Review

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

Xufeng Duan , Zhaoqian Yao , Xinyu Zhou , Zihao Fu , Beijia Kan , Yibo Wang , Bei Xiao , Hengyuan Zhang , Xuemei Tang , Ercong Nie , Zhenguang G. Cai *

Posted Date: 20 July 2026

doi: 10.20944/preprints202607.1441.v1

Keywords: mechanistic interpretability; Large Language Models; Language Abilities

 doi:10.20944/preprints202607.1441.v1

Review

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

Xufeng Duan [1], Zhaoqian Yao [1], Xinyu Zhou [2,3], Zihao Fu [1], Beijia Kan [1], Yibo Wang [1], Bei Xiao [1], Hengyuan Zhang [4], Xuemei Tang [5], Ercong Nie [6], Zhenguang G. Cai [1,*]

[1] The Chinese University of Hong Kong
[2] Université Paris Cité
[3] Sorbonne Université
[4] University of Hong Kong
[5] Hong Kong Polytechnic University
[6] Shanghai Jiao Tong University
* Correspondence: 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, and causal 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.

Keywords: mechanistic interpretability; Large Language Models; Language Abilities

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

Mechanistic interpretability (MI) aims to clarify this distinction by decomposing a model's blackbox 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,

 doi:10.20944/preprints202607.1441.v1

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 formal semantics (e.g., argument 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.

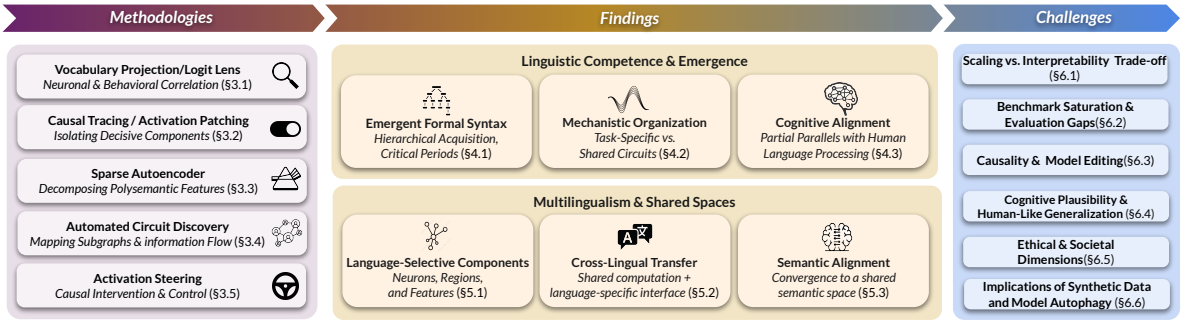

Figure 1. The overview of this survey.

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.

Table 1. Representative surveys in MI and their focus.

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

This account occupies a middle position between two extremes. While mechanistic findings rule out a pure n-gram or rote-memorization explanation by identifying explicit pathways for linguistic competence (e.g., syntax), they do not justify claims of human-like symbolic competence. The discovered mechanisms remain distributed, frequency-conditioned, and model-specific. Therefore, we treat

 doi:10.20944/preprints202607.1441.v1

mechanistic evidence as a window into how specific models implement linguistic behavior, rather than proof of 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, distinguishing 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

In this paper, we review key methods for analyzing how LLMs represent linguistic knowledge (see Table 3 for a summary and a running example for all methods in A.1.4).

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):

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

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). 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, Bau, et al., 2023; K. Wang et al., 2022; F. Zhang & Nanda, 2024). If the intervention restores the model's performance on a target task, that component is considered causally decisive (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:

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

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

 doi:10.20944/preprints202607.1441.v1

Linguistic Leverage

Patching provides direct causal evidence for linguistic tasks. Its validity depends on demonstrating generalizability across diverse lexical items, templates, and distractors to avoid identifying prompt-specific pathways.

2.3. 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; Z. He et al., 2025; Huben et al., 2024; D. Shu et al., 2025; Y. Tang et al., 2026). SAEs reconstruct an activation $\mathbf{x}$ as a linear combination of features:

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

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?

2.4. Circuit Discovery

While SAEs disentangle representations into interpretable features, these units can also be composed into higher-level computational mechanisms. Building on activation patching, **Circuit Discovery** (Mondorf et al., 2025; F. Zhang & Nanda, 2024) 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.

 doi:10.20944/preprints202607.1441.v1

2.5. Activation Steering

Methodology has expanded from observation to control, enabling direct intervention in a model's linguistic behavior. **Activation Steering** involves injecting vectors (Karny et al., 2025; Potertì et al., 2025; Todd et al., 2024; Z. Wu, Yu, 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) (M. Wang, Xu, et al., 2025) 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 (J. 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, Yao, et al. (2025) identify "critical periods" in training, specific windows where rapid internal syntactic specialization occurs. Ou et al. (2025) and Bayazit et al. (2025) 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 Y. Liu et al. (2025), 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 (Chang, Tu, & Bergen, 2024; Wei et al., 2021). 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

 doi:10.20944/preprints202607.1441.v1

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., 2025). 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, Trinley, et al., 2025; Huo et al., 2024; T. Tang et al., 2024; Z. Wu 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 (Gurnee et al., 2023; Olah et al., 2020), whereas complex reasoning requires interactions among numerous polysemantic units (Marks et al., 2025).

What We Learn

Current circuit evidence favors reusable subroutines over fully isolated linguistic modules. Duplicate-token detection, agreement routing, and mover/inhibition patterns suggest that models can reuse low-level mechanisms while attaching task-specific execution heads.

3.3. Cognitive View: Partial Alignment with Human Language Processing

Recent works (AlKhamissi, Tuckute, Tang, et al., 2025; Honarmand et al., 2025; Kumar et al., 2024) suggest that the relationship between LLMs and human language processing can be characterized as partial alignment: systematic correspondences emerge at specific representational levels, while substantial divergences persist elsewhere.

Alignment is most relevant for this survey at the level of language-selective structure. AlKhamissi, Tuckute, Bosselut, and Schrimpf (2025) identify language-selective units across a wide range of LLMs, which exhibit stronger alignment with neural activity in human language regions than randomly sampled areas (Kumar et al., 2024). At a finer grain, Duan, Zhou, et al. (2025) use theory-driven psycholinguistic paradigms to probe neuron-level mechanisms, reporting that when a model exhibits human-like behavior in tasks such as sound–gender association, a small subset of neurons can be causally identified.

At the representational level, hidden states can predict neural responses during naturalistic language comprehension (C. Gao et al., 2025; Hosseini et al., 2024). Furthermore, this alignment exhibits a layered correspondence between model depth and the temporal hierarchy of neural responses (Goldstein et al., 2025). However, this alignment is selective. Tracking models across training, AlKhamissi, Tuckute, Tang, et al. (2025) report that brain alignment correlates more strongly with formal linguistic competence than with functional abilities such as reasoning.

What We Learn

Cognitive alignment findings are suggestive but bounded. The most relevant evidence concerns formal linguistic competence and language-selective representations; broader pragmatics and human-like cognition remain less directly explained by current MI.

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

 doi:10.20944/preprints202607.1441.v1

across languages. Current evidence points to a hybrid organization of multilingual representations. Language-specific components appear to govern language identity and surface-form processing, whereas shared representations are more closely associated with transfer and semantic alignment.

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. T. 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. Z. 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, X. Liu et al. (2025) 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 alongside 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. (2025) identify monosemantic language-specific features whose ablation selectively degrades individual languages. Building on this, Zhong et al. (2025) 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. Y. 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, Y. Zhao et al. (2024) report that targeted tuning of those neurons can selectively improve a language without broadly affecting others. To map this flow to the entire network, Harrasse et al. (2025) utilize cross-layer transcoders, finding that while internal representations are largely shared, language-specific

 doi:10.20944/preprints202607.1441.v1

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. C. Zhang et al. (2025) report that multilingual embeddings cluster near-synonymous tokens across languages in overlapping regions, while W. 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

 doi:10.20944/preprints202607.1441.v1

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 polysemanticity 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 (Jing et al., 2025; Marks 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 (Arora et al., 2024; J. Hu & Levy, 2023) 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 underrepresented contexts, risking interpretability methods that overfit to narrow linguistic variations. Developing robust, culturally sensitive frameworks for low-resource languages remains a critical challenge (Chang, Arnett, et al., 2024; Zeng et al., 2025).

5.3. Causality and Model Editing

A core goal of MI is causal intervention. Techniques such as attribution patching (Ferrando & Voita, 2024) and neuron or feature steering (Banayeeanzade et al., 2025; Lindsey et al., 2025) 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, Tuckute, Tang, et al. (2025) 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; Duan, Yao, et al., 2025; Evanson et al., 2023), 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 inter-

 doi:10.20944/preprints202607.1441.v1

pretability (N. 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, Arnett, et al., 2024). 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, Arnett, et al., 2024; Üveges & 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, Bau, et al., 2023). 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 (Merullo et al., 2024; K. Wang et al., 2022). 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, Tu, & Bergen, 2024; 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 (T. Tang et al., 2024; Z. Zhang et al., 2024) and proposed shared semantic representations such as a model-internal lingua franca (Z. Wu, Yu, et al., 2025; Zeng et al., 2025).

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, Arnett, et al., 2024; 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

 doi:10.20944/preprints202607.1441.v1

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.

A. Appendix

A.1. Related Works

A.1.1. Foundational Interpretability Studies

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

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; Pires et al., 2019; S. Wu & Dredze, 2019). These models distribute linguistic information across layers for increased compositional flexibility (Hao & Linzen, 2023), with early surveys (Sajjad et al., 2021) synthesizing these insights to establish a basis for mechanistic interpretability.

A.1.2. Benchmarks and Linguistic Probing

Researchers evaluate grammatical competence using controlled benchmarks, primarily employing the minimal pairs paradigm to compare model likelihoods for grammatical vs. ungrammatical sentences (Someya & Oseki, 2023; Song et al., 2022; Taktasheva et al., 2024; Warstadt et al., 2020; Xiang et al., 2021) (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.1.3. 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 (Huang et al., 2023; Sajjad et al., 2021). 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 (Elhage et al., 2021; Mohebbi et al., 2024; Olah et al., 2020).

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

A.1.4. Running Example: Subject–Verb Agreement

Subject–verb agreement illustrates how the methods above form a cumulative evidential chain rather than separate tool descriptions. Behavioral minimal pairs first test whether a model prefers *The keys are* over *The keys is*, but such success may still reflect local frequency or lexical memorization. Vocabulary projection can then show when number-consistent verb predictions become available across layers. Activation patching tests whether specific heads, MLP activations, or residual-stream states are causally necessary for transferring number information from the subject to the verb position.

 doi:10.20944/preprints202607.1441.v1

SAEs and probes can identify candidate number features, while circuit discovery asks whether those features compose into a reusable pathway that survives attractor nouns and lexical substitutions. A strong mechanistic claim therefore requires more than one positive result: it requires causal necessity, generalization beyond templates, and evidence against simpler frequency-based explanations.

