Layer-Wise Distinction of Slang Types in Large Language Models

Anonymous ACL submission

Abstract

Slang is often treated as a marginal or stylistic phenomenon in NLP, yet it poses a systematic challenge to language models by simultaneously involving lexical novelty, constructional meaning, and contextual semantic shift. In this work, we propose a linguistically motivated typology of slang—neologisms, constructional slang, and semantic shifts—and investigate whether these categories exhibit differential processing patterns across layers of large language models (LLMs). Using a controlled dataset with literal counterparts and a mechanistic probing framework based on sparse autoencoders, we analyze layer-wise representation differences between slang and non-slang usages in an open-weight transformer model. Our results reveal two key findings. First, slang types differ in where representational divergence peaks: semantic shift slang peaks in deep layers (Layer 46), while neologisms and constructional slang both peak in middle layers (Layer 24), suggesting that meaning disambiguation engages deeper representational resources than lexical or constructional novelty. Second, slang types differ in magnitude: neologisms exhibit the highest divergence from literal counterparts across all layers, while constructional slang shows the lowest, reflecting distinct processing challenges —unfamiliar surface forms versus familiar constituents with non-compositional meaning. Rather than assuming a fixed mapping between layers and linguistic modules, our findings point to an emergent alignment where peak location and magnitude capture orthogonal aspects of slang processing. This study highlights the value of linguistically grounded categorization for understanding how LLMs process non-standard language and contributes to ongoing discussions on the interpretability of neural language models.

1 Introduction

Large Language Models (LLMs) have demonstrated an exceptional ability to generate fluent and contextually appropriate language across a wide range of registers. Aside from literal and formal usage, LLMs are also increasingly proficient at producing sentences with nonliteral meanings, stylistic markings, and context dependencies. This advancement raises a question for mechanistic interpretability: how are complex, context-dependent language phenomena internally represented in LLMs? While prior work has distinguished where linguistic features such as syntax or semantics emerge within models, little research has examined language phenomena that combine lexical novelty, pragmatics, and social context indexing.

One such phenomenon is slang. Slang words often stem from assigning novel meanings to existing lexical items that diverge from their literal meanings. The interpretation of slang words oftentimes depends on discourse context, social register, and constructions. Linguists have therefore characterized slang not as a singular lexical category, but a multi-dimensional phenomenon combining semantic shift, persona indexing, and constructional meaning.

Despite slang's linguistic complexity, it has rarely been used to study internal representations of LLMs. While existing work treats slang as a generation or detection problem, this study builds on previous research on contextual semantic integration and asks whether different types of slang exhibit distinct layer-wise representations, essentially treating slang as a probe for contextual semantic integration.

Based on prior findings that early layers capture lexical features while deeper layers handle semantic integration (Jing et al., 2025), we hypothesize that different slang types may engage representational resources at different depths. Specifically, we predict that neologisms—as novel lexical forms—may show early divergence, constructional slang may peak in middle layers where syntactic templates are processed, and semantic shift slang may require deep-layer processing for meaning disambiguation. We additionally examine whether slang types differ in the overall magnitude of representational divergence, which may reflect qualitatively different processing challenges.

To evaluate these hypotheses, we conduct an analysis of slang and literal usage in transformer-based language models using sparse autoencoders (SAEs) as an interpretability lens. SAEs provide a sparse coordinate system for analyzing representational features within LLMs, enabling systematic comparison of slang–literal divergence across layers.

2 Related Works

2.1 Layer-wise Linguistic Representations in LLMs

A growing body of work investigates how linguistic information is distributed across transformer layers. Early probing studies on BERT established that different layers tend to capture different types of linguistic information: surface features in lower layers, syntactic information in middle layers, and semantic features in upper layers (Jawahar et al., 2019; Tenney et al., 2019). Rogers et al. (2020) provide a comprehensive survey of these findings, while cautioning that probe accuracy reflects information *accessibility* rather than confirming that models actively *use* such information during processing.

More recent work extends this line of inquiry to larger decoder-based models. Jing et al. (2025) propose LinguaLens, a framework for analyzing linguistic mechanisms using sparse autoencoders, finding that features related to morphology and syntax tend to show stronger activation in earlier layers, while semantic and pragmatic features emerge more prominently in deeper layers. Rather than assuming a strict mapping between layers and linguistic modules, we treat such layer-wise patterns as emergent tendencies that may vary across linguistic phenomena—a perspective we apply to our analysis of slang.

