

5.1 Scaling vs. Interpretability Trade-Off

While scaling enhances performance, it challenges fine-grained interpretability. Lieberum et al. (2023) observe that circuit-level analysis becomes increasingly intractable at scale, suggesting emergent behaviors may resist decomposition. Furthermore, Méloux et al. (2025) argue that extensive poly-semanticity in complex models precludes unique interpretations. Conversely, Conmy et al. (2023) provide evidence that automated circuit discovery can scale beyond fully manual analysis, indicating that methodological advances may offset some complexity. However, the tractability of such approaches, and techniques like SAE (Marks et al., 2025; Jing et al., 2025), for massive-scale models remains unproven.

5.2 Evaluation Gaps

Although benchmarks like CausalGym (Arora et al., 2024), Holmes (Waldis et al., 2024) and BHASA (Leong et al., 2023) have expanded linguistic coverage, significant gaps persist. Minimal-pair and targeted syntactic tests (Hu and Levy, 2023; Arora et al., 2024) effectively probe local grammar but often neglect higher-level reasoning, discourse, and pragmatics. Moreover, the focus on high-resource languages limits the detection of failures in under-represented contexts, risking interpretability methods that overfit to narrow linguistic variations. Developing robust, culturally sensitive frameworks for low-resource languages remains a critical challenge (Chang et al., 2024a; Zeng et al., 2025).

5.3 Causality and Model Editing

A core goal of MI is causal intervention. Techniques such as attribution patching (Ferrando and Voita, 2024) and neuron or feature steering (Lindsey et al., 2025; Banayeeanzade et al., 2026) test whether specific components can be causally linked

to linguistic behavior. However, these interventions often lack stability: localized changes can propagate through the network, producing unintended side effects (Stickland et al., 2024).

Superposition further complicates causal control. As shown by Elhage et al. (2022), polysemantic neurons encode multiple features simultaneously, making it difficult to edit one function without disrupting others. Developing reliable methods to identify modular, independently editable circuits and organize them into a safe, reusable “circuit library” remains an open challenge.

5.4 Cognitive Plausibility and Human-Like Generalization

Alignment between LLMs and human language processing remains partial. AlKhamissi et al. (2025b) observe that while specific units mirror human language networks, divergence increases as models integrate broader reasoning capabilities. Training dynamics also echo human acquisition via staged morphological and syntactic learning (Choshen et al., 2022; Evanson et al., 2023; Duan et al., 2025a), yet generalization remains limited: LLMs transfer argument-structure patterns mainly when the relevant contextual relation was seen in pre-training (Wilson et al., 2023). Distinguishing shared cognitive implementation from behavioral convergence is critical for interpretability (Zhao et al., 2025). Despite calls for analyzing critical training transitions (Arora et al., 2024), the fundamental cognitive plausibility of LLM competence remains an open question.

5.5 Ethical and Societal Dimensions

MI also raises ethical considerations, particularly regarding bias, fairness, and language equity. Neuron-level analyses can expose how models encode stereotypes or toxic associations, enabling targeted mitigation strategies (Iqbal et al., 2025; Li et al., 2024). From a global perspective, interpretability highlights persistent linguistic inequities: multilingual models remain strongly English-centric (Guo et al., 2025), with uneven performance across languages (Chang et al., 2024a). Developing interpretable and equitable models for underrepresented languages, remains an underexplored but critical challenge.

5.6 Implications of Synthetic Data and Model Collapse

As LLMs are increasingly trained on machine-generated data, interpretability faces new uncertainties. Guo et al. (2024) show that iterative training on synthetic text can lead to lexical or syntactic homogenization, potentially simplifying internal circuits. Moreover, large-scale synthetic corpora training may also distort internal representations in ways that obscure the origins of linguistic behavior (Chang et al., 2024a; Üveges and Ring, 2025). Whether mechanistic insights derived from current models will generalize to future systems trained predominantly on synthetic data remains an open and pressing question.

6 Conclusion

This survey reviewed MI advances for understanding linguistic abilities in LLMs, from static probing to circuit discovery (Conmy et al., 2023) and causal intervention (Meng et al., 2023a). Together, these methods provide evidence about candidate mechanisms underlying grammatical and semantic competence and support a hybrid view in which LLMs can exhibit structured abstraction while still relying on surface heuristics.

The reviewed evidence is difficult to reconcile with a pure memorization account: several studies identify reusable circuits for grammatical phenomena such as indirect object identification and subject–verb agreement (Wang et al., 2022; Merullo et al., 2024). At the same time, these capabilities are implemented through distributed, polysemantic representations (Elhage et al., 2022; Marks et al., 2025) and frequency-dependent learning dynamics (Chang et al., 2024b; Wei et al., 2021), which differ from classical symbolic accounts of grammar. In multilingual settings, this complexity is further reflected in evidence for language-specific components (Tang et al., 2024; Zhang et al., 2024) and proposed shared semantic representations such as a model-internal lingua franca (Zeng et al., 2025; Wu et al., 2025a).

Looking ahead, progress will hinge on addressing the trade-off between model scale and interpretability (Lieberum et al., 2023), developing evaluation benchmarks that extend beyond local syntactic competence (Diego-Simón et al., 2025), expanding multilingual replication beyond high-resource languages, and triangulating MI with behavioral and cognitive evidence.

Limitations

While this survey attempts to provide a critical synthesis of MI as applied to linguistic competence, several constraints limit its scope.

First, our synthesis is inherently restricted by the **monolingual bias** in existing MI literature. While we dedicate a section to multilingualism, the majority of available research focuses on English or high-resource languages; consequently, our conclusions regarding internal linguistic mechanisms may not generalize to typologically diverse or low-resource languages where data scarcity affects representational development (Chang et al., 2024a; Guo et al., 2025).

Second, we acknowledge a **methodological risk of inflation** in our synthesis. Because MI research often relies on “toy” settings or templated prompts to isolate circuits, the significance of these findings might be over-represented in this survey when compared to the model’s behavior on complex, naturalistic language. As argued by Makelov et al. (2024), what we identify as a “linguistic circuit” may be a compensatory pathway, and our survey is limited by the current field’s inability to definitively distinguish between these two.

Finally, we restricted our scope to Transformer-based architectures. As new paradigms like State Space Models (e.g., Mamba) emerge, the applicability of attention-based interpretability techniques (Olah et al., 2020) will need to be re-evaluated.