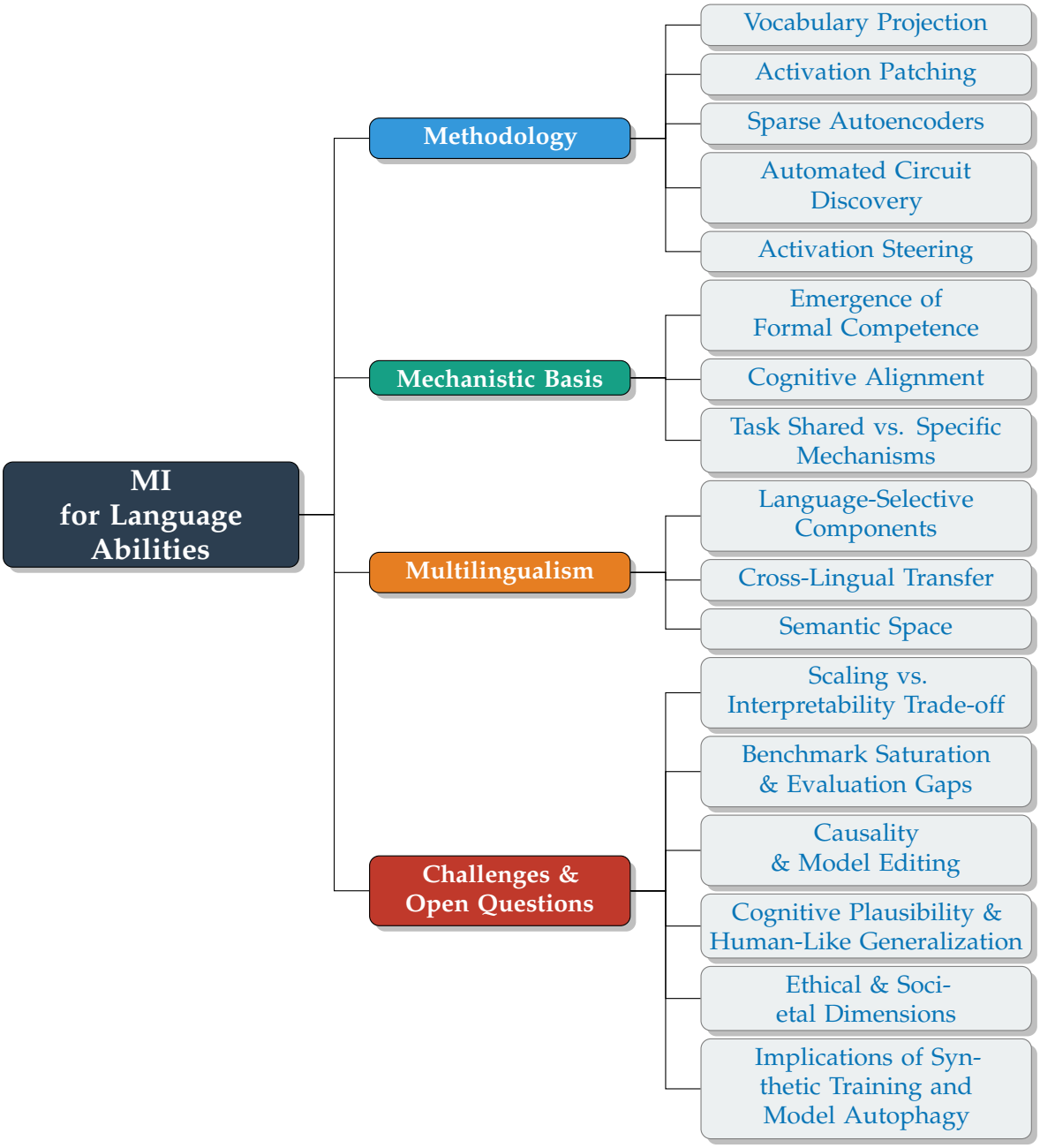

Figure 2. A taxonomy of the MI landscape for LLMs. Click on the nodes to jump to the corresponding section.

 doi:10.20944/preprints202607.1441.v1

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 Warstadt et al. (2020)	English	67k	67
CLiMP Xiang et al. (2021)	Chinese	16k	16
SLING Song et al. (2022)	Chinese	38k	38
ZhoBLiMP Y. Liu et al. (2025)	Chinese	35k	118
JBLiMP Someya and Oseki (2023)	Japanese	331	39
RuBLiMP Taktasheva et al. (2024)	Russian	45k	45
BLiMP-NL Suijkerbuijk et al. (2025)	Dutch	8.4k	84
QFrBLiMP Beauchemin et al. (2025)	Quebec-French	1761	20
TurBLiMP Başar et al. (2025)	Turkish	16k	16
UrBLiMP Adeeba et al. (2025)	Urdu	5696	19
Irish-BLiMP McGiff et al. (2025)	Irish	1020	102
Arabic MPs Alrajhi et al. (2022)	Arabic	3000	9
BLiMP-IT Barbini et al. (2025)	Italian	2899	78
CLAMS Mueller et al. (2020)	English, French, German, Hebrew and Russian	229.9k	7
MultiBLiMP 1.0 Jumelet et al. (2025)	101 languages	128321	2
BHS Kryvosheieva and Levy (2024)	Basque, Hindi, Swahili	300	3

Table 3. Comparison of core interpretability methods and intervention methods in LLMs. Each method is characterized by its representative studies, key idea, strengths and limitations.

Methodology	Representative Works	Key Idea	Strengths and Limitations
Vocabulary Projection	**Logit Lens** (Geva et al., 2022; Nostalgebraist, 2020); **Tuned Lens** (Belrose et al., 2025); **Future Lens** (Pal et al., 2023).	Decodes intermediate hidden states into vocabulary tokens to reveal the model's evolving predictions layer-by-layer.	**Strengths:** Maps word vectors across languages rapidly; no model retraining needed. **Limitations:** Ineffective for distant languages; performance depends on initial dictionary quality; struggles with new words or complex tasks.
Patching	**Activation Patching/Attribution Patching** (Arora et al., 2024; Ferrando & Voita, 2024; Meng, Bau, et al., 2023; Nanda, 2023; Syed et al., 2024; F. Zhang & Nanda, 2024)	Identifies causally decisive components by swapping activations between inputs or using gradient-based approximations.	**Strengths:** Pinpoints layers/-heads encoding key information; enables transplant experiments; shows causal roles of activations. **Limitations:** Sensitive to superposition; computationally costly.

 doi:10.20944/preprints202607.1441.v1

Methodology	Representative Works	Key Idea	Strengths and Limitations
Sparse Autoencoders	**SAE, Transcoder, Cross-layer Transcoder, Crosscoder and Binary Autoencoder** (Bricken et al., 2023; H. Cho et al., 2025; Elhage et al., 2022; Huben et al., 2024)	Decomposes polysemantic neurons into sparse, interpretable feature directions, resolving superposition.	**Strengths:** Disentangles polysemantic neurons; reveals interpretable subspaces; quantifies feature entanglement. **Limitations:** Requires training or specialized modeling; imperfect disentanglement.
Automated Circuit Discovery	**Circuit Tracing** (Ameisen et al., 2025; Conmy et al., 2023; Marks et al., 2025)	Automatically finds minimal subgraphs (component- or feature-level) that implement specific tasks.	**Strengths:** Traces information flow via minimal subgraphs; reveals emergent "circuit reuse"; mechanistically grounded. **Limitations:** Hard to automate fully; scalability remains challenging.
Activation Steering	**ActAdd, CAA, STA** (Fang et al., 2025; Meng, Bau, et al., 2023; Meng, Sharma, et al., 2023; Rimsky et al., 2024; Todd et al., 2024; Turner et al., 2024; M. Wang, Xu, et al., 2025)	Manipulates model behavior via inference-time activation injection (Steering) or permanent weight modification (Editing).	**Strengths:** Directly alters internal behavior; useful for debiasing and control; connects interpretability with intervention. **Limitations:** Can disrupt unrelated behaviors; hard to find minimal safe edits.

Table 4. Representative papers in MI. We categorize works into **Method** (foundational techniques, tools, frameworks, or benchmarks) and **Application** (studies leveraging these methods to investigate specific linguistic capabilities). The table spans five core domains: Probing, Vocabulary Projection (Logit Lens), Activation Patching, Sparse Autoencoders (SAE), and Circuit Discovery.

Paper Title	Type	Venue
Benchmarks		
MIB: A Mechanistic Interpretability Benchmark (Mueller, Geiger, et al., 2025)	Method	ICML
AXBENCH: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders (Z. Wu, Arora, et al., 2025)	Method	ICML
SAEBench: A Comprehensive Benchmark for Sparse Autoencoders (Karvonen et al., 2025)	Method	ICML
InterpBench: Semi-Synthetic Transformers for Evaluating MI Techniques (Gupta et al., 2025)	Method	NIPS
Thank You, Stingray: Multilingual Large Language Models Can Not (Yet) Disambiguate Cross-Lingual Word Senses (Cahyawijaya et al., 2025)	Method	NAACL
CausalGym: Benchmarking causal interpretability methods on linguistic tasks (Arora et al., 2024)	Method	ACL

 doi:10.20944/preprints202607.1441.v1

Paper Title	Type	Venue
Holmes: A Benchmark to Assess the Linguistic Competence of Language Models (Waldis et al., 2024)	Method	TACL
SyntaxGym: An Online Platform for Targeted Evaluation of Language Models (Gauthier et al., 2020)	Method	ACL
Probing		
Lexical Popularity: Quantifying the Impact of Pre-training for LLM Performance (Ruzzetti et al., 2026)	Method	Arxiv
Steering Embedding Models with Geometric Rotation: Mapping Semantic Relationships Across Languages and Models (Freenor & Alvarez, 2026)	Method	Arxiv
Emergence of Phonemic, Syntactic, and Semantic Representations in Artificial Neural Networks (Orhan et al., 2026)	Application	Arxiv
ShifCon: Enhancing Non-Dominant Language Capabilities with a Shift-based Multilingual Contrastive Framework (H. Zhang et al., 2025)	Method	ACL
Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models (M. Wang, Adel, et al., 2025)	Application	ACL
Probing Syntax in Large Language Models: Successes and Remaining Challenges (Diego-Simón et al., 2025)	Application	COLM
Probing Internal Representations of Multi-Word Verbs in Large Language Models (Kissane et al., 2025)	Application	MWE
Can Cross-Lingual Transferability of Multilingual Transformers Be Activated Without End-Task Data? (Z. Chi et al., 2023)	Method	ACL
Finding Neurons in a Haystack: Case Studies with Sparse Probing (Gurnee et al., 2023)	Application	Arxiv
Probing Classifiers: Promises, Shortcomings, and Advances (Belinkov, 2022)	Method	CL
First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT (Muller et al., 2021)	Application	EACL
Finding Universal Grammatical Relations in Multilingual BERT (E. A. Chi et al., 2020)	Application	ACL
On the Language Neutrality of Pre-trained Multilingual Representations (Libovický et al., 2020)	Application	EMNLP
Emergent Linguistic Structure in Artificial Neural Networks Trained by Self-supervision (Manning et al., 2020)	Application	PNAS
A Structural Probe for Finding Syntax in Word Representations (Hewitt & Manning, 2019)	Application	NAACL
Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information (Giulianelli et al., 2018)	Application	BlackboxNLP
Vocabulary Projection (Logit Lens)		
Eliciting Latent Predictions from Transformers with the Tuned Lens (Belrose et al., 2025)	Method	Arxiv
Future Lens: Anticipating Subsequent Tokens from a Single Hidden State (Pal et al., 2023)	Method	CoNLL
Interpreting GPT: the logit lens (Nostalgebraist, 2020)	Method	Blog