2.2 Sparse Autoencoders for Mechanistic Interpretability

Traditional probing classifiers face well-documented limitations: high probe accuracy may reflect the probe's own capacity rather than genuine model representations, and probes reveal correlation rather than causation (Hewitt and Liang, 2019; Belinkov, 2022). A further challenge is polysemanticity, where individual neurons respond to multiple unrelated concepts (Elhage et al., 2022), complicating neuron-level analysis.

Sparse Autoencoders (SAEs) offer a complementary approach by decomposing dense activations into sparse, more interpretable feature spaces (Bricken et al., 2023; Cunningham et al., 2024). Under a sparsity constraint, SAEs learn projections where each dimension ideally captures a single semantic concept, mitigating polysemanticity concerns. Jing et al. (2025) demonstrate that SAE-extracted features can capture linguistic competencies and enable controlled output steering through targeted interventions.

For our study, SAEs provide a principled method to compare representational differences between slang and literal usages. We select Gemma-3-12B-IT paired with Gemma-Scope-2 SAEs, prioritizing interpretability and reproducibility over raw performance, as open-weight models with publicly available SAEs enable systematic mechanistic analysis.

2.3 Computational Approaches to Slang

Slang presents unique challenges for NLP systems due to its flexible semantics, rapid evolution, and context-dependent interpretation.

2.3.1 Detection and Identification.

Pei et al. (2019) formulate slang processing as a sequence labeling task, identifying Part-of-Speech transformation as a salient feature—slang words are twice as likely to undergo syntactic category shifts compared to standard vocabulary. Their distinction between newly extended senses and newly created words (neologisms) informs our typology.

2.3.2 Generation and Interpretation.

Sun et al. (2021) develop a computational framework modeling speaker word choice by relating conventional and slang senses. Their approach successfully predicts historical emergence of slang, demonstrating that slang follows principled patterns of semantic extension. Follow-up work (Sun et al., 2022) models slang interpretation through contrastive semantic spaces.

2.3.3 Semantic Shift and Sense Competition.

A key mechanism underlying slang interpretation is the suppression of dominant literal meanings in favor of novel slang senses. This process parallels research on lexical ambiguity resolution, where context must inhibit competing word senses (Rodd et al., 2005). For slang terms like *cook* or *slay*, successful interpretation requires suppressing the dominant conventional meaning—a deeper semantic operation than simply recognizing a novel lexical form. Sun and Xu (2022) trace how slang senses compete with and sometimes displace conventional meanings over time, suggesting that semantic shift slang may engage representational resources associated with meaning disambiguation.

2.3.4 Benchmarking LLM Knowledge.

Sun and Xu (2024) systematically evaluate LLM knowledge of slang across detection, identification, and interpretation tasks, finding significant performance variation across slang types. This motivates our investigation into whether such variations reflect underlying differences in layer-wise representations.

2.4 Functional Localization in Neural Language Models

Research on functional localization has identified neural components associated with particular linguistic functions. Tang et al. (2024) introduce Language Activation Probability Entropy (LAPE) to identify language-specific neurons, finding that multilingual proficiency concentrates in a small subset of neurons in top and bottom layers. Critically, selectively activating or deactivating these neurons can steer output language.

Extending this to stylistic variation, Lai et al. (2024) identify style-specific neurons for text style transfer, demonstrating causal relationships between neuron activation and stylistic output through targeted deactivation. However, they note that such interventions can impact fluency, highlighting complex interactions between representational components.

These studies focus on relatively stable markers such as language identity or formal register. Our work extends this paradigm to slang, which presents a distinct challenge: the same lexical form must activate different representational pathways based on contextual factors. Furthermore, rather than analyzing individual neurons, we leverage SAEs to examine distributed feature representations, providing finer-grained analysis of how slang processing organizes across layers. We frame our findings as observed tendencies rather than architectural claims, acknowledging that layer-wise patterns reflect emergent properties that may not generalize across all model families.

3 Methodology

The methodology of this study investigates whether linguistically distinct types of slang exhibit systematic differences in how they are represented across the layers of a large language model. We treat slang as a probe for contextual semantic integration and examine how slang–literal distinctions emerge within layers. To accomplish this, we construct controlled slang–literal minimal pairs, extract layer-wise representations from a decoder-based transformer, project these representations into a sparse latent space using a pretrained sparse autoencoder (SAE), and quantify representational separation across layers and slang categories.

3.1 Slang Typology and Data Construction

This study employs a linguistically motivated typology consisting of three categories: neologisms, constructional slang, and semantic shift. For each category, this study compiles a dataset pairing slang instances with literal counterparts that preserve meaning while removing slang-specific cues. Dataset sizes vary by category due to availability constraints (see Section 4.2).