References

- Farah Adeeba, Brian Dillon, Hassan Sajjad, and Rajesh Bhatt. 2025. *UrbLimp: A benchmark for evaluating the linguistic competence of large language models in urdu*. Preprint, arXiv:2508.01006.
- Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025a. *The LLM language network: A neuroscientific approach for identifying causally task-relevant units*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, Albuquerque, New Mexico. Association for Computational Linguistics.
- Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025b. *From language to cognition: How llms outgrow the human language network*. Preprint, arXiv:2503.01830.
- Wafa Abdullah Alrajhi, Hend Al-Khalifa, and Abdulmalik AlSalman. 2022. *Assessing the linguistic knowledge in Arabic pre-trained language models using minimal pairs*. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 185–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, and 8 others. 2025. *Circuit tracing: Revealing computational graphs in language models*. Anthropic Technical Report.
- Lyzander Marciano Andrylie, Inaya Rahmanisa, Mahardika Krisna Ihsani, Alfian Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2026. *Sparse autoencoders can capture language-specific concepts across diverse languages*. Preprint, arXiv:2507.11230.
- Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. *Causalgym: Benchmarking causal interpretability methods on linguistic tasks*. Preprint, arXiv:2402.12560.
- Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. 2026. *Psychological steering in LLMs: An evaluation of effectiveness and trustworthiness*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1719–1771, San Diego, California, United States. Association for Computational Linguistics.
- Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Achille Fusco, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi, and Cristiano Chesi. 2025. *BLiMP-IT: Harnessing automatic minimal pair generation for Italian language model evaluation*. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 64–71, Cagliari, Italy. CEUR Workshop Proceedings.
- Deniz Bayazit, Aaron Mueller, and Antoine Bosselut. 2026. *Crosscoding through time: Tracking emergence & consolidation of linguistic representations throughout LLM pretraining*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1353–1377, San Diego, California, United States. Association for Computational Linguistics.
- Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. *TurbLimp: A turkish benchmark of linguistic minimal pairs*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 16506–16521. Association for Computational Linguistics.

- David Beauchemin, Pier-Luc Veilleux, Richard Khoury, and Johanna-Pascale Roy. 2026. [QFrBLiMP: a Quebec-French benchmark of linguistic minimal pairs](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4274–4304, Rabat, Morocco. Association for Computational Linguistics.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. [Eliciting latent predictions from transformers with the tuned lens](#). *Preprint*, arXiv:2303.08112.
- Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. [Is incoherence surprising? targeted evaluation of coherence prediction from language models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.
- Usha Bhalla, Alex Oesterling, Claudio Mayrunk Verdun, Himabindu Lakkaraju, and Flavio Calmon. 2026. [Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability](#). In *The Fourteenth International Conference on Learning Representations*.
- Sasha Boguraev and Kyle Mahowald. 2026. [Causal drawbridges: Characterizing gradient blocking of syntactic islands in transformer lms](#). *Preprint*, arXiv:2604.13950.
- Sasha Boguraev, Christopher Potts, and Kyle Mahowald. 2025. [Causal interventions reveal shared structure across English filler-gap constructions](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25021–25042, Suzhou, China. Association for Computational Linguistics.
- Gemma Boleda. 2025. [LLMs as a synthesis between symbolic and distributed approaches to language](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9365–9379, Suzhou, China. Association for Computational Linguistics.
- Johan J Bolhuis, Stephen Crain, Sandiway Fong, and Andrea Moro. 2024. [Three reasons why ai doesn’t model human language](#). *Nature*, 627(8004):489.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Holt McCauley, et al. 2023. [Towards monosemanticity: Decomposing language models with dictionary learning](#). *Transformer Circuits Thread*.
- Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. [Large language models share representations of latent grammatical concepts across typologically diverse languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.
- Daniel Brubaker, William Sheffield, Junyi Jessy Li, and Kanishka Misra. 2026. [Wugnatives: Novel entity inferences of language models from discourse connectives](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6109–6127, Rabat, Morocco. Association for Computational Linguistics.
- Samuel Cahyawijaya, Ruochen Zhang, Jan Christian Blaise Cruz, Holy Lovenia, Elisa Gilbert, Hiroki Nomoto, and Alham Fikri Aji. 2025. [Thank you, stingray: Multilingual large language models can not \(yet\) disambiguate cross-lingual word senses](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3228–3250, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhenguang G. Cai, Xufeng Duan, David A. Haslett, Shuqi Wang, and Martin J. Pickering. 2024. [Do large language models resemble humans in language use?](#) *Preprint*, arXiv:2303.08014.
- Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024a. [When is multilinguality a curse? language modeling for 250 high- and low-resource languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.
- Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024b. [Characterizing learning curves during language model pre-training: Learning, forgetting, and stability](#). *Transactions of the Association for Computational Linguistics*, 12:1346–1362.
- Jiawei Chen, Wentao Chen, Jing Su, Jingjing Xu, Hongyu Lin, Mengjie Ren, Yaojie Lu, Xianpei Han, and Le Sun. 2025. [The rise and down of babel tower: Investigating the evolution process of multilingual code large language model](#). In *The Thirteenth International Conference on Learning Representations*.
- Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. [Finding universal grammatical relations in multilingual bert](#). *Preprint*, arXiv:2005.04511.
- Zewen Chi, Heyan Huang, and Xian-Ling Mao. 2023. [Can cross-lingual transferability of multilingual transformers be activated without end-task data?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12572–12584, Toronto, Canada. Association for Computational Linguistics.

- Odysseas S. Chlapanis, Orfeas Menis Mastromichalakis, and Christos H. Papadimitriou. 2026. [The grounding gap: How llms anchor the meaning of abstract concepts differently from humans](#). *Preprint*, arXiv:2605.08837.
- Hakaze Cho, Haolin Yang, Brian M. Kurkoski, and Naoya Inoue. 2025. [Binary autoencoder for mechanistic interpretability of large language models](#). *Preprint*, arXiv:2509.20997.
- Ikhyun Cho and Julia Hockenmaier. 2025. [Analyzing multilingualism in large language models with sparse autoencoders](#). In *Second Conference on Language Modeling*.
- Noam Chomsky. 2023. [Noam chomsky: The false promise of chatgpt](#). *The New York Times*. Opinion.
- Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. [The grammar-learning trajectories of neural language models](#). *Preprint*, arXiv:2109.06096.
- Cheng-Ting Chou, George Liu, Jessica Sun, Cole Blondin, Kevin Zhu, Vasu Sharma, and Sean O’Brien. 2025. [Causal language control in multilingual transformers via sparse feature steering](#). *Preprint*, arXiv:2507.13410.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. [Dola: Decoding by contrasting layers improves factuality in large language models](#). *Preprint*, arXiv:2309.03883.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. [Towards automated circuit discovery for mechanistic interpretability](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Xinnan Dai, Chung-Hsiang Lo, Kai Guo, Shenglai Zeng, Dongsheng Luo, and Jiliang Tang. 2025. [Uncovering graph reasoning in decoder-only transformers with circuit tracing](#). *Preprint*, arXiv:2509.20336.
- Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. [Unveiling language-specific features in large language models via sparse autoencoders](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.
- Atrey Desai and Sathvik Nair. 2026. [Filling in the mechanisms: How do LMs learn filler-gap dependencies under developmental constraints?](#) In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 34799–34815, San Diego, California, United States. Association for Computational Linguistics.
- Pablo J. Diego-Simón, Emmanuel Chemla, Jean-Rémi King, and Yair Lakretz. 2025. [Probing syntax in large language models: Successes and remaining challenges](#). *Preprint*, arXiv:2508.03211.
- Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhengguang G. Cai. 2024. [Hlb: Benchmarking llms’ humanlikeness in language use](#). *Preprint*, arXiv:2409.15890.
- Xufeng Duan, Zhaoqian Yao, Yunhao Zhang, Shaonan Wang, and Zhengguang G. Cai. 2025a. [How syntax specialization emerges in language models](#). *Preprint*, arXiv:2505.19548.
- Xufeng Duan, Xinyu Zhou, Bei Xiao, and Zhengguang Cai. 2025b. [Unveiling language competence neurons: A psycholinguistic approach to model interpretability](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10148–10157, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. [Transcoders find interpretable llm feature circuits](#). *Preprint*, arXiv:2406.11944.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. [Toy models of superposition](#). *Preprint*, arXiv:2209.10652.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. [A mathematical framework for transformer circuits](#). *Transformer Circuits Thread*.
- Linnea Evanson, Yair Lakretz, and Jean-Rémi King. 2023. [Language acquisition: do children and language models follow similar learning stages?](#) *Preprint*, arXiv:2306.03586.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. [Alphaedit: Null-space constrained model editing for language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Javier Ferrando and Elena Voita. 2024. [Information flow routes: Automatically interpreting language models at scale](#). *Preprint*, arXiv:2403.00824.
- Sheridan Feucht, Eric Todd, Byron Wallace, and David Bau. 2025. [The dual-route model of induction](#). *Preprint*, arXiv:2504.03022.
- Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. [Causal analysis of syntactic agreement mechanisms in neural language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on*

- Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.
- Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. 2026. [Natural language autoencoders produce unsupervised explanations of llm activations](#). *Transformer Circuits Thread*.
- Michael Freenor and Lauren Alvarez. 2026. [Mapping semantic & syntactic relationships with geometric rotation](#). *Preprint*, arXiv:2510.09790.
- Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. 2025. [Increasing alignment of large language models with language processing in the human brain](#). *Nature computational science*, pages 1–11.
- Yutong Gao, Qinglin Meng, Yuan Zhou, and Liangming Pan. 2026. [Towards intrinsic interpretability of large language models: a survey of design principles and architectures](#). *Preprint*, arXiv:2604.16042.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. [SyntaxGym: An online platform for targeted evaluation of language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.
- Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. [Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. [Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.
- Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, et al. 2025. [Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models](#). *Nature communications*, 16(1):10529.
- Nora Graichen, Iria de Dios-Flores, and Gemma Boleda. 2026. [The grammar of transformers: A systematic review of interpretability research on syntactic knowledge in language models](#). *Preprint*, arXiv:2601.19926.
- Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. [Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms](#). *Preprint*, arXiv:2410.15956.
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. [The curious decline of linguistic diversity: Training language models on synthetic text](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.
- Rohan Gupta, Iván Arcuschin, Thomas Kwa, and Adrià Garriga-Alonso. 2025. [Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques](#). *Preprint*, arXiv:2407.14494.
- Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. 2026. [Mixing mechanisms: How language models retrieve bound entities in-context](#). In *The Fourteenth International Conference on Learning Representations*.
- Daniil Gurgurov, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2026. [Sparse subnetwork enhancement for underrepresented languages in large language models](#). *Preprint*, arXiv:2510.13580.
- Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025. [Language arithmetics: Towards systematic language neuron identification and manipulation](#). *Preprint*, arXiv:2507.22608.
- Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. [Universal neurons in GPT2 language models](#). *Transactions on Machine Learning Research*.
- Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. [Finding neurons in a haystack: Case studies with sparse probing](#). *Preprint*, arXiv:2305.01610.
- Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, Chris Olah, and Jack Lindsey. 2026. [Verbalizable representations form a global workspace in language models](#). *Transformer Circuits Thread*. Anthropic Interpretability Research.
- Michael Hanna and Aaron Mueller. 2025. [Incremental sentence processing mechanisms in autoregressive transformer language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3181–3203, Albuquerque, New Mexico. Association for Computational Linguistics.

- Sophie Hao and Tal Linzen. 2023. [Verb conjugation in transformers is determined by linear encodings of subject number](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4531–4539, Singapore. Association for Computational Linguistics.
- Abir Harrasse, Florent Draye, Punya Syon Pandey, Zhi-jing Jin, and Bernhard Schölkopf. 2026. [Tracing multilingual representations in llms with cross-layer transcoders](#). *Preprint*, arXiv:2511.10840.
- Linyang He, Peili Chen, Ercong Nie, Yuaning Li, and Jonathan R. Brennan. 2024a. [Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs](#). *Preprint*, arXiv:2403.17299.
- Zhengfu He, Xuyang Ge, Qiong Tang, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024b. [Dictionary learning improves patch-free circuit discovery in mechanistic interpretability: A case study on othello-gpt](#). *Preprint*, arXiv:2402.12201.
- Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, and Mengnan Du. 2025. [Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models](#). *Preprint*, arXiv:2502.11356.
- Stefan Heimersheim and Neel Nanda. 2024. [How to use and interpret activation patching](#). *Preprint*, arXiv:2404.15255.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. [Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training](#). *Neurobiology of Language*, 5(1):43–63.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. [A systematic assessment of syntactic generalization in neural language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.
- Jennifer Hu and Roger P. Levy. 2023. [Prompting is not a substitute for probability measurements in large language models](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. [Between circuits and](#)
- Chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709, Vienna, Austria. Association for Computational Linguistics.
- Jing Huang, Atticus Geiger, Karel D’Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. [Rigorously assessing natural language explanations of neurons](#). *Preprint*, arXiv:2309.10312.
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. [Sparse autoencoders find highly interpretable features in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. [Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 6801–6816. Association for Computational Linguistics.
- Amna Iqbal, Muhammad Younas, Saman Iftikhar, Faekeha Fatima, and Rabia Saleem. 2025. [Spam detection using hybrid model on fusion of spammer behavior and linguistics features](#). *Egyptian Informatics Journal*, 29:100605.
- Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. [LinguaLens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28232–28251, Suzhou, China. Association for Computational Linguistics.
- Thomas Jiralerspong and Trenton Bricken. 2026. [Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between llms](#). *Preprint*, arXiv:2602.11729.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. [Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *Preprint*, arXiv:2504.02768.
- Subhash Kantamneni, Joshua Engels, Senthoooran Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. [Are sparse autoencoders useful? a case study in sparse probing](#). *Preprint*, arXiv:2502.16681.
- Sheer Karny, Anthony Baez, and Pat Pataranutaporn. 2025. [Neural transparency: Mechanistic interpretability interfaces for anticipating model behaviors for personalized ai](#). *Preprint*, arXiv:2511.00230.
- Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. 2025. [Saebench:](#)