 doi:10.20944/preprints202607.1441.v1

Paper Title	Type	Venue
Unraveling Syntax: How Language Models Learn Context-Free Grammars (Schulz et al., 2025)	Application	Arxiv
The Semantic Hub Hypothesis: Language Models Share Semantic Representations (Z. Wu, Yu, et al., 2025)	Application	ICLR
Do Multilingual LLMs Think in English? (Schut et al., 2025)	Application	Workshop
Do Llamas Work in English? On the Latent Language of Multilingual Transformers (Wendler et al., 2024)	Application	ACL
DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models (Chuang et al., 2024)	Application	ICLR
Sparse AutoEncoders (SAE)		
Are Sparse Autoencoders Useful? A Case Study in Sparse Probing (Kantamneni et al., 2025)	Method	ICML
Binary Autoencoder for Mechanistic Interpretability of Large Language Models (H. Cho et al., 2025)	Method	Arxiv
LinguaLens: Towards Interpreting Linguistic Mechanisms via Sparse Auto-Encoder (Jing et al., 2025)	Method	EMNLP
SAIF: A Sparse Autoencoder Framework for Interpreting and Steering Instruction Following of Language Models (Z. He et al., 2025)	Method	Arxiv
A Unified Theory of Sparse Dictionary Learning in Mechanistic Interpretability: Piecewise Biconvexity and Spurious Minima (Y. Tang et al., 2026)	Method	Arxiv
The Grounding Gap: How LLMs Anchor the Meaning of Abstract Concepts Differently from Humans (Chlapanis et al., 2026)	Application	Arxiv
From Syntax to Emotion: A Mechanistic Analysis of Emotion Inference in LLMs (B. Shu et al., 2026)	Application	Arxiv
Cross-Architecture Model Diffing with Crosscoders: Unsupervised Discovery of Differences Between LLMs (Jiralerspong & Bricken, 2026)	Application	Arxiv
What's the plan? Metrics for implicit planning in LLMs and their application to rhyme generation and question answering (Maar et al., 2026)	Application	ICLR
Crosscoding Through Time: Tracking Emergence of Linguistic Representations (Bayazit et al., 2025)	Application	ICML
Large Language Models Share Representations of Latent Grammatical Concepts (Brinkmann et al., 2025)	Application	NAACL
Incremental Sentence Processing Mechanisms in Autoregressive Transformer Language Models (Hanna & Mueller, 2025)	Application	NAACL
Sparse Autoencoders Can Capture Language-Specific Concepts Across Diverse Languages (Andrylie et al., 2025)	Application	Arxiv
Unveiling Language-Specific Features via Sparse Autoencoders (Deng et al., 2025)	Application	ACL
Analyzing Multilingualism in Large Language Models with Sparse Autoencoders (I. Cho & Hockenmaier, 2025)	Application	COLM
Sparse Autoencoders Can Capture Language-Specific Concepts (Andrylie et al., 2025)	Application	Arxiv
Tracing Multilingual Representations in LLMs with Cross-Layer Transcoders (Harrasse et al., 2025)	Application	Arxiv

 doi:10.20944/preprints202607.1441.v1

Paper Title	Type	Venue
Causal Language Control in Multilingual Transformers via Sparse Feature Steering (Chou et al., 2025)	Application	Arxiv
Semantic Convergence: Investigating Shared Representations Across Scaled LLMs (Rufail et al., 2025)	Application	SRW
Extended Abstract for "Linguistic Universals": Emergent Shared Features in Independent Monolingual Language Models via Sparse Autoencoders (E. Zhou & Salhan, 2025)	Application	MRL
Sparse Autoencoders Find Highly Interpretable Features in Language Models (Huben et al., 2024)	Method	ICLR
Transcoders Find Interpretable LLM Feature Circuits (Dunefsky et al., 2024)	Method	NIPS
Activation Patching		
Mixing Mechanisms: How Language Models Retrieve Bound Entities In-Context (Gur-Arieh et al., 2025)	Method	Arxiv
How to use and interpret activation patching (Heimersheim & Nanda, 2024)	Method	Arxiv
Is This the Subspace You Are Looking for? An Interpretability Illusion (Makelov et al., 2024)	Method	ICLR
Towards Best Practices of Activation Patching in Language Models (F. Zhang & Nanda, 2024)	Method	ICLR
Function Vectors in Large Language Models (Todd et al., 2024)	Method	ICLR
Attribution Patching Outperforms Automated Circuit Discovery (Syed et al., 2024)	Method	Blackbox
Attribution Patching: Activation Patching At Industrial Scale (Nanda, 2023)	Method	BLOG
Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models (Finlayson et al., 2021)	Method	IJCNLP
The Dual-Route Model of Induction (Feucht et al., 2025)	Application	COLM
Neuron Analysis		
From Directions to Regions: Decomposing Activations in Language Models via Local Geometry(Shafran et al., 2026)	Method	Arxiv
LANDeRMT: Dectecting and Routing Language-Aware Neurons for Selectively Finetuning LLMs to Machine Translation (Zhu et al., 2024)	Method	ACL
Task-Specific Skill Localization in Fine-tuned Language Models (Panigrahi et al., 2023)	Method	ICML
Language Arithmetics: Towards Systematic Language Neuron Identification and Manipulation (Gurgurov, Trinley, et al., 2025)	Application	AACL
Cross-Lingual Generalization and Compression (Riemenschneider & Frank, 2025)	Application	ACL
How Syntax Specialization Emerges in Language Models (Duan, Yao, et al., 2025)	Application	Arxiv
Language Lives in Sparse Dimensions: Interpretable Multilingual Control (Zhong et al., 2025)	Application	Arxiv
Inducing Dyslexia in Vision Language Models (Honarmand et al., 2025)	Application	Arxiv

 doi:10.20944/preprints202607.1441.v1

Paper Title	Type	Venue
Different types of syntactic agreement recruit the same units (Kryvosheieva et al., 2025)	Application	Arxiv
Sparse Subnetwork Enhancement for Underrepresented Languages in Large Language Models (Gurgurov, van Genabith, & Ostermann, 2025)	Application	Arxiv
Multilingual Knowledge Editing with Language-Agnostic Factual Neurons (X. Zhang et al., 2025)	Application	COLING
From Language to Cognition: How LLMs Outgrow the Human Language Network (AlKhamissi, Tuckute, Tang, et al., 2025)	Application	EMNLP
The Transfer Neurons Hypothesis (Tezuka & Inoue, 2025)	Application	EMNLP
Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models (Nie et al., 2025)	Application	EMNLP
The Rise and Down of Babel Tower: Investigating the Evolution Process of Multilingual Code Large Language Model (Chen et al., 2025)	Application	ICLR
The LLM Language Network: A Neuroscientific Approach (AlKhamissi, Tuckute, Bosselut, & Schrimpf, 2025)	Application	NAACL
Language-Specific Neurons Do Not Facilitate Cross-Lingual Transfer (Mondal et al., 2025)	Application	Workshop
Language-Specific Neurons: The Key to Multilingual Capabilities (T. Tang et al., 2024)	Application	ACL
Unveiling Linguistic Regions in Large Language Models (Z. Zhang et al., 2024)	Application	ACL
Converging to a Lingua Franca: Evolution of Linguistic Regions (Zeng et al., 2025)	Application	COLING
Unveiling Language Competence Neurons: A Psycholinguistic Approach (Duan, Zhou, et al., 2025)	Application	COLING
Linguistic Minimal Pairs Elicit Linguistic Similarity (X. Zhou et al., 2025)	Application	COLING
Neuron-Level Knowledge Attribution in Large Language Models (Yu & Ananiadou, 2024)	Application	EMNLP
Neuron Specialization: Leveraging Intrinsic Task Modularity for Multilingual Machine Translation (Tan et al., 2024)	Application	EMNLP
Decoding Probing: Revealing Internal Linguistic Structures (L. He et al., 2024)	Application	LREC
On the Multilingual Ability: Finding and Controlling Language-Specific Neurons (Kojima et al., 2024)	Application	NAACL
How do Large Language Models Handle Multilingualism? (Y. Zhao et al., 2024)	Application	NIPS
Universal Neurons in GPT2 Language Models (Gurnee et al., 2024)	Application	TMLR
Same Neurons, Different Languages (Stańczak et al., 2022)	Application	NAACL
Importance-based Neuron Allocation for Multilingual Neural Machine Translation (Xie et al., 2021)	Application	ACL
Circuit Discovery		
Circuit Tracing: Revealing Computational Graphs in Language Models (Ameisen et al., 2025)	Method	Anthropic
Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs (Marks et al., 2025)	Method	ICLR

 doi:10.20944/preprints202607.1441.v1

Paper Title	Type	Venue
Scaling Sparse Feature Circuits For Studying In-Context Learning (Kharlapenko et al., 2025)	Method	ICML
Dictionary Learning Improves Patch-Free Circuit Discovery in Mechanistic Interpretability: A Case Study on Othello-GPT (Z. He et al., 2024)	Method	Arxiv
Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small (K. Wang et al., 2022)	Method	ICLR
Towards Automated Circuit Discovery for Mechanistic Interpretability (Conmy et al., 2023)	Method	NIPS
Between Circuits and Chomsky: Pre-pretraining on Formal Languages (M. Y. Hu et al., 2025)	Application	ACL
The Same But Different: Structural Similarities in Multilingual LM (R. Zhang et al., 2025)	Application	ICLR
Circuit Component Reuse Across Tasks in Transformer Language Models (Merullo et al., 2024)	Application	ICLR

B. Details of Findings

B.1. Emergence of Formal Linguistic Competence

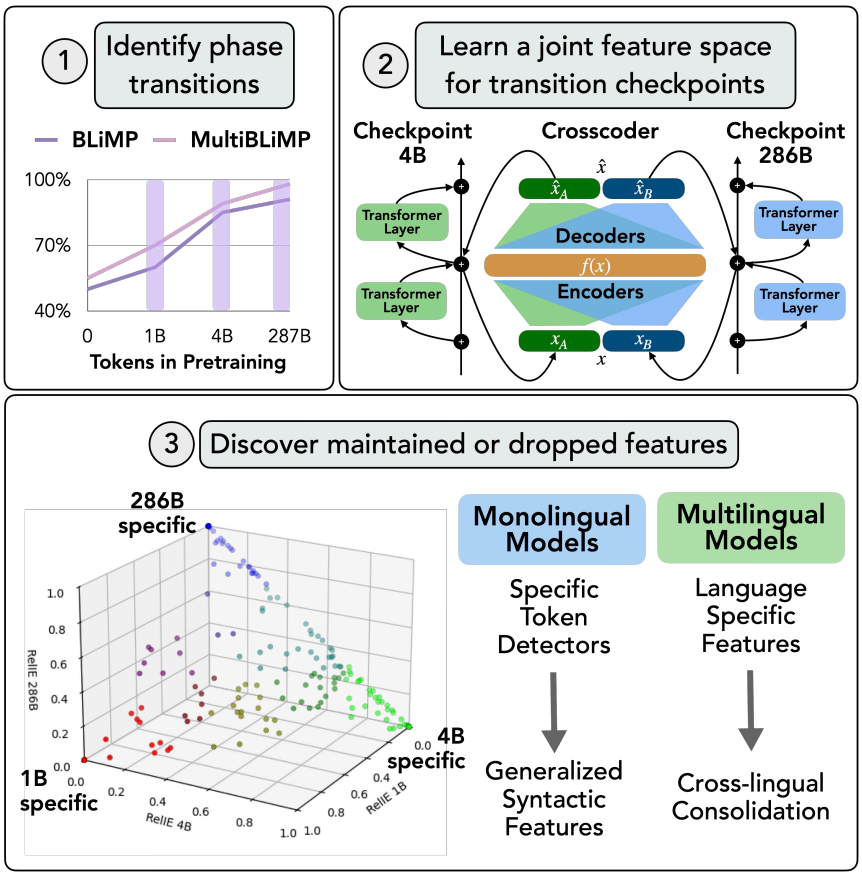

Figure 3. Overview of the analysis in Bayazit et al. (2025). The study investigates how formal linguistic competence changes during model pretraining. It first identifies performance phase transitions on benchmarks such as BLiMP across training scales, then uses Crosscoder to learn a shared feature space 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.

 doi:10.20944/preprints202607.1441.v1

B.2. Partial Alignment with Human Language Processing

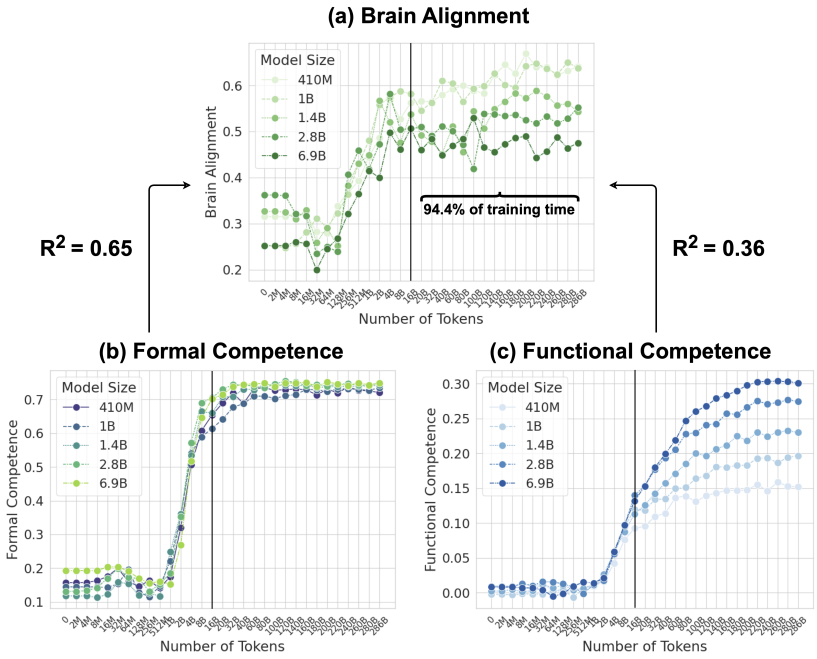

Figure 4. Overview of the analysis in AlKhamissi, Tuckute, Tang, et al. (2025). The study relates neural alignment to formal linguistic competence and functional abilities during training. In the reported results, brain alignment rises early and then plateaus, while functional performance follows a different 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.