3.1.1 Semantic Shift

Semantic-shift slang consists of lexical items whose conventional dictionary meanings are repurposed in informal contexts (e.g., *cook*, *slay*). Instances are drawn from an existing Gen Z slang pair dataset, where the same lexical form appears in both literal and slang contexts. Literal counterparts are provided by the dataset and each entry is further validated through cross-referencing with Green's Dictionary of Slang and Urban Dictionary to ensure consistency.

3.1.2 Neologisms

Neologisms are newly coined lexical items without established dictionary entries at the time of collection (e.g., *rizz*). Instances are sourced from an existing Gen Z slang dataset and filtered to exclude proper nouns, typos, and entries found as headwords in standard dictionaries (e.g. Oxford, Merriam-Webster) at the time of collection. This study additionally requires agreement between dataset-provided glosses and Urban Dictionary definitions. Since no conventional literal form exists for neologisms, we generate paraphrases by prompting a large language model to rephrase the provided gloss into standard English while preserving meaning and removing slang-specific cues.

3.1.3 Constructional Slang

Constructional slang consists of multi-word schematic expressions whose meanings are not fully derivable from the constituents (e.g., *It's giving X*). This study extracts multi-word slang expressions from an Urban Dictionary data dump and uses LLM-assisted filtering to distinguish productive constructional patterns from fixed idioms. For each constructional instance, we generate paraphrases of example sentences that preserve content while removing the constructions.

3.2 Model and Representations

This paper analyzes the slang representations from **Gemma-3-12B-IT** (Team, 2025), an open source, decoder-based transformer language model. This paper chooses this model for its reproducibility for mechanistic analysis. To facilitate interpretability, this paper uses **Gemma-Scope-2**, a pre-trained sparse autoencoder aligned with the Gemma architecture (McDougall et al., 2025).

For each input instance, this study extracts layer-wise hidden representations at a fixed hook location corresponding to the residual stream. Following standard practice in sentence-level analysis, representations are pooled at the final token position to capture the model's integrated sentence representation. While this approach may smooth over token-level details, it provides a consistent comparison framework across slang types with varying surface forms. Importantly, this methodology does not assume that individual SAE features correspond to discrete linguistic units. Instead, the SAE provides a sparse coordinate system that enables systematic comparison of representational structure across layers.

3.3 Slang-Literal Separation Analysis

Our primary analysis measures where in the model slang and literal usages diverge most strongly. For each layer, we project hidden representations into the SAE feature space and compute a slang–literal separation score, defined as the aggregated activation difference between paired slang and literal instances. We then compare separation profiles across layers and across slang categories.

This analysis allows us to assess whether different types of slang exhibit distinct depth-wise patterns of representational divergence, without assuming a strict mapping between layers and linguistic modules. Rather than interpreting separation as evidence of discrete processing stages, we treat observed layer-wise trends as emergent tendencies in how the model integrates lexical, constructional, and contextual semantic information.

4 Experiment and Analyses

4.1 Experimental Setup

This paper evaluates layer-wise representational differences between slang and literal usages across three slang categories: semantic shift, neologisms, and constructional slang. All experiments are conducted using Gemma-3-12B-IT, a 48-layer decoder-based transformer (Team, 2025). Representations are extracted from each transformer block at a fixed hook location corresponding to the residual stream.

To conduct systematic comparison across layers, this paper projects hidden representations into the latent space of Gemma-Scope-2, a pre-trained sparse autoencoder aligned with the target model (McDougall et al., 2025). The SAE provides a high-dimensional but sparse feature basis that enables comparison of representational structure without relying on supervised probes.

For each input instance, this paper extracts the representation of the slang sentence versus the literal sentences. All analyses are conducted independently for each layer.

4.2 Datasets Overview

Three datasets were compiled corresponding to our slang taxonomy, each containing paired slang and literal sentences. (See Table 1)

Category	N	Source
Semantic Shift	1002	Kaggle
Neologism	1000	Urban Dictionary & HuggingFace
Constructional	37	Urban Dictionary (filtered)

Table 1: Dataset Overview

Semantic Shift pairs were sourced from a Kaggle dataset of Gen-Z slang with provided literal counterparts (Gamage, 2025). **Neologism** pairs were constructed by extracting novel lexical items from a HuggingFace dataset, accompanied by extracting novel lexical items from Urban Dictionary that lack entries in standard dictionaries, with literal paraphrases generated via LLM and verified against Urban Dictionary glosses. **Constructional Slang** examples were filtered from Urban Dictionary to retain only productive syntactic templates (e.g., *X is giving Y*) rather than fixed idioms, with literal paraphrases generated for each instance. The smaller size of the constructional slang dataset reflects the relative scarcity of productive slang constructions compared to single-word slang.