- A comprehensive benchmark for sparse autoencoders in language model interpretability. *Preprint*, arXiv:2503.09532.
- Ihor Kendiukhov. 2025. A review of developmental interpretability in large language models. *Preprint*, arXiv:2508.15841.
- Dmitrii Kharlapenko, Stepan Shabalin, Fazl Barez, Arthur Conmy, and Neel Nanda. 2025. Scaling sparse feature circuit finding for in-context learning. *Preprint*, arXiv:2504.13756.
- Hassane Kissane, Achim Schilling, and Patrick Krauss. 2025. Probing internal representations of multiword verbs in large language models. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 7–13, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.
- Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *Preprint*, arXiv:2404.02431.
- Daria Kryvosheieva, Andrea Gregor de Varda, Evelina Fedorenko, and Greta Tuckute. 2026. Different types of syntactic agreement recruit the same units within large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–227, San Diego, California, United States. Association for Computational Linguistics.
- Daria Kryvosheieva and Roger Levy. 2024. Controlled evaluation of syntactic knowledge in multilingual language models. *Preprint*, arXiv:2411.07474.
- Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. 2024. Shared functional specialization in transformer-based language models and the human brain. *Nature communications*, 15(1):5523.
- Ryoma Kumon and Hitomi Yanaka. 2026. Fine-grained analysis of shared syntactic mechanisms in language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44866–44891, San Diego, California, United States. Association for Computational Linguistics.
- Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.
- Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *Preprint*, arXiv:2309.06085.
- Xiaochen Li, Zheng-Xin Yong, and Stephen H. Bach. 2024. Preference tuning for toxicity mitigation generalizes across languages. *Preprint*, arXiv:2406.16235.
- Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.
- Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *Preprint*, arXiv:2307.09458.
- Johnny Lin. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. Software available from neuronpedia.org.
- Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Anthropic Technical Report*.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Wei Liu, Ming Xiang, and Nai Ding. 2025a. Active use of latent tree-structured sentence representation in humans and large language models. *Nature Human Behaviour*, pages 1–14.
- Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024. Unraveling Babel: Exploring multilingual activation patterns of LLMs and their applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11855–11881, Miami, Florida, USA. Association for Computational Linguistics.
- Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du, Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and Xiaoqi Jia. 2026. Focusing on language: Revealing and exploiting language attention heads in multilingual large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32195–32203.
- Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu.

- 2025b. [A systematic assessment of language models with linguistic minimal pairs in chinese](#). *Preprint*, arXiv:2411.06096.
- Jim Maar, Denis Paperno, Callum Stuart McDougall, and Neel Nanda. 2026. [What’s the plan? metrics for implicit planning in llms and their application to rhyme generation and question answering](#). *Preprint*, arXiv:2601.20164.
- Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. [Is this the subspace you are looking for? an interpretability illusion for subspace activation patching](#). In *The Twelfth International Conference on Learning Representations*.
- Christopher D Manning, Kevin Clark, John Hewitt, Urvasi Khandelwal, and Omer Levy. 2020. [Emergent linguistic structure in artificial neural networks trained by self-supervision](#). *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.
- Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. [Sparse feature circuits: Discovering and editing interpretable causal graphs in language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Josh McGiff, Tung Khanh Tran, William Mulcahy, Dáibhidh Ó Luínín, Jake Dalzell, Róisín Ní Bhroin, Adam Burke, Barry O’Sullivan, Hoang D. Nguyen, and Nikola S. Nikolov. 2026. [Irish-blimp: A linguistic benchmark for evaluating human and language model performance in a low-resource setting](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4575–4586, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. 2025. [Everything, everywhere, all at once: Is mechanistic interpretability identifiable?](#) In *The Thirteenth International Conference on Learning Representations*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023a. [Locating and editing factual associations in gpt](#). *Preprint*, arXiv:2202.05262.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023b. [Mass-editing memory in a transformer](#). *Preprint*, arXiv:2210.07229.
- Wei Han Meng, Hongzhu Guo, Yi Jing, Dewen Liu, Zijun Yao, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2026. [Saevalizer: Generating explanations for sparse autoencoder features via representation verbalization](#). *Preprint*, arXiv:2608.13538.
- Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. [Circuit component reuse across tasks in transformer language models](#). In *The Twelfth International Conference on Learning Representations*.
- Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. [COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.
- Hosein Mohebbi, Jaap Jumelet, Michael Hanna, Afra Alishahi, and Willem Zuidema. 2024. [Transformer-specific interpretability](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 21–26, St. Julian’s, Malta. Association for Computational Linguistics.
- Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhanian, and Preethi Jyothi. 2025. [Language-specific neurons do not facilitate cross-lingual transfer](#). In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.
- Philipp Mondorf, Mingyang Wang, Sebastian Gerstner, Ahmad Dawar Hakimi, Yihong Liu, Leonor Veloso, Shijia Zhou, Hinrich Schuetze, and Barbara Plank. 2025. [BlackboxNLP-2025 MIB shared task: Exploring ensemble strategies for circuit localization methods](#). In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 537–542, Suzhou, China. Association for Computational Linguistics.
- Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2026. [The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis](#). *Computational Linguistics*, 52(1):331–378.
- Aaron Mueller, Atticus Geiger, Sarah Wiegrefe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, and 4 others. 2025. [Mib: A mechanistic interpretability benchmark](#). *Preprint*, arXiv:2504.13151.
- Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. [Cross-linguistic syntactic evaluation of word prediction models](#). *Preprint*, arXiv:2005.00187.
- Aaron Mueller, Yu Xia, and Tal Linzen. 2022. [Causal analysis of syntactic agreement neurons in multilingual language models](#). *Preprint*, arXiv:2210.14328.
- Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. [First align, then predict: Understanding the cross-lingual ability of multilingual BERT](#). In *Proceedings of the 16th Conference of the*

- European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.
- Neel Nanda. 2023. [Attribution patching: Activation patching at industrial scale](#). *Neel Nanda’s Blog*.
- Neel Nanda and Joseph Bloom. 2022. Transformerlens. <https://github.com/TransformerLensOrg/TransformerLens>.
- Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2025. [Mechanistic understanding and mitigation of language confusion in English-centric large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 690–706, Suzhou, China. Association for Computational Linguistics.
- Nostalgebraist. 2020. [Interpreting gpt: The logit lens](#). LessWrong blog.
- C. Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. [Zoom in: An introduction to circuits](#). volume 5.
- Pierre Orhan, Pablo Diego-Simón, Emmmanuel Chemla, Yair Lakretz, Yves Boubenec, and Jean-Rémi King. 2026. [Emergence of phonemic, syntactic, and semantic representations in artificial neural networks](#). *Preprint*, arXiv:2601.18617.
- Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. [How do llms acquire new knowledge? a knowledge circuits perspective on continual pre-training](#). *Preprint*, arXiv:2502.11196.
- Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. 2023. [Future lens: Anticipating subsequent tokens from a single hidden state](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, page 548–560. Association for Computational Linguistics.
- Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. [Task-specific skill localization in fine-tuned language models](#). *Preprint*, arXiv:2302.06600.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Daniele Poterì, Andrea Seveso, and Fabio Mercorio. 2025. [Can role vectors affect LLM behaviour?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17735–17747, Suzhou, China. Association for Computational Linguistics.
- Daking Rai, Mor Geva, and Ziyu Yao. 2026. [Data-driven circuit discovery for interpretability of language models](#). *Preprint*, arXiv:2605.09129.
- Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2025. [A practical review of mechanistic interpretability for transformer-based language models](#). *Preprint*, arXiv:2407.02646.
- Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. [Explainability and interpretability of multilingual large language models: A survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20454–20486, Suzhou, China. Association for Computational Linguistics.
- Frederick Riemenschneider and Anette Frank. 2025. [Cross-lingual generalization and compression: From language-specific to shared neurons](#). *Preprint*, arXiv:2506.01629.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Juan Diego Rodriguez, Aaron Mueller, and Kanishka Misra. 2025. [Characterizing the role of similarity in the property inferences of language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11515–11533, Albuquerque, New Mexico. Association for Computational Linguistics.
- Andrew Rufail, Sanjana Rathore, Daniel Son, Adrian Simon, Soham Dave, Daniel Zhang, Cole Blondin, Sean O’Brien, and Kevin Zhu. 2025. [Semantic convergence: Investigating shared representations across scaled LLMs](#). In *ACL 2025 Student Research Workshop*.
- Elena Sofia Ruzzetti, Fabio Massimo Zanzotto, and Tommaso Caselli. 2026. [Lexical popularity: Quantifying the impact of pre-training for LLM performance](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1209–1230, Rabat, Morocco. Association for Computational Linguistics.
- Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. [Neuron-level interpretation of deep NLP models: A survey](#). *Transactions of the Association for Computational Linguistics*, 10:1285–1303.
- Laura Ying Schulz, Daniel Mitropolsky, and Tomaso Poggio. 2026. [Unraveling syntax: Language modeling and the substructure of grammars](#). In *Forty-third International Conference on Machine Learning*.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do multilingual LLMs think in english?](#) In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