B.3. Task-Specific Mechanisms vs. Shared Circuits

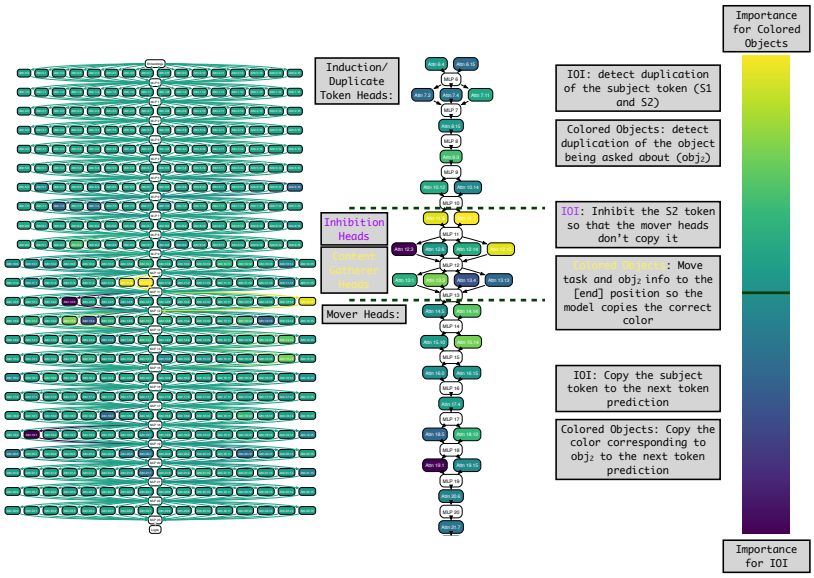

Figure 5. Overview of the circuit analysis in Merullo et al. (2024) for Colored Objects and Indirect Object Identification (IOI). The study reports overlap in heads involved in duplicate-token detection across the two tasks, while downstream heads use the duplicate signal differently. This supports a 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.

 doi:10.20944/preprints202607.1441.v1

B.4. Language-Selective Components

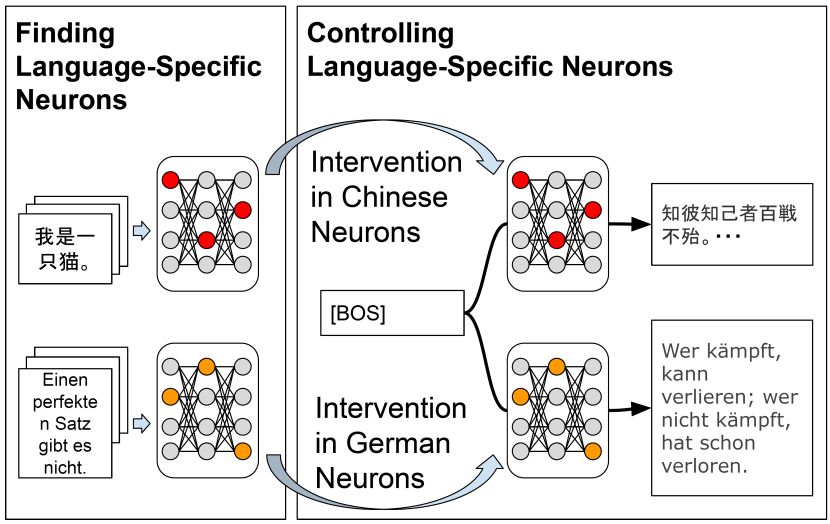

Figure 6. Overview of the language-selective neuron analysis in Kojima et al. (2024). The study identifies neurons with stronger responses to particular languages (e.g., Chinese or German) and tests whether enhancing or suppressing them changes output language probabilities. We treat these interventions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

B.5. Cross-Lingual Transfer

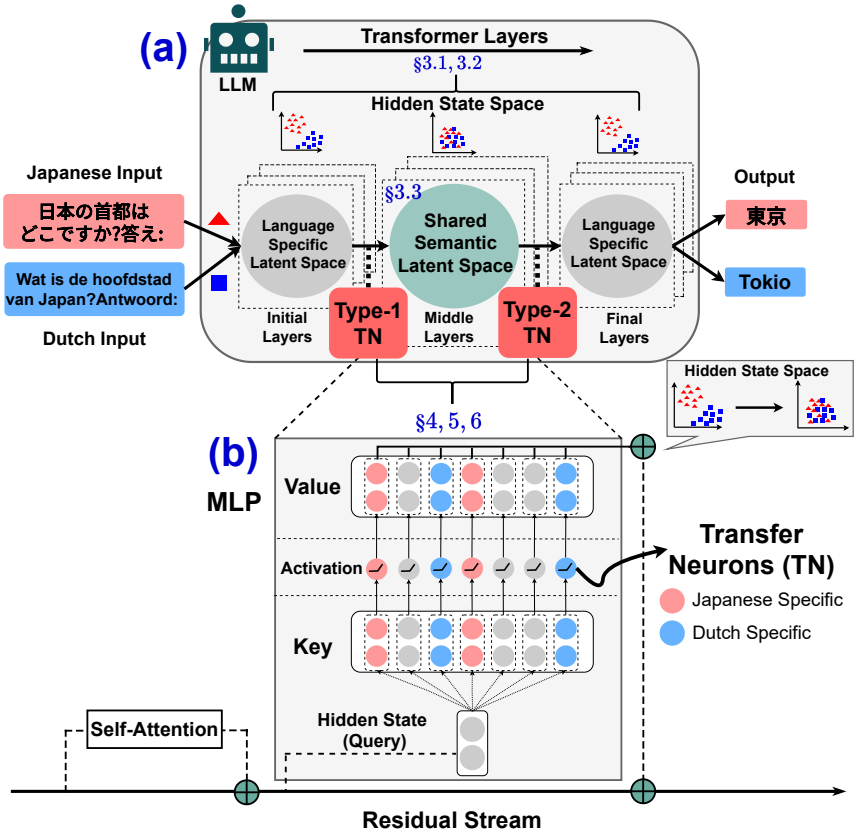

Figure 7. Overview of the Transfer Neurons Hypothesis proposed by Tezuka and Inoue (2025). The hypothesis models cross-lingual transfer as movement between language-specific latent spaces and a shared semantic latent space, mediated by specialized MLP neurons. This provides a candidate mechanism for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

 doi:10.20944/preprints202607.1441.v1

B.6. Mechanisms of Semantic Alignment

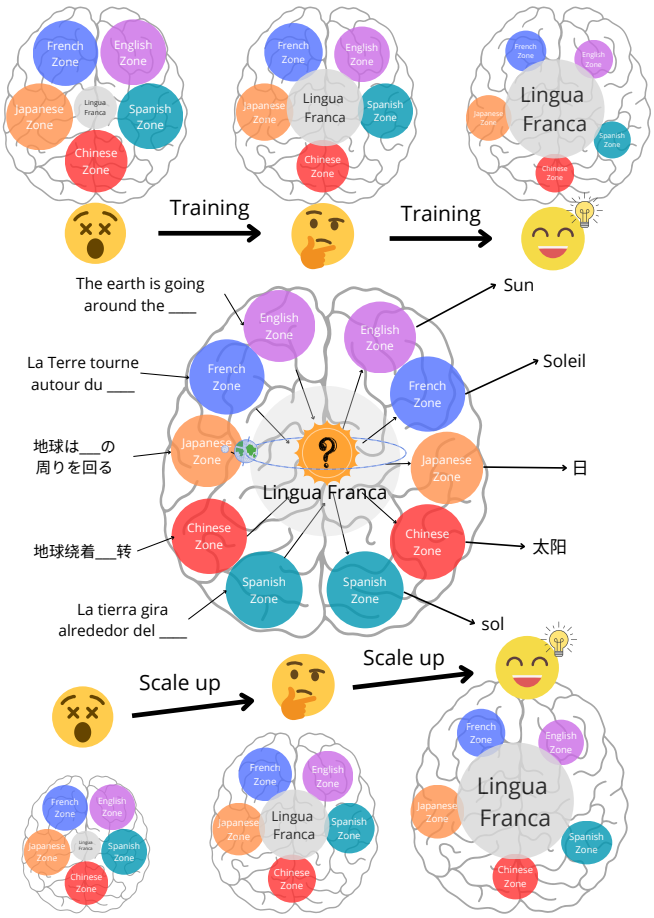

Figure 8. Overview of the "Lingua Franca" hypothesis in Zeng et al. (2025). The study proposes that multilingual models can map surface forms from different languages into a partially shared latent semantic space, with stronger alignment observed in larger or more extensively trained models. We present 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.

References

Adeeba, F., Dillon, B., Sajjad, H., & Bhatt, R. (2025). *Urblimp: A benchmark for evaluating the linguistic competence of large language models in urdu.* Available online: https://arxiv.org/abs/2508.01006 (accessed on).

AlKhamissi, B., Tuckute, G., Bosselut, A., & Schrimpf, M. (2025). *The llm language network: A neuroscientific approach for identifying causally task-relevant units.* Available online: https://arxiv.org/abs/2411.02280 (accessed on).

AlKhamissi, B., Tuckute, G., Tang, Y., Binhuraib, T., Bosselut, A., & Schrimpf, M. (2025). *From language to cognition: How llms outgrow the human language network.* Available online: https://arxiv.org/abs/2503.01830 (accessed on).

Alrajhi, W. A., Al-Khalifa, H., & AlSalman, A. (2022, December). Assessing the Linguistic Knowledge in Arabic Pre-trained Language Models Using Minimal Pairs. In H. Bouamor et al. (Eds.), *Proceedings of the seventh arabic natural language processing workshop (wanlp)* (pp. 185–193). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. Available online: https://aclanthology.org/2022.wanlp-1.17/ (accessed on). https://doi.org/10.18653/v1/2022.wanlp-1.17.

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. (2025, March). Circuit Tracing: Revealing Computational Graphs in Language Models. *Anthropic Technical Report*. Available online: https://transformer-circuits.pub/2025/attribution-graphs/methods.html (accessed on).

 doi:10.20944/preprints202607.1441.v1

Andrylie, L. M., Rahmanisa, I., Ihsani, M. K., Wicaksono, A. F., Wibowo, H. A., & Aji, A. F. (2025). *Sparse autoencoders can capture language-specific concepts across diverse languages.* Available online: https://arxiv.org/abs/2507.11230 (accessed on).

Arora, A., Jurafsky, D., & Potts, C. (2024). *Causalgym: Benchmarking causal interpretability methods on linguistic tasks.* Available online: https://arxiv.org/abs/2402.12560 (accessed on).

Banayeeanzade, A., Tak, A. N., Bahrani, F., Bolourani, A., Blas, L., Ferrara, E., Gratch, J., & Karimireddy, S. P. (2025). *Psychological steering in llms: An evaluation of effectiveness and trustworthiness.* Available online: https://arxiv.org/abs/2510.04484 (accessed on).

Barbini, M., Piccini Bianchessi, M. L., Bressan, V., Fusco, A., Neri, S., Rossi, S., Sgrizzi, T., & Chesi, C. (2025, September). BLiMP-IT: Harnessing Automatic Minimal Pair Generation for Italian Language Model Evaluation. In C. Bosco, E. Jezek, M. Polignano, & M. Sanguinetti (Eds.), *Proceedings of the eleventh italian conference on computational linguistics (clic-it 2025)* (pp. 64–71). Cagliari, Italy: CEUR Workshop Proceedings. Available online: https://aclanthology.org/2025.clicit-1.8/ (accessed on).

Bayazit, D., Mueller, A., & Bosselut, A. (2025). *Crosscoding through time: Tracking emergence & consolidation of linguistic representations throughout llm pretraining.* Available online: https://arxiv.org/abs/2509.05291 (accessed on).

Başar, E., Padovani, F., Jumelet, J., & Bisazza, A. (2025). TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs. In *Proceedings of the 2025 conference on empirical methods in natural language processing* (p. 16506–16521). Association for Computational Linguistics. Available online: http://dx.doi.org/10.18653/v1/2025.emnlp-main.834 (accessed on). https://doi.org/10.18653/v1/2025.emnlp-main.834.

Beauchemin, D., Veilleux, P.-L., Khoury, R., & Roy, J.-P. (2025). *Qfrblimp: a quebec-french benchmark of linguistic minimal pairs.* Available online: https://arxiv.org/abs/2509.25664 (accessed on).

Belinkov, Y. (2022, March). Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, *48*(1), 207–219. Available online: https://aclanthology.org/2022.cl-1.7/ (accessed on). https://doi.org/10.1162/coli_a_00422.

Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., & Steinhardt, J. (2025). *Eliciting latent predictions from transformers with the tuned lens.* Available online: https://arxiv.org/abs/2303.08112 (accessed on).

Bolhuis, J. J., Crain, S., Fong, S., & Moro, A. (2024). Three reasons why AI doesn't model human language. *Nature*, *627*(8004), 489. Available online: https://doi.org/10.1038/d41586-024-00824-z (accessed on).

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., McCauley, H., et al. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*. Available online: https://transformer-circuits.pub/2023/monosemantic-features (accessed on).

Brinkmann, J., Wendler, C., Bartelt, C., & Mueller, A. (2025, April). Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *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)* (pp. 6131–6150). Albuquerque, New Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.naacl-long.312/ (accessed on). https://doi.org/10.18653/v1/2025.naacl-long.312.

Cahyawijaya, S., Zhang, R., Cruz, J. C. B., Lovenia, H., Gilbert, E., Nomoto, H., & Aji, A. F. (2025, April). Thank You, Stingray: Multilingual Large Language Models Can Not (Yet) Disambiguate Cross-Lingual Word Senses. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Findings of the association for computational linguistics: Naacl 2025* (pp. 3228–3250). Albuquerque, New Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.findings-naacl.178/ (accessed on). https://doi.org/10.18653/v1/2025.findings-naacl.178.

Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., & Pickering, M. J. (2024). *Do large language models resemble humans in language use?* Available online: https://arxiv.org/abs/2303.08014 (accessed on).

Chang, T. A., Arnett, C., Tu, Z., & Bergen, B. (2024, November). When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 4074–4096). Miami, Florida, USA: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.emnlp-main.236/ (accessed on). https://doi.org/10.18653/v1/2024.emnlp-main.236.

Chang, T. A., Tu, Z., & Bergen, B. K. (2024, 11). Characterizing Learning Curves During Language Model Pre-Training: Learning, Forgetting, and Stability. *Transactions of the Association for Computational Linguistics*,

 doi:10.20944/preprints202607.1441.v1

12, 1346-1362. Available online: https://doi.org/10.1162/tacl_a_00708 (accessed on). https://doi.org/10.1162/tacl_a_00708.

Chen, J., Chen, W., Su, J., Xu, J., Lin, H., Ren, M., Lu, Y., Han, X., & Sun, L. (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*. Available online: https://openreview.net/forum?id=eznTVIM3bs (accessed on).

Chi, E. A., Hewitt, J., & Manning, C. D. (2020). *Finding universal grammatical relations in multilingual bert*. Available online: https://arxiv.org/abs/2005.04511 (accessed on).

Chi, Z., Huang, H., & Mao, X.-L. (2023, July). Can Cross-Lingual Transferability of Multilingual Transformers Be Activated Without End-Task Data? In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 12572–12584). Toronto, Canada: Association for Computational Linguistics. Available online: https://aclanthology.org/2023.findings-acl.796/ (accessed on). https://doi.org/10.18653/v1/2023.findings-acl.796.

Chlapanis, O. S., Mastromichalakis, O. M., & Papadimitriou, C. H. (2026). *The grounding gap: How llms anchor the meaning of abstract concepts differently from humans*. Available online: https://arxiv.org/abs/2605.08837 (accessed on).

Cho, H., Yang, H., Kurkoski, B. M., & Inoue, N. (2025). *Binary autoencoder for mechanistic interpretability of large language models*. Available online: https://arxiv.org/abs/2509.20997 (accessed on).

Cho, I., & Hockenmaier, J. (2025). Analyzing Multilingualism in Large Language Models with Sparse Autoencoders. In *Second conference on language modeling*. Available online: https://openreview.net/forum?id=NmGSvZoU3K (accessed on).

Chomsky, N. (2023, March 8). Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. Available online: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html (accessed on). (Opinion)

Choshen, L., Hacohen, G., Weinshall, D., & Abend, O. (2022). *The grammar-learning trajectories of neural language models*. Available online: https://arxiv.org/abs/2109.06096 (accessed on).

Chou, C.-T., Liu, G., Sun, J., Blondin, C., Zhu, K., Sharma, V., & O'Brien, S. (2025). *Causal language control in multilingual transformers via sparse feature steering*. Available online: https://arxiv.org/abs/2507.13410 (accessed on).

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., & He, P. (2024). *Dola: Decoding by contrasting layers improves factuality in large language models*. Available online: https://arxiv.org/abs/2309.03883 (accessed on).

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Thirty-seventh conference on neural information processing systems*. Available online: https://openreview.net/forum?id=89ia77nZ8u (accessed on).

Deng, B., Wan, Y., Yang, B., Zhang, Y., & Feng, F. (2025, July). Unveiling Language-Specific Features in Large Language Models via Sparse Autoencoders. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4563–4608). Vienna, Austria: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.acl-long.229/ (accessed on). https://doi.org/10.18653/v1/2025.acl-long.229.

Diego-Simón, P. J., Chemla, E., King, J.-R., & Lakretz, Y. (2025). *Probing syntax in large language models: Successes and remaining challenges*. Available online: https://arxiv.org/abs/2508.03211 (accessed on).

Duan, X., Xiao, B., Tang, X., & Cai, Z. G. (2024). *Hlb: Benchmarking llms' humanlikeness in language use*. Available online: https://arxiv.org/abs/2409.15890 (accessed on).

Duan, X., Yao, Z., Zhang, Y., Wang, S., & Cai, Z. G. (2025). *How syntax specialization emerges in language models*. Available online: https://arxiv.org/abs/2505.19548 (accessed on).

Duan, X., Zhou, X., Xiao, B., & Cai, Z. (2025, January). Unveiling Language Competence Neurons: A Psycholinguistic Approach to Model Interpretability. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 10148–10157). Abu Dhabi, UAE: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.coling-main.677/ (accessed on).

Dunefsky, J., Chlenski, P., & Nanda, N. (2024). *Transcoders find interpretable llm feature circuits*. Available online: https://arxiv.org/abs/2406.11944 (accessed on).

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). *Toy models of superposition*. Available online: https://arxiv.org/abs/2209.10652 (accessed on).

 doi:10.20944/preprints202607.1441.v1

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*. Available online: https://transformer-circuits.pub/2021/framework/index.html (accessed on).

Evanson, L., Lakretz, Y., & King, J.-R. (2023). *Language acquisition: do children and language models follow similar learning stages?* Available online: https://arxiv.org/abs/2306.03586 (accessed on).

Fang, J., Jiang, H., Wang, K., Ma, Y., Shi, J., Wang, X., He, X., & Chua, T.-S. (2025). AlphaEdit: Null-Space Constrained Model Editing for Language Models. In *The thirteenth international conference on learning representations*. Available online: https://openreview.net/forum?id=HvSytvg3Jh (accessed on).

Ferrando, J., & Voita, E. (2024). *Information flow routes: Automatically interpreting language models at scale.* Available online: https://arxiv.org/abs/2403.00824 (accessed on).

Feucht, S., Todd, E., Wallace, B., & Bau, D. (2025). *The dual-route model of induction.* Available online: https://arxiv.org/abs/2504.03022 (accessed on).

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., & Belinkov, Y. (2021, August). Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 1828–1843). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.acl-long.144/ (accessed on). https://doi.org/10.18653/v1/2021.acl-long.144.

Freenor, M., & Alvarez, L. (2026). *Mapping semantic & syntactic relationships with geometric rotation.* Available online: https://arxiv.org/abs/2510.09790 (accessed on).

Gao, C., Ma, Z., Chen, J., Li, P., Huang, S., & Li, J. (2025). Increasing alignment of large language models with language processing in the human brain. *Nature computational science*, 1–11. Available online: https://www.nature.com/articles/s43588-025-00863-0 (accessed on).

Gao, Y., Meng, Q., Zhou, Y., & Pan, L. (2026). *Towards intrinsic interpretability of large language models:a survey of design principles and architectures.* Available online: https://arxiv.org/abs/2604.16042 (accessed on).

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020, July). SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 70–76). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2020.acl-demos.10/ (accessed on). https://doi.org/10.18653/v1/2020.acl-demos.10.

Geva, M., Caciularu, A., Wang, K., & Goldberg, Y. (2022, December). Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 30–45). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Available online: https://aclanthology.org/2022.emnlp-main.3/ (accessed on). https://doi.org/10.18653/v1/2022.emnlp-main.3.

Giulianelli, M., Harding, J., Mohnert, F., Hupkes, D., & Zuidema, W. (2018, November). Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information. In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 240–248). Brussels, Belgium: Association for Computational Linguistics. Available online: https://aclanthology.org/W18-5426/ (accessed on). https://doi.org/10.18653/v1/W18-5426.

Goldstein, A., Ham, E., Schain, M., Nastase, S. A., Aubrey, B., Zada, Z., Grinstein-Dabush, A., Gazula, H., Feder, A., Doyle, W., 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. Available online: https://doi.org/10.1038/s41467-025-65518-0 (accessed on).

Graichen, N., de Dios-Flores, I., & Boleda, G. (2026). *The grammar of transformers: A systematic review of interpretability research on syntactic knowledge in language models.* Available online: https://arxiv.org/abs/2601.19926 (accessed on).

Guo, Y., Conia, S., Zhou, Z., Li, M., Potdar, S., & Xiao, H. (2025). *Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms.* Available online: https://arxiv.org/abs/2410.15956 (accessed on).

Guo, Y., Shang, G., Vazirgiannis, M., & Clavel, C. (2024, June). The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the association for computational linguistics: Naacl 2024* (pp. 3589–3604). Mexico City, Mexico: Association for

 doi:10.20944/preprints202607.1441.v1

Computational Linguistics. Available online: https://aclanthology.org/2024.findings-naacl.228/ (accessed on). https://doi.org/10.18653/v1/2024.findings-naacl.228.