- Or David Shafran, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. 2026. [From directions to regions: Decomposing activations in language models via local geometry](#). In *Forty-third International Conference on Machine Learning*.
- Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, and 10 others. 2025. [Open problems in mechanistic interpretability](#). *Preprint*, arXiv:2501.16496.
- Bangzhao Shu, Arinjay Singh, and Mai ElSherief. 2026. [Why are some emotions harder for llms? uncovering the causal mechanisms of emotion inference via sparse autoencoders](#). *Preprint*, arXiv:2604.25866.
- Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. [A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models](#). *Preprint*, arXiv:2503.05613.
- Aaditya K. Singh, Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. 2023. [The transient nature of emergent in-context learning in transformers](#). *Preprint*, arXiv:2311.08360.
- Taiga Someya and Yohei Oseki. 2023. [JBLiMP: Japanese benchmark of linguistic minimal pairs](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.
- Daniel Son, Sanjana Rathore, Andrew Rufail, Adrian Simon, Daniel Zhang, Soham Dave, Cole Blondin, Kevin Zhu, and Sean O’Brien. 2025. [Semantic convergence: Investigating shared representations across scaled llms](#). *Preprint*, arXiv:2507.22918.
- Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. [SLING: Sino linguistic evaluation of large language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Karolina Stańczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. [Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models](#). *Preprint*, arXiv:2205.02023.
- Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R. Bowman. 2024. [Steering without side effects: Improving post-deployment control of language models](#). *Preprint*, arXiv:2406.15518.
- Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. [Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation](#). *Computational Linguistics*, pages 1–35.
- Aaquib Syed, Can Rager, and Arthur Conmy. 2024. [Attribution patching outperforms automated circuit discovery](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.
- Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. [RuBLiMP: Russian benchmark of linguistic minimal pairs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics.
- Shaomu Tan, Di Wu, and Christof Monz. 2024. [Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [Language-specific neurons: The key to multilingual capabilities in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.
- Yiming Tang, Arash Lagzian, Srinivas Anumasa, Qiran Zou, Yingtao Zhu, Ye Zhang, Trang Nguyen, Yih-Chung Tham, Ehsan Adeli, Ching-Yu Cheng, Yilun Du, and Dianbo Liu. 2026a. [Human-like content analysis for generative ai with language-grounded sparse encoders](#). *Preprint*, arXiv:2508.18236.
- Yiming Tang, Qinglin Qi, Zhaoqian Yao, Harshvardhan Saini, and Dianbo Liu. 2026b. [Emergent hierarchical monosemantic neurons from the group-contrastive forward-forward algorithm](#). *Preprint*, arXiv:2607.16295.
- Yiming Tang, Harshvardhan Saini, Zhaoqian Yao, Zheng Lin, Yizhen Liao, Jingyi Cui, Yisen Wang, Mengnan Du, and Dianbo Liu. 2026c. [A unified theory of sparse dictionary learning in mechanistic interpretability: Piecewise biconvexity and spurious minima](#). *Preprint*, arXiv:2512.05534.
- Hinata Tezuka and Naoya Inoue. 2025. [The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms](#). *Preprint*, arXiv:2509.17030.
- Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. [Function vectors in large language models](#). In *The Twelfth International Conference on Learning Representations*.

- Katharina Trinley, Jesujoba O. Alabi, Dietrich Klakow, and Vagrant Gautam. 2026. [A mechanistic understanding of pronoun fidelity in llms](#). *Preprint*, arXiv:2606.16407.
- Igor Tufanov, Karen Hambardzumyan, Javier Ferrando, and Elena Voita. 2024. [Lm transparency tool: Interactive tool for analyzing transformer language models](#). *Preprint*, arXiv:2404.07004.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. [Steering language models with activation engineering](#). *Preprint*, arXiv:2308.10248.
- Aastha A K Verma, Anwoy Chatterjee, Mehak Gupta, and Tanmoy Chakraborty. 2026. [Multilingual language models encode script over linguistic structure](#). *Preprint*, arXiv:2604.05090.
- Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. [Holmes \$\Sigma\$ a benchmark to assess the linguistic competence of language models](#). *Transactions of the Association for Computational Linguistics*, 12:1616–1647.
- Dingdong WANG, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen M. Meng. 2026. [MMSU: A massive multi-task spoken language understanding and reasoning benchmark](#). In *The Fourteenth International Conference on Learning Representations*.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. [Interpretability in the wild: a circuit for indirect object identification in gpt-2 small](#). *Preprint*, arXiv:2211.00593.
- Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, and Ningyu Zhang. 2025a. [Beyond prompt engineering: Robust behavior control in llms via steering target atoms](#). *Preprint*, arXiv:2505.20322.
- Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025b. [Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. [Frequency effects on syntactic rule learning in transformers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do llamas work in English? on the latent language of multilingual transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Michael Wilson, Jackson Petty, and Robert Frank. 2023. [How abstract is linguistic generalization in large language models? experiments with argument structure](#). *Transactions of the Association for Computational Linguistics*, 11:1377–1395.
- Shijie Wu and Mark Dredze. 2019. [Beto, bentz, becas: The surprising cross-lingual effectiveness of bert](#). *Preprint*, arXiv:1904.09077.
- Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025a. [The semantic hub hypothesis: Language models share semantic representations across languages and modalities](#). *Preprint*, arXiv:2411.04986.
- Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025b. [Axbench: Steering llms? even simple baselines outperform sparse autoencoders](#). *Preprint*, arXiv:2501.17148.
- Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024. [ReFT: Representation fine-tuning for language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. [CLiMP: A benchmark for Chinese language model evaluation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.
- Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. [Importance-based neuron allocation for multilingual neural machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5725–5737, Online. Association for Computational Linguistics.
- Zhenxiong Yan, Suhang Yao, Yu Liu, and Wenqiang Jin. 2026. [Circuitprint: Mechanistic circuit fingerprints for large language models](#). In *Forty-third International Conference on Machine Learning*.
- Zeping Yu and Sophia Ananiadou. 2024. [Neuron-level knowledge attribution in large language models](#). In *Proceedings of the 2024 Conference on Empirical*

- Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.
- Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. [Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.
- Crystina Zhang, Jing Lu, Vinh Q. Tran, Tal Schuster, Donald Metzler, and Jimmy Lin. 2025a. [Tomato, tomahto, tomate: Do multilingual language models understand based on subword-level semantic concepts?](#) In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1821–1837, Albuquerque, New Mexico. Association for Computational Linguistics.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *The Twelfth International Conference on Learning Representations*.
- Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Yiyao Yu, Feng Yao, Renliang Sun, Yujiu Yang, and Furu Wei. 2025b. [ShifCon: Enhancing non-dominant language capabilities with a shift-based multilingual contrastive framework](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4818–4841, Vienna, Austria. Association for Computational Linguistics.
- Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Ercong Nie, Xufeng Duan, Feijiang Han, Qibo Xue, Zeping Yu, Chenming Shang, Xiao Liang, Jing Xiong, Hui Shen, Chaofan Tao, Zhengwu Liu, Senjie Jin, and 10 others. 2026. [Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models](#). *Preprint*, arXiv:2601.14004.
- Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025c. [The same but different: Structural similarities and differences in multilingual language modeling](#). In *The Thirteenth International Conference on Learning Representations*.
- Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025d. [Multilingual knowledge editing with language-agnostic factual neurons](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xu-anjing Huang. 2024. [Unveiling linguistic regions in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand. Association for Computational Linguistics.
- Nan Zhao, Xufeng Duan, and Zhenguang G. Cai. 2025. [The missing half of language learning in current developmental language models: Exogenous and endogenous linguistic input](#). *Open Mind*, 9:1543–1549.
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. [How do large language models handle multilingualism?](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. [Beyond english-centric llms: What language do multilingual language models think in?](#) *Preprint*, arXiv:2408.10811.
- Chengzhi Zhong, Fei Cheng, Qianying Liu, Yugo Murawaki, Chenhui Chu, and Sadao Kurohashi. 2026. [Language lives in sparse dimensions: Toward interpretable and efficient multilingual control for large language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7958–7970, Rabat, Morocco. Association for Computational Linguistics.
- Ej Zhou and Suchir Salhan. 2025. [Extended abstract for “linguistic universals”: Emergent shared features in independent monolingual language models via sparse autoencoders](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 128–130, Suzhuo, China. Association for Computational Linguistics.
- Xinyu Zhou, Delong Chen, Samuel Cahyawijaya, Xufeng Duan, and Zhenguang Cai. 2025. [Linguistic minimal pairs elicit linguistic similarity in large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888, Abu Dhabi, UAE. Association for Computational Linguistics.
- Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024. [LANDeRMT: Detecting and routing language-aware neurons for selectively finetuning LLMs to machine translation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148, Bangkok, Thailand. Association for Computational Linguistics.
- István Üveges and Orsolya Ring. 2025. [Evaluating the impact of synthetic data on emotion classification: A linguistic and structural analysis](#). *Information*, 16(4).

A Appendix

A.1 Related Works

Early interpretability research on RNNs and LSTMs used behavioral diagnostics and hidden-state analysis to show that individual neurons capture long-distance syntactic dependencies (Linzen et al., 2016; Lakretz et al., 2019).

As focus shifted to transformers (e.g., BERT, GPT), studies utilized causal and correlational methods to link attention heads and hidden states to linguistic phenomena (Finlayson et al., 2021; Mueller et al., 2022; Wu and Dredze, 2019; Pires et al., 2019). These models distribute linguistic information across layers for increased compositional flexibility (Hao and Linzen, 2023), with early surveys (Sajjad et al., 2022) synthesizing these insights to establish a basis for mechanistic interpretability.

A.1.1 Benchmarks and Linguistic Probing

Targeted evaluation often compares minimally different inputs whose contrast isolates a linguistic variable. Although many widely used suites focus on morphosyntax (Warstadt et al., 2020; Someya and Oseki, 2023; Xiang et al., 2021; Song et al., 2022; Taktasheva et al., 2024), the minimal-pair paradigm is not restricted to syntax. It has also been used to test conceptual and semantic property knowledge (Misra et al., 2023), discourse coherence (Beyer et al., 2021), and inferences licensed by discourse connectives (Brubaker et al., 2026). Minimal pairs therefore provide controlled behavioral contrasts across multiple levels of linguistic analysis; by themselves, however, they do not identify the internal mechanism producing the preference. (See Table 2 for a summary).

To minimize surface heuristic reliance, more rigorous approaches include SyntaxGym (Gauthier et al., 2020) for standardized evaluation, Holmes (Waldis et al., 2024) for classifier-based hidden-state probing, and HLB (Duan et al., 2024) for assessing psycholinguistic humanlikeness. Furthermore, BHASA (Leong et al., 2023) extends these diagnostics to Southeast Asian languages.

However, because these benchmarks measure behavioral outcomes rather than causal mechanisms, they risk rewarding spurious correlations (Arora et al., 2024). This limitation drives the shift toward MI, which targets the internal computational structures that causally generate linguistic behavior.

A.1.2 Shift from Model-Agnostic to Model-Specific Interpretability

To move beyond model behavior, interpretability has shifted from model-agnostic saliency methods to transformer-specific techniques (Sajjad et al., 2022; Huang et al., 2023). By explicitly targeting structural components like attention heads, feed-forward layers, and the residual stream, these methods enable precise causal interventions and analysis of information flow (Mohebbi et al., 2024; Olah et al., 2020; Elhage et al., 2021).

This transition is bolstered by integrated toolkits such as the LM Transparency Tool (Tufanov et al., 2024), Neuronpedia (Lin, 2023), and TransformerLens (Nanda and Bloom, 2022).

B Supplementary Tables and Figures

Table 2: A comparative summary of minimal pair datasets. The list includes monolingual and multilingual benchmarks, detailing their target language, total number of samples, and the count of specific linguistic paradigms evaluated.

Dataset	Language	Size	Paradigms
BLiMP	English	67k	67
CLiMP	Chinese	16k	16
SLING	Chinese	38k	38
ZhoBLiMP	Chinese	35k	118
JBLiMP	Japanese	331	39
RuBLiMP	Russian	45k	45
BLiMP-NL	Dutch	8.4k	84
QFrBLiMP	Quebec-French	1761	20
TurBLiMP	Turkish	16k	16
UrBLiMP	Urdu	5696	19
Irish-BLiMP	Irish	1020	102
Arabic MPs	Arabic	3000	9
BLiMP-IT	Italian	2899	78
CLAMS	English, French, German, Hebrew and Russian	229.9k	7
MultiBLiMP 1.0	101 languages	128321	2
BHS	Basque, Hindi, Swahili	300	3
COMPS	English	49280	3645

Table 3: Comparison of core interpretability methods and intervention methods in LLMs.