Gupta, R., Arcuschin, I., Kwa, T., & Garriga-Alonso, A. (2025). *Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques.* Available online: https://arxiv.org/abs/2407.14494 (accessed on).

Gur-Arieh, Y., Geva, M., & Geiger, A. (2025). *Mixing mechanisms: How language models retrieve bound entities in-context.* Available online: https://arxiv.org/abs/2510.06182 (accessed on).

Gurgurov, D., Trinley, K., Ghussin, Y. A., Baeumel, T., van Genabith, J., & Ostermann, S. (2025). *Language arithmetics: Towards systematic language neuron identification and manipulation.* Available online: https://arxiv.org/abs/2507.22608 (accessed on).

Gurgurov, D., van Genabith, J., & Ostermann, S. (2025). *Sparse subnetwork enhancement for underrepresented languages in large language models.* Available online: https://arxiv.org/abs/2510.13580 (accessed on).

Gurnee, W., Horsley, T., Guo, Z. C., Kheirkhah, T. R., Sun, Q., Hathaway, W., Nanda, N., & Bertsimas, D. (2024). Universal Neurons in GPT2 Language Models. *Transactions on Machine Learning Research*. Available online: https://openreview.net/forum?id=ZeI104QZ8I (accessed on).

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., & Bertsimas, D. (2023). *Finding neurons in a haystack: Case studies with sparse probing.* Available online: https://arxiv.org/abs/2305.01610 (accessed on).

Hanna, M., & Mueller, A. (2025, April). Incremental Sentence Processing Mechanisms in Autoregressive Transformer Language Models. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *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)* (pp. 3181–3203). Albuquerque, New Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.naacl-long.164/ (accessed on). https://doi.org/10.18653/v1/2025.naacl-long.164.

Hao, S., & Linzen, T. (2023, December). Verb Conjugation in Transformers Is Determined by Linear Encodings of Subject Number. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 4531–4539). Singapore: Association for Computational Linguistics. Available online: https://aclanthology.org/2023.findings-emnlp.300/ (accessed on). https://doi.org/10.18653/v1/2023.findings-emnlp.300.

Harrasse, A., Draye, F., Jin, Z., & Schölkopf, B. (2025). *Tracing multilingual representations in llms with cross-layer transcoders.* Available online: https://arxiv.org/abs/2511.10840 (accessed on).

He, L., Chen, P., Nie, E., Li, Y., & Brennan, J. R. (2024). *Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs.* Available online: https://arxiv.org/abs/2403.17299 (accessed on).

He, Z., Ge, X., Tang, Q., Sun, T., Cheng, Q., & Qiu, X. (2024). *Dictionary learning improves patch-free circuit discovery in mechanistic interpretability: A case study on othello-gpt.* Available online: https://arxiv.org/abs/2402.12201 (accessed on).

He, Z., Zhao, H., Qiao, Y., Yang, F., Payani, A., Ma, J., & Du, M. (2025). *Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models.* Available online: https://arxiv.org/abs/2502.11356 (accessed on).

Heimersheim, S., & Nanda, N. (2024). *How to use and interpret activation patching.* Available online: https://arxiv.org/abs/2404.15255 (accessed on).

Hewitt, J., & Manning, C. D. (2019, June). A Structural Probe for Finding Syntax in Word Representations. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4129–4138). Minneapolis, Minnesota: Association for Computational Linguistics. Available online: https://aclanthology.org/N19-1419/ (accessed on). https://doi.org/10.18653/v1/N19-1419.

Honarmand, M., Sharma, A., AlKhamissi, B., Mehrer, J., & Schrimpf, M. (2025). *Inducing dyslexia in vision language models.* Available online: https://arxiv.org/abs/2509.24597 (accessed on).

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024, 04). 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. Available online: https://doi.org/10.1162/nol_a_00137 (accessed on). https://doi.org/10.1162/nol_a_00137.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020, July). A Systematic Assessment of Syntactic Generalization in Neural Language Models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1725–1744). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2020.acl-main.158/ (accessed on). https://doi.org/10.18653/v1/2020.acl-main.158.

 doi:10.20944/preprints202607.1441.v1

Hu, J., & Levy, R. P. (2023). Prompting is not a substitute for probability measurements in large language models. In *The 2023 conference on empirical methods in natural language processing*. Available online: https://openreview.net/forum?id=hMqRphmoM9 (accessed on).

Hu, M. Y., Petty, J., Shi, C., Merrill, W., & Linzen, T. (2025, July). Between Circuits and Chomsky: Pre-pretraining on Formal Languages Imparts Linguistic Biases. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9691–9709). Vienna, Austria: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.acl-long.478/ (accessed on). https://doi.org/10.18653/v1/2025.acl-long.478.

Huang, J., Geiger, A., D'Oosterlinck, K., Wu, Z., & Potts, C. (2023). *Rigorously assessing natural language explanations of neurons*. Available online: https://arxiv.org/abs/2309.10312 (accessed on).

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., & Sharkey, L. (2024). Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *The twelfth international conference on learning representations*. Available online: https://openreview.net/forum?id=F76bwRSLeK (accessed on).

Huo, J., Yan, Y., Hu, B., Yue, Y., & Hu, X. (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* (p. 6801–6816). Association for Computational Linguistics. Available online: http://dx.doi.org/10.18653/v1/2024.emnlp-main.387 (accessed on). https://doi.org/10.18653/v1/2024.emnlp-main.387.

Iqbal, A., Younas, M., Iftikhar, S., Fatima, F., & Saleem, R. (2025). Spam detection using hybrid model on fusion of spammer behavior and linguistics features. *Egyptian Informatics Journal*, *29*, 100605. Available online: https://www.sciencedirect.com/science/article/pii/S1110866524001683 (accessed on). https://doi.org/https://doi.org/10.1016/j.eij.2024.100605.

Jing, Y., Yao, Z., Guo, H., Ran, L., Wang, X., Hou, L., & Li, J. (2025, November). LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 28232–28251). Suzhou, China: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.emnlp-main.1433/ (accessed on). https://doi.org/10.18653/v1/2025.emnlp-main.1433.

Jiralerspong, T., & Bricken, T. (2026). *Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between llms*. Available online: https://arxiv.org/abs/2602.11729 (accessed on).

Jumelet, J., Weissweiler, L., Nivre, J., & Bisazza, A. (2025). *Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs*. Available online: https://arxiv.org/abs/2504.02768 (accessed on).

Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., & Nanda, N. (2025). *Are sparse autoencoders useful? a case study in sparse probing*. Available online: https://arxiv.org/abs/2502.16681 (accessed on).

Karny, S., Baez, A., & Pataranutaporn, P. (2025). *Neural transparency: Mechanistic interpretability interfaces for anticipating model behaviors for personalized ai*. Available online: https://arxiv.org/abs/2511.00230 (accessed on).

Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C., Ayonrinde, K., Till, D., Wearden, M., Conmy, A., Marks, S., & Nanda, N. (2025). *Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability*. Available online: https://arxiv.org/abs/2503.09532 (accessed on).

Kendiukhov, I. (2025). *A review of developmental interpretability in large language models*. Available online: https://arxiv.org/abs/2508.15841 (accessed on).

Kharlapenko, D., Shabalin, S., Barez, F., Conmy, A., & Nanda, N. (2025). *Scaling sparse feature circuit finding for in-context learning*. Available online: https://arxiv.org/abs/2504.13756 (accessed on).

Kissane, H., Schilling, A., & Krauss, P. (2025, May). Probing Internal Representations of Multi-Word Verbs in Large Language Models. In A. K. Ojha et al. (Eds.), *Proceedings of the 21st workshop on multiword expressions (mwe 2025)* (pp. 7–13). Albuquerque, New Mexico, U.S.A.: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.mwe-1.2/ (accessed on). https://doi.org/10.18653/v1/2025.mwe-1.2.

Kojima, T., Okimura, I., Iwasawa, Y., Yanaka, H., & Matsuo, Y. (2024). *On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons*. Available online: https://arxiv.org/abs/2404.02431 (accessed on).

Kryvosheieva, D., de Varda, A., Fedorenko, E., & Tuckute, G. (2025). *Different types of syntactic agreement recruit the same units within large language models*. Available online: https://arxiv.org/abs/2512.03676 (accessed on).

 doi:10.20944/preprints202607.1441.v1

Kryvosheieva, D., & Levy, R. (2024). *Controlled evaluation of syntactic knowledge in multilingual language models.* Available online: https://arxiv.org/abs/2411.07474 (accessed on).

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. *Nature communications*, *15*(1), 5523. Available online: https://doi.org/10.1038/s41467-024-49173-5 (accessed on).

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019, June). The emergence of number and syntax units in LSTM language models. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 11–20). Minneapolis, Minnesota: Association for Computational Linguistics. Available online: https://aclanthology.org/N19-1002/ (accessed on). https://doi.org/10.18653/v1/N19-1002.

Leong, W. Q., Ngui, J. G., Susanto, Y., Rengarajan, H., Sarveswaran, K., & Tjhi, W. C. (2023). *Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models.* Available online: https://arxiv.org/abs/2309.06085 (accessed on).

Li, X., Yong, Z.-X., & Bach, S. H. (2024). *Preference tuning for toxicity mitigation generalizes across languages.* Available online: https://arxiv.org/abs/2406.16235 (accessed on).

Libovický, J., Rosa, R., & Fraser, A. (2020, November). On the Language Neutrality of Pre-trained Multilingual Representations. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1663–1674). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2020.findings-emnlp.150/ (accessed on). https://doi.org/10.18653/v1/2020.findings-emnlp.150.

Lieberum, T., Rahtz, M., Kramár, J., Nanda, N., Irving, G., Shah, R., & Mikulik, V. (2023). *Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla.* Available online: https://arxiv.org/abs/2307.09458 (accessed on).

Lin, J. (2023). *Neuronpedia: Interactive reference and tooling for analyzing neural networks.* Available online: https://www.neuronpedia.org (accessed on). (Software available from neuronpedia.org)

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. (2025, March). On the Biology of a Large Language Model. *Anthropic Technical Report*. Available online: https://transformer-circuits.pub/2025/attribution-graphs/biology.html (accessed on).

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. Available online: https://aclanthology.org/Q16-1037/ (accessed on). https://doi.org/10.1162/tacl_a_00115.

Liu, W., Xiang, M., & Ding, N. (2025). Active use of latent tree-structured sentence representation in humans and large language models. *Nature Human Behaviour*, 1–14. Available online: https://doi.org/10.1038/s41562-025-02297-0 (accessed on).

Liu, W., Xu, Y., Xu, H., Chen, J., Hu, X., & Wu, J. (2024, November). Unraveling Babel: Exploring Multilingual Activation Patterns of LLMs and Their Applications. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 11855–11881). Miami, Florida, USA: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.emnlp-main.662/ (accessed on). https://doi.org/10.18653/v1/2024.emnlp-main.662.

Liu, X., Song, Q., Zhou, Q., Du, H., Xu, S., Jiang, W., Zhang, W., & Jia, X. (2025). *Focusing on language: Revealing and exploiting language attention heads in multilingual large language models.* Available online: https://arxiv.org/abs/2511.07498 (accessed on).

Liu, Y., Shen, Y., Zhu, H., Xu, L., Qian, Z., Song, S., Zhang, K., Tang, J., Zhang, P., Yang, B., Wang, R., & Hu, H. (2025). *A systematic assessment of language models with linguistic minimal pairs in chinese.* Available online: https://arxiv.org/abs/2411.06096 (accessed on).

Maar, J., Paperno, D., McDougall, C. S., & Nanda, N. (2026). *What's the plan? metrics for implicit planning in llms and their application to rhyme generation and question answering*. Available online: https://arxiv.org/abs/2601.20164 (accessed on).

Makelov, A., Lange, G., Geiger, A., & Nanda, N. (2024). Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. In *The twelfth international conference on learning representations.* Available online: https://openreview.net/forum?id=Ebt7JgMHv1 (accessed on).

 doi:10.20944/preprints202607.1441.v1

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. Available online: https://doi.org/10.1073/pnas.1907367117 (accessed on).

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2025). Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *The thirteenth international conference on learning representations*. Available online: https://openreview.net/forum?id=I4e82CIDxv (accessed on).

McGiff, J., Tran, K.-T., Mulcahy, W., Luinín, D. Ó., Dalzell, J., Bhroin, R. N., Burke, A., O'Sullivan, B., Nguyen, H. D., & Nikolov, N. S. (2025). *Irish-blimp: A linguistic benchmark for evaluating human and language model performance in a low-resource setting*. Available online: https://arxiv.org/abs/2510.20957 (accessed on).

Méloux, M., Maniu, S., Portet, F., & Peyrard, M. (2025). Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable? In *The thirteenth international conference on learning representations*. Available online: https://openreview.net/forum?id=5IWJBStfU7 (accessed on).

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2023). *Locating and editing factual associations in gpt*. Available online: https://arxiv.org/abs/2202.05262 (accessed on).

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., & Bau, D. (2023). *Mass-editing memory in a transformer*. Available online: https://arxiv.org/abs/2210.07229 (accessed on).

Merullo, J., Eickhoff, C., & Pavlick, E. (2024). Circuit Component Reuse Across Tasks in Transformer Language Models. In *The twelfth international conference on learning representations*. Available online: https://openreview.net/forum?id=fpoAYV6Wsk (accessed on).

Mohebbi, H., Jumelet, J., Hanna, M., Alishahi, A., & Zuidema, W. (2024, March). Transformer-specific Interpretability. In M. Mesgar & S. Loáiciga (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: Tutorial abstracts* (pp. 21–26). St. Julian's, Malta: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.eacl-tutorials.4/ (accessed on). https://doi.org/10.18653/v1/2024.eacl-tutorials.4.

Mondal, S. K., Sen, S., Singhania, A., & Jyothi, P. (2025, May). Language-Specific Neurons Do Not Facilitate Cross-Lingual Transfer. In A. Drozd, J. Sedoc, S. Tafreshi, A. Akula, & R. Shu (Eds.), *The sixth workshop on insights from negative results in nlp* (pp. 46–62). Albuquerque, New Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.insights-1.6/ (accessed on). https://doi.org/10.18653/v1/2025.insights-1.6.

Mondorf, P., Wang, M., Gerstner, S., Hakimi, A. D., Liu, Y., Veloso, L., Zhou, S., Schütze, H., & Plank, B. (2025). *Blackboxnlp-2025 mib shared task: Exploring ensemble strategies for circuit localization methods*. Available online: https://arxiv.org/abs/2510.06811 (accessed on).

Mueller, A., Brinkmann, J., Li, M., Marks, S., Pal, K., Prakash, N., Rager, C., Sankaranarayanan, A., Sharma, A. S., Sun, J., Todd, E., Bau, D., & Belinkov, Y. (2025). *The quest for the right mediator: Surveying mechanistic interpretability through the lens of causal mediation analysis*. Available online: https://arxiv.org/abs/2408.01416 (accessed on).

Mueller, A., Geiger, A., Wiegreffe, S., Arad, D., Arcuschin, I., Belfki, A., Chan, Y. S., Fiotto-Kaufman, J., Haklay, T., Hanna, M., Huang, J., Gupta, R., Nikankin, Y., Orgad, H., Prakash, N., Reusch, A., Sankaranarayanan, A., Shao, S., Stolfo, A., … Belinkov, Y. (2025). *Mib: A mechanistic interpretability benchmark*. Available online: https://arxiv.org/abs/2504.13151 (accessed on).

Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, N., & Linzen, T. (2020). *Cross-linguistic syntactic evaluation of word prediction models*. Available online: https://arxiv.org/abs/2005.00187 (accessed on).

Mueller, A., Xia, Y., & Linzen, T. (2022). *Causal analysis of syntactic agreement neurons in multilingual language models*. Available online: https://arxiv.org/abs/2210.14328 (accessed on).

Muller, B., Elazar, Y., Sagot, B., & Seddah, D. (2021, April). First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 2214–2231). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.eacl-main.189/ (accessed on). https://doi.org/10.18653/v1/2021.eacl-main.189.

Nanda, N. (2023). Attribution Patching: Activation Patching at Industrial Scale. *Neel Nanda's Blog*. Available online: https://www.neelnanda.io/mechanistic-interpretability/attribution-patching (accessed on).

Nanda, N., & Bloom, J. (2022). *Transformerlens*. https://github.com/TransformerLensOrg/TransformerLens.

Nie, E., Schmid, H., & Schuetze, H. (2025, November). Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the association for computational linguistics: Emnlp 2025* (pp. 690–706). Suzhou, China:

 doi:10.20944/preprints202607.1441.v1

Association for Computational Linguistics. Available online: https://aclanthology.org/2025.findings-emnlp.37/ (accessed on). https://doi.org/10.18653/v1/2025.findings-emnlp.37.

Nostalgebraist. (2020). Interpreting GPT: The logit lens. *Blog Post*.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, *5*(3), e00024–001.

Orhan, P., Diego-Simón, P., Chemla, E., Lakretz, Y., Boubenec, Y., & King, J.-R. (2026). *Emergence of phonemic, syntactic, and semantic representations in artificial neural networks.* Available online: https://arxiv.org/abs/2601.18617 (accessed on).

Ou, Y., Yao, Y., Zhang, N., Jin, H., Sun, J., Deng, S., Li, Z., & Chen, H. (2025). *How do llms acquire new knowledge? a knowledge circuits perspective on continual pre-training.* Available online: https://arxiv.org/abs/2502.11196 (accessed on).

Pal, K., Sun, J., Yuan, A., Wallace, B., & Bau, D. (2023). Future Lens: Anticipating Subsequent Tokens from a Single Hidden State. In *Proceedings of the 27th conference on computational natural language learning (conll)* (p. 548–560). Association for Computational Linguistics. Available online: http://dx.doi.org/10.18653/v1/2023.conll-1.37 (accessed on). https://doi.org/10.18653/v1/2023.conll-1.37.

Panigrahi, A., Saunshi, N., Zhao, H., & Arora, S. (2023). *Task-specific skill localization in fine-tuned language models.* Available online: https://arxiv.org/abs/2302.06600 (accessed on).

Pires, T., Schlinger, E., & Garrette, D. (2019, July). How Multilingual is Multilingual BERT? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4996–5001). Florence, Italy: Association for Computational Linguistics. Available online: https://aclanthology.org/P19-1493/ (accessed on). https://doi.org/10.18653/v1/P19-1493.

Potertì, D., Seveso, A., & Mercorio, F. (2025, November). Can Role Vectors Affect LLM Behaviour? In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the association for computational linguistics: Emnlp 2025* (pp. 17735–17747). Suzhou, China: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.findings-emnlp.963/ (accessed on). https://doi.org/10.18653/v1/2025.findings-emnlp.963.

Rai, D., Zhou, Y., Feng, S., Saparov, A., & Yao, Z. (2025). *A practical review of mechanistic interpretability for transformer-based language models.* Available online: https://arxiv.org/abs/2407.02646 (accessed on).

Riemenschneider, F., & Frank, A. (2025). *Cross-lingual generalization and compression: From language-specific to shared neurons.* Available online: https://arxiv.org/abs/2506.01629 (accessed on).

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. (2024, August). Steering Llama 2 via Contrastive Activation Addition. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 15504–15522). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.acl-long.828/ (accessed on). https://doi.org/10.18653/v1/2024.acl-long.828.

Rufail, A., Rathore, S., Son, D., Simon, A., Dave, S., Zhang, D., Blondin, C., O'Brien, S., & Zhu, K. (2025). Semantic Convergence: Investigating Shared Representations Across Scaled LLMs. In *Acl 2025 student research workshop*. Available online: https://openreview.net/forum?id=oOxrKNo1lQ (accessed on).

Ruzzetti, E. S., Zanzotto, F. M., & Caselli, T. (2026, March). Lexical Popularity: Quantifying the Impact of Pre-training for LLM Performance. In V. Demberg, K. Inui, & L. Marquez (Eds.), *Proceedings of the 19th conference of the European chapter of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 1209–1230). Rabat, Morocco: Association for Computational Linguistics. Available online: https://aclanthology.org/2026.eacl-long.55/ (accessed on). https://doi.org/10.18653/v1/2026.eacl-long.55.

Sajjad, H., Durrani, N., & Dalvi, F. (2021). Neuron-level Interpretation of Deep NLP Models: A Survey. *Transactions of the Association for Computational Linguistics*, *10*, 1285-1303. Available online: https://arxiv.org/abs/2108.13138 (accessed on).

Schulz, L. Y., Mitropolsky, D., & Poggio, T. (2025). *Unraveling syntax: How language models learn context-free grammars.* Available online: https://arxiv.org/abs/2510.02524 (accessed on).

Schut, L., Gal, Y., & Farquhar, S. (2025). Do Multilingual LLMs Think In English? In *Iclr 2025 workshop on building trust in language models and applications.* Available online: https://openreview.net/forum?id=I8BOtOPcOv (accessed on).

Shafran, O., Ronen, S., Fahn, O., Ravfogel, S., Geiger, A., & Geva, M. (2026). *From directions to regions: Decomposing activations in language models via local geometry.* Available online: https://arxiv.org/abs/2602.02464 (accessed on).

 doi:10.20944/preprints202607.1441.v1

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., … McGrath, T. (2025). *Open problems in mechanistic interpretability.* Available online: https://arxiv.org/abs/2501.16496 (accessed on).

Shu, B., Singh, A., & ElSherief, M. (2026). *From syntax to emotion: A mechanistic analysis of emotion inference in llms.* Available online: https://arxiv.org/abs/2604.25866 (accessed on).

Shu, D., Wu, X., Zhao, H., Rai, D., Yao, Z., Liu, N., & Du, M. (2025). *A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models.* Available online: https://arxiv.org/abs/2503.05613 (accessed on).

Singh, A. K., Chan, S. C. Y., Moskovitz, T., Grant, E., Saxe, A. M., & Hill, F. (2023). *The transient nature of emergent in-context learning in transformers.* Available online: https://arxiv.org/abs/2311.08360 (accessed on).

Someya, T., & Oseki, Y. (2023, May). JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In A. Vlachos & I. Augenstein (Eds.), *Findings of the association for computational linguistics: Eacl 2023* (pp. 1581–1594). Dubrovnik, Croatia: Association for Computational Linguistics. Available online: https://aclanthology.org/2023.findings-eacl.117/ (accessed on). https://doi.org/10.18653/v1/2023.findings-eacl.117.

Son, D., Rathore, S., Rufail, A., Simon, A., Zhang, D., Dave, S., Blondin, C., Zhu, K., & O'Brien, S. (2025). *Semantic convergence: Investigating shared representations across scaled llms.* Available online: https://arxiv.org/abs/2507.22918 (accessed on).

Song, Y., Krishna, K., Bhatt, R., & Iyyer, M. (2022, December). SLING: Sino Linguistic Evaluation of Large Language Models. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 4606–4634). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Available online: https://aclanthology.org/2022.emnlp-main.305/ (accessed on). https://doi.org/10.18653/v1/2022.emnlp-main.305.

Stańczak, K., Ponti, E., Hennigen, L. T., Cotterell, R., & Augenstein, I. (2022). *Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models.* Available online: https://arxiv.org/abs/2205.02023 (accessed on).

Stickland, A. C., Lyzhov, A., Pfau, J., Mahdi, S., & Bowman, S. R. (2024). *Steering without side effects: Improving post-deployment control of language models.* Available online: https://arxiv.org/abs/2406.15518 (accessed on).