Representative Works & Key Idea	Strengths & Limitations
Vocabulary Projection (Nostalgebraist, 2020; Geva et al., 2022; Belrose et al., 2025; Pal et al., 2023; Gurnee et al., 2026) Decodes intermediate hidden states into vocabulary tokens to reveal the model’s evolving predictions layer by layer.	Strengths: Simple and training-free; provides an intuitive layer-wise view of how token predictions evolve. Limitations: Intermediate residual representations are not calibrated for direct decoding; apparent decoded tokens do not establish causal involvement.
Patching (Meng et al., 2023a; Zhang and Nanda, 2024; Nanda, 2023; Syed et al., 2024; Ferrando and Voita, 2024; Arora et al., 2024) Identifies causally relevant components by swapping activations between inputs or using gradient-based approximations.	Strengths: Pinpoints layers or heads carrying task-relevant information; enables transplant experiments. Limitations: Sensitive to superposition and intervention design; exhaustive patching can be computationally costly.
Causal Abstraction (Rodriguez et al., 2025; Boguraev et al., 2025; Desai and Nair, 2026) Aligns a hypothesized linguistic variable with a neural subspace and tests the alignment through counterfactual interventions.	Strengths: Tests explicit high-level causal hypotheses. Limitations: Results depend on the proposed abstraction, search procedure, intervention set, and generalization controls.
Sparse Autoencoders (Elhage et al., 2022; Bricken et al., 2023; Huben et al., 2024; Cho et al., 2025; Fraser-Taliente et al., 2026) Decomposes polysemantic representations into sparse, interpretable feature directions, aiming to resolve superposition.	Strengths: Can disentangle polysemantic representations, reveal interpretable subspaces, and quantify feature entanglement. Limitations: Requires additional training or specialized modeling, and disentanglement is imperfect.
Automated Circuit Discovery (Ameisen et al., 2025; Conmy et al., 2023; Marks et al., 2025) Automatically identifies compact component- or feature-level subgraphs associated with particular model behaviors.	Strengths: Traces information flow through subgraphs and can reveal circuit reuse. Limitations: Full automation remains difficult, and scalability and robustness across settings remain challenging.
Activation Steering (Turner et al., 2024; Todd et al., 2024; Rimsky et al., 2024; Wang et al., 2025a) Manipulates model behavior through inference-time interventions on internal activations.	Strengths: Directly alters internal behavior and connects interpretability with controlled intervention. Limitations: Steering can affect unrelated behaviors, and identifying robust, selective directions remains difficult.

C Details of Findings

C.1 Emergence of Formal Linguistic Competence

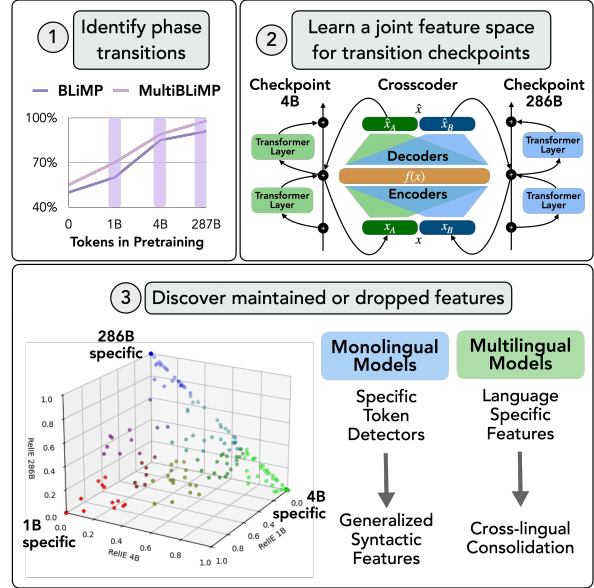


Figure 3: Overview of the analysis in Bayazit et al. (2026). The study investigates how formal linguistic competence changes during model pretraining. It first identifies performance phase transitions on benchmarks such as BLiMP across training scales, then uses Crosscoder to learn a shared feature space for selected checkpoints. The analysis suggests that some syntactic features, including language-specific features acquired early in training, become more consolidated as scale and cross-lingual exposure increase, while other features become less salient. We treat this as evidence for training-dependent representational change, not as proof of symbolic rule acquisition.

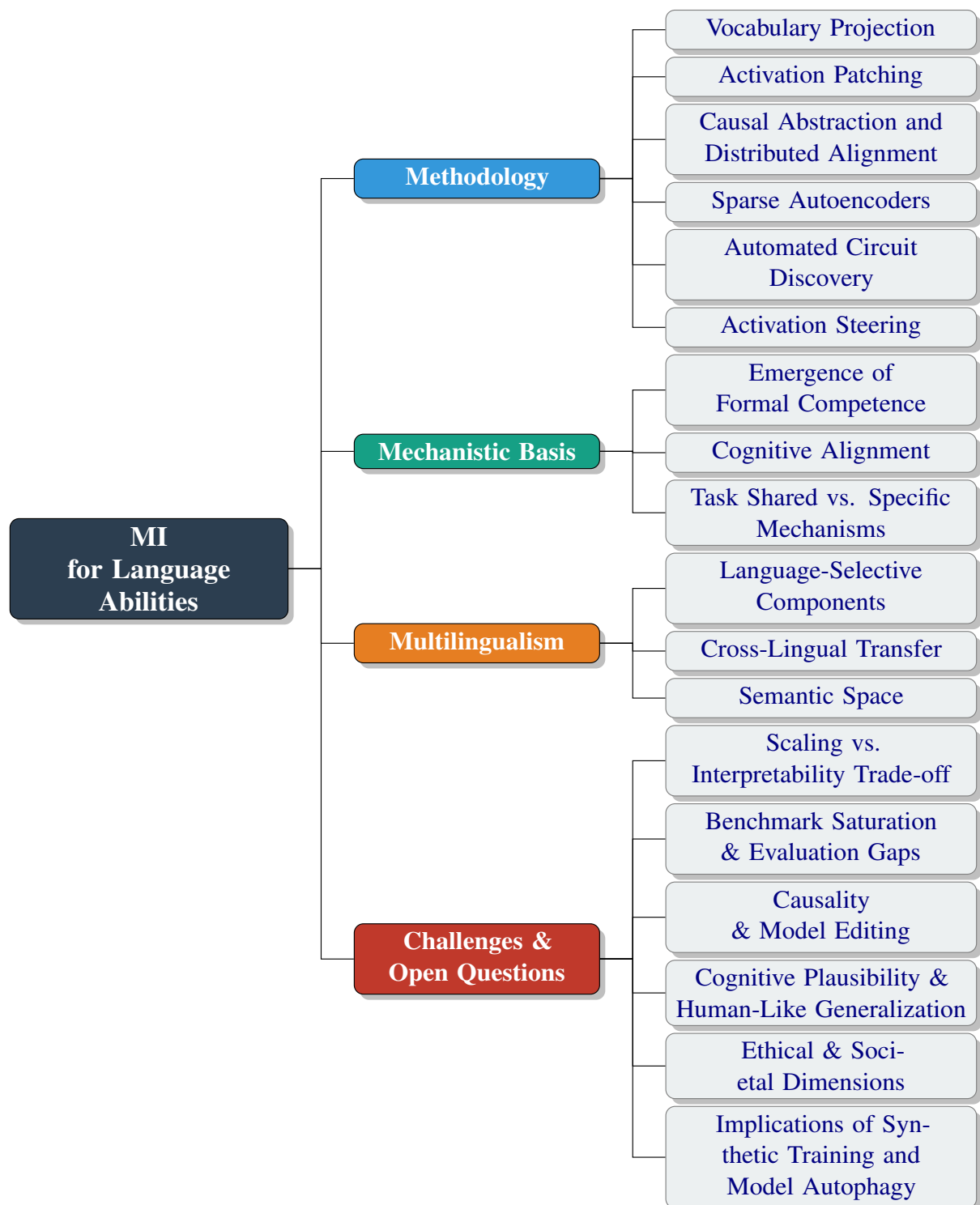


Figure 2: A taxonomy of the MI landscape for LLMs. Click on the nodes to jump to the corresponding section.