Suijkerbuijk, M., Prins, Z., Kloots, M. d. H., Zuidema, W., & Frank, S. L. (2025, 05). BLiMP-NL: A Corpus of Dutch Minimal Pairs and Acceptability Judgments for Language Model Evaluation. *Computational Linguistics*, 1-35. Available online: https://doi.org/10.1162/coli_a_00559 (accessed on). https://doi.org/10.1162/coli_a_00559.

Syed, A., Rager, C., & Conmy, A. (2024, November). Attribution Patching Outperforms Automated Circuit Discovery. In Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, & H. Chen (Eds.), *Proceedings of the 7th blackboxnlp workshop: Analyzing and interpreting neural networks for nlp* (pp. 407–416). Miami, Florida, US: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.blackboxnlp-1.25/ (accessed on). https://doi.org/10.18653/v1/2024.blackboxnlp-1.25.

Taktasheva, E., Bazhukov, M., Koncha, K., Fenogenova, A., Artemova, E., & Mikhailov, V. (2024, November). RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 9268–9299). Miami, Florida, USA: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.emnlp-main.522/ (accessed on). https://doi.org/10.18653/v1/2024.emnlp-main.522.

Tan, S., Wu, D., & Monz, C. (2024, November). Neuron Specialization: Leveraging Intrinsic Task Modularity for Multilingual Machine Translation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 6506–6527). Miami, Florida, USA: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.emnlp-main.374/ (accessed on). https://doi.org/10.18653/v1/2024.emnlp-main.374.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., & Wen, J.-R. (2024, August). Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5701–5715). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.acl-long.309/ (accessed on). https://doi.org/10.18653/v1/2024.acl-long.309.

 doi:10.20944/preprints202607.1441.v1

Tang, Y., Saini, H., Yao, Z., Lin, Z., Liao, Y., Cui, J., Wang, Y., Du, M., & Liu, D. (2026). *A unified theory of sparse dictionary learning in mechanistic interpretability: Piecewise biconvexity and spurious minima.* Available online: https://arxiv.org/abs/2512.05534 (accessed on).

Tezuka, H., & Inoue, N. (2025). *The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms.* Available online: https://arxiv.org/abs/2509.17030 (accessed on).

Todd, E., Li, M., Sharma, A. S., Mueller, A., Wallace, B. C., & Bau, D. (2024). Function Vectors in Large Language Models. In *The twelfth international conference on learning representations.* Available online: https://openreview.net/forum?id=AwyxtyMwaG (accessed on).

Tufanov, I., Hambardzumyan, K., Ferrando, J., & Voita, E. (2024). *Lm transparency tool: Interactive tool for analyzing transformer language models.* Available online: https://arxiv.org/abs/2404.07004 (accessed on).

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2024). *Steering language models with activation engineering.* Available online: https://arxiv.org/abs/2308.10248 (accessed on).

Waldis, A., Perlitz, Y., Choshen, L., Hou, Y., & Gurevych, I. (2024). Holmes: A Benchmark to Assess the Linguistic Competence of Language Models. *Transactions of the Association for Computational Linguistics*, *12*, 1616–1647. Available online: https://aclanthology.org/2024.tacl-1.88/ (accessed on). https://doi.org/10.1162/tacl_a_00718.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). *Interpretability in the wild: a circuit for indirect object identification in gpt-2 small.* Available online: https://arxiv.org/abs/2211.00593 (accessed on).

Wang, M., Adel, H., Lange, L., Liu, Y., Nie, E., Strötgen, J., & Schuetze, H. (2025, July). Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5075–5094). Vienna, Austria: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.acl-long.253/ (accessed on). https://doi.org/10.18653/v1/2025.acl-long.253.

Wang, M., Xu, Z., Mao, S., Deng, S., Tu, Z., Chen, H., & Zhang, N. (2025). *Beyond prompt engineering: Robust behavior control in llms via steering target atoms.* Available online: https://arxiv.org/abs/2505.20322 (accessed on).

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. Available online: https://aclanthology.org/2020.tacl-1.25/ (accessed on). https://doi.org/10.1162/tacl_a_00321.

Wei, J., Garrette, D., Linzen, T., & Pavlick, E. (2021, November). Frequency Effects on Syntactic Rule Learning in Transformers. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 932–948). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.emnlp-main.72/ (accessed on). https://doi.org/10.18653/v1/2021.emnlp-main.72.

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024, August). Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 15366–15394). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.acl-long.820/ (accessed on). https://doi.org/10.18653/v1/2024.acl-long.820.

Wilson, M., Petty, J., & Frank, R. (2023, 11). How Abstract Is Linguistic Generalization in Large Language Models? Experiments with Argument Structure. *Transactions of the Association for Computational Linguistics*, *11*, 1377-1395. Available online: https://doi.org/10.1162/tacl_a_00608 (accessed on). https://doi.org/10.1162/tacl_a_00608.

Wu, S., & Dredze, M. (2019). *Beto, bentz, becas: The surprising cross-lingual effectiveness of bert.* Available online: https://arxiv.org/abs/1904.09077 (accessed on).

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., & Potts, C. (2025). *Axbench: Steering llms? even simple baselines outperform sparse autoencoders.* Available online: https://arxiv.org/abs/2501.17148 (accessed on).

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., & Potts, C. (2024). ReFT: Representation Finetuning for Language Models. In *The thirty-eighth annual conference on neural information processing systems.* Available online: https://openreview.net/forum?id=fykjplMc0V (accessed on).

 doi:10.20944/preprints202607.1441.v1

Wu, Z., Yu, X. V., Yogatama, D., Lu, J., & Kim, Y. (2025). *The semantic hub hypothesis: Language models share semantic representations across languages and modalities.* Available online: https://arxiv.org/abs/2411.04986 (accessed on).

Xiang, B., Yang, C., Li, Y., Warstadt, A., & Kann, K. (2021, April). CLiMP: A Benchmark for Chinese Language Model Evaluation. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 2784–2790). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.eacl-main.242/ (accessed on). https://doi.org/10.18653/v1/2021.eacl-main.242.

Xie, W., Feng, Y., Gu, S., & Yu, D. (2021, August). Importance-based Neuron Allocation for Multilingual Neural Machine Translation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 5725–5737). Online: Association for Computational Linguistics. Available online: https://aclanthology.org/2021.acl-long.445/ (accessed on). https://doi.org/10.18653/v1/2021.acl-long.445.

Yu, Z., & Ananiadou, S. (2024, November). Neuron-Level Knowledge Attribution in Large Language Models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 3267–3280). Miami, Florida, USA: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.emnlp-main.191/ (accessed on). https://doi.org/10.18653/v1/2024.emnlp-main.191.

Zeng, H., Han, S., Chen, L., & Yu, K. (2025, January). Converging to a Lingua Franca: Evolution of Linguistic Regions and Semantics Alignment in Multilingual Large Language Models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 10602–10617). Abu Dhabi, UAE: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.coling-main.707/ (accessed on).

Zhang, C., Lu, J., Tran, V. Q., Schuster, T., Metzler, D., & Lin, J. (2025, April). Tomato, Tomahto, Tomate: Do Multilingual Language Models Understand Based on Subword-Level Semantic Concepts? In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Findings of the association for computational linguistics: Naacl 2025* (pp. 1821–1837). Albuquerque, New Mexico: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.findings-naacl.98/ (accessed on). https://doi.org/10.18653/v1/2025.findings-naacl.98.

Zhang, F., & Nanda, N. (2024). Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *The twelfth international conference on learning representations*. Available online: https://openreview.net/forum?id=Hf17y6u9BC (accessed on).

Zhang, H., Shang, C., Wang, S., Zhang, D., Yu, Y., Yao, F., Sun, R., Yang, Y., & Wei, F. (2025, July). ShifCon: Enhancing Non-Dominant Language Capabilities with a Shift-based Multilingual Contrastive Framework. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4818–4841). Vienna, Austria: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.acl-long.239/ (accessed on). https://doi.org/10.18653/v1/2025.acl-long.239.

Zhang, H., Zhang, Z., Wang, M., Su, Z., Wang, Y., Wang, Q., Yuan, S., Nie, E., Duan, X., Han, F., Xue, Q., Yu, Z., Shang, C., Liang, X., Xiong, J., Shen, H., Tao, C., Liu, Z., Jin, S., … Wong, N. (2026). *Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models.* Available online: https://arxiv.org/abs/2601.14004 (accessed on).

Zhang, R., Yu, Q., Zang, M., Eickhoff, C., & Pavlick, E. (2025). The Same but Different: Structural Similarities and Differences in Multilingual Language Modeling. In *The thirteenth international conference on learning representations*. Available online: https://openreview.net/forum?id=NCrFA7dq8T (accessed on).

Zhang, X., Liang, Y., Meng, F., Zhang, S., Chen, Y., Xu, J., & Zhou, J. (2025, January). Multilingual Knowledge Editing with Language-Agnostic Factual Neurons. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 5775–5788). Abu Dhabi, UAE: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.coling-main.385/ (accessed on).

Zhang, Z., Zhao, J., Zhang, Q., Gui, T., & Huang, X. (2024, August). Unveiling Linguistic Regions in Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6228–6247). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.acl-long.338/ (accessed on). https://doi.org/10.18653/v1/2024.acl-long.338.

 doi:10.20944/preprints202607.1441.v1

Zhao, N., Duan, X., & Cai, Z. G. (2025, 09). The Missing Half of Language Learning in Current Developmental Language Models: Exogenous and Endogenous Linguistic Input. *Open Mind*, *9*, 1543-1549. Available online: https://doi.org/10.1162/OPMI.a.33 (accessed on). https://doi.org/10.1162/OPMI.a.33.

Zhao, Y., Zhang, W., Chen, G., Kawaguchi, K., & Bing, L. (2024). How do Large Language Models Handle Multilingualism? In *The thirty-eighth annual conference on neural information processing systems.* Available online: https://openreview.net/forum?id=ctXYOoAgRy (accessed on).

Zhong, C., Cheng, F., Liu, Q., Jiang, J., Wan, Z., Chu, C., Murawaki, Y., & Kurohashi, S. (2024). *Beyond english-centric llms: What language do multilingual language models think in?* Available online: https://arxiv.org/abs/2408.10811 (accessed on).

Zhong, C., Cheng, F., Liu, Q., Murawaki, Y., Chu, C., & Kurohashi, S. (2025). *Language lives in sparse dimensions: Toward interpretable and efficient multilingual control for large language models.* Available online: https://arxiv.org/abs/2510.07213 (accessed on).

Zhou, E., & Salhan, S. (2025, November). Extended Abstract for "Linguistic Universals": Emergent Shared Features in Independent Monolingual Language Models via Sparse Autoencoders. In D. I. Adelani et al. (Eds.), *Proceedings of the 5th workshop on multilingual representation learning (mrl 2025)* (pp. 128–130). Suzhuo, China: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.mrl-main.9/ (accessed on). https://doi.org/10.18653/v1/2025.mrl-main.9.

Zhou, X., Chen, D., Cahyawijaya, S., Duan, X., & Cai, Z. (2025, January). Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 6866–6888). Abu Dhabi, UAE: Association for Computational Linguistics. Available online: https://aclanthology.org/2025.coling-main.459/ (accessed on).

Zhu, S., Pan, L., Li, B., & Xiong, D. (2024, August). LANDeRMT: Dectecting and Routing Language-Aware Neurons for Selectively Finetuning LLMs to Machine Translation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 12135–12148). Bangkok, Thailand: Association for Computational Linguistics. Available online: https://aclanthology.org/2024.acl-long.656/ (accessed on). https://doi.org/10.18653/v1/2024.acl-long.656.

Üveges, I., & Ring, O. (2025). Evaluating the Impact of Synthetic Data on Emotion Classification: A Linguistic and Structural Analysis. *Information*, *16*(4). Available online: https://www.mdpi.com/2078-2489/16/4/330 (accessed on). https://doi.org/10.3390/info16040330.