C.2 Partial Alignment with Human Language Processing

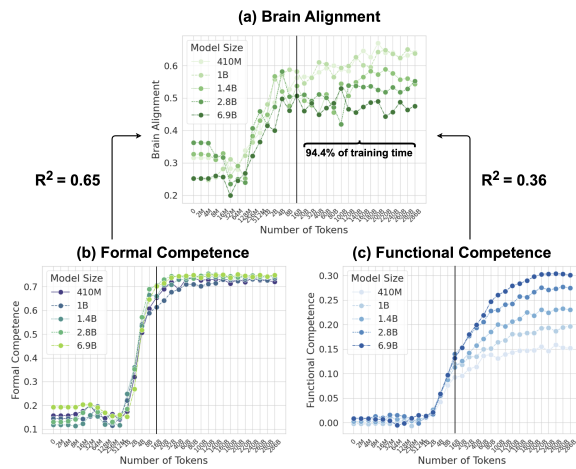


Figure 4: Overview of the analysis in AlKhamissi et al. (2025b). The study relates neural alignment to formal linguistic competence and functional abilities during training. In the reported results, brain alignment rises early and then plateaus, while functional performance follows a different trajectory. We interpret this as evidence that alignment is selective and bounded: some language-related representations become more similar to human neural responses, but this does not imply broad human-like cognition or a shared implementation of semantic and pragmatic abilities.

C.3 Task-Specific Mechanisms vs. Shared Circuits

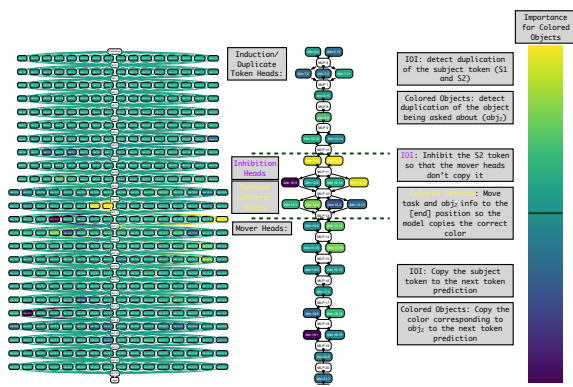


Figure 5: Overview of the circuit analysis in Merullo et al. (2024) for Colored Objects and Indirect Object Identification (IOI). The study reports overlap in heads involved in duplicate-token detection across the two tasks, while downstream heads use the duplicate signal differently. This supports a limited “shared subroutine, task-specific use” interpretation: some components can be reused across tasks with related structure, but the result should not be generalized to all linguistic phenomena or all models without further validation.

C.4 Language-Selective Components

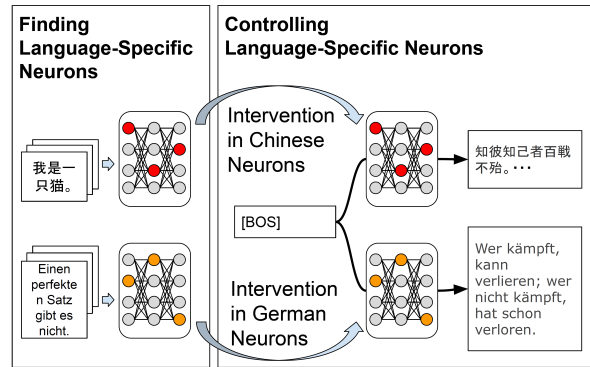


Figure 6: Overview of the language-selective neuron analysis in Kojima et al. (2024). The study identifies neurons with stronger responses to particular languages (e.g., Chinese or German) and tests whether enhancing or suppressing them changes output language probabilities. We treat these interventions as evidence for language-sensitive components, while noting that steering output language is not the same as fully explaining multilingual competence.

C.5 Cross-Lingual Transfer

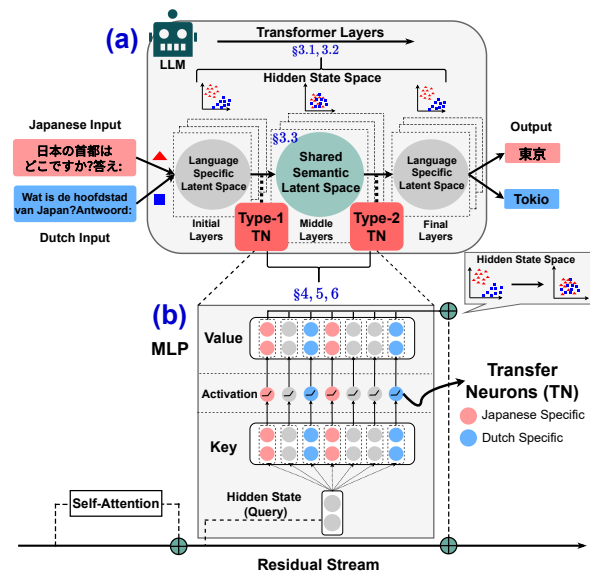


Figure 7: Overview of the Transfer Neurons Hypothesis proposed by Tezuka and Inoue (2025). The hypothesis models cross-lingual transfer as movement between language-specific latent spaces and a shared semantic latent space, mediated by specialized MLP neurons. This provides a candidate mechanism for cross-lingual routing, but its scope and causal sufficiency remain open questions across models, languages, and tasks.

C.6 Mechanisms of Semantic Alignment

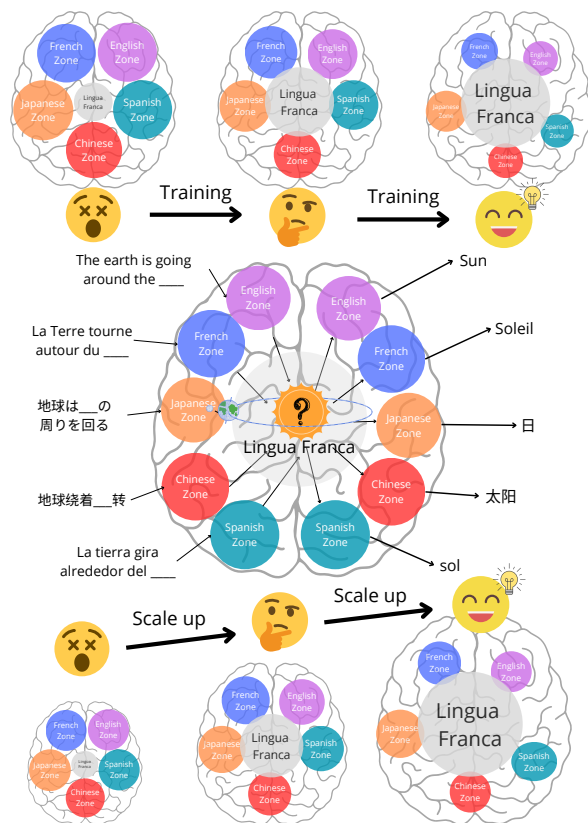


Figure 8: Overview of the “Lingua Franca” hypothesis in Zeng et al. (2025). The study proposes that multi-lingual models can map surface forms from different languages into a partially shared latent semantic space, with stronger alignment observed in larger or more extensively trained models. We present this as a useful hypothesis about semantic alignment, while leaving open how consistently it holds for low-resource languages, typologically distant languages, and non-translation tasks.