4.3 Feature Extraction

For each sentence pair, this experiment:

1. Tokenizes both slang and literal sentences
2. Extracts activations from all 48 layers via forward hooks on the residual stream
3. Projects activations through the corresponding layer's SAE encoder to acquire sparse feature vectors
4. Pools representations at the final token position, capturing the model's integrated sentence representation
5. Computes distance metrics between slang and literal feature vectors

This study computes two complementary distance metrics: L2 (Euclidean distance), which captures total magnitude of activation difference, and cosine distance, which captures directional differences in feature space.

The cosine distance is particularly informative for our analysis as it measures *which* features differ rather than *how much* total activation differs, controlling for the natural increase in activation magnitude in deeper layers.

4.4 Analysis Framework

We partition the 48 layers into three regions following the LinguaLens framework (Jing et al., 2025): early layers (0-15), middle layers (16-31), deep layers (32-47). Early layers are hypothesized to be associated with lexical access and morphology, while middle and deep layers are claimed to associate with syntax and pragmatics, and semantics and rhetoric, respectively.

For each slang category, we identify the layer with peak representational divergence, region-wise mean divergence, and statistical significance of between-region differences.

4.5 Results

4.5.1 Layer-wise Activation Differences

Table 2 presents the peak layer for both L2 and cosine distance metrics, and Table 3 shows the region-wise means for both metrics mentioned above.

Category	Metric	Peak Layer	Depth
Semantic Shift	L2	46	97.9%
Semantic Shift	Cosine	46	97.9%
Neologism	L2	47	100%
Neologism	Cosine	24	51.1%
Construction	L2	47	100%
Construction	Cosine	24	51.1%

Table 2: Peak layer (argmax) and depth for each metric and slang category.

Category	Metric	Early	Mid	Deep
Semantic Shift	L2	269.1	2790.0	10660.6
Semantic Shift	Cosine	0.212	0.244	0.298
Neologism	L2	280.2	3423.1	11522.9
Neologism	Cosine	0.221	0.434	0.462
Construction	L2	183.4	2,520.5	8,991.5
Construction	Cosine	0.095	0.229	0.250

Table 3: Region-wise mean divergence (Early: 0–15; Middle: 16–31; Deep: 32–47).

As observed above, **all three slang categories show monotonically increasing L2 distance with layer depth**, peaking at layer 46-47. This reflects the cumulative nature of representational differences through the forward pass rather than localized processing.

On the other hand, when controlling for activation magnitude, semantic shift slang peaks in deep layers (Layer 46), whereas neologisms and constructions peak in middle layers (Layer 24). This dissociation indicates qualitatively different processing signatures.

Notably, for neologisms, the deep-layer region mean (0.462) exceeds the middle-layer mean (0.434), despite the global peak occurring at Layer 24. This reflects the shape of the activation curve: neologisms show a sharp peak at Layer 24, followed by a sustained plateau through deep layers. The peak identifies the layer of maximum divergence, while the region means capture average divergence across each third of the network.

4.5.2 Statistical Analysis

All cross-region differences were highly significant. (See Table 4)

Category	Metric	F-statistic	p-value
Semantic Shift	Cosine	21.81	$2.38 \times 10^{\{-7\}}$
Neologism	Cosine	43.51	$3.04 \times 10^{\{-11\}}$
Construction	Cosine	87.53	$3.09 \times 10^{\{-16\}}$

Table 4: One-way ANOVA comparing cosine distance across three layer regions (Early, Middle, Deep).

Post-hoc pairwise t-tests reveal distinct patterns:

Cat	Early vs Middle	Middle vs Deep	Early vs Deep
S	t = −2.18, p = .037	t = −4.56, p < .001	t = −6.81, p < .001
N	t = −6.51, p < .001	t = −1.51, p = .142	t = −7.69, p < .001
C	t = −9.24, p < .001	t = −2.06, p = .048	t = −11.87, p < .001

Table 5: Bonferroni-corrected pairwise t-tests on cosine distance between layer regions.

For **semantic shift**, all pairwise comparisons are significant, with the largest effect between middle and deep layers, consistent with the hypothesis that semantic disambiguation occurs in deeper layers.

For **neologisms**, the middle-to-deep transition is not significant (p = .142), indicating that representational divergence plateaus after middle layers. For **constructions**, the middle-to-deep difference reaches significance (p = .048), though the effect size is smaller than the early-to-middle transition. This suggests that representational divergence is largely established by middle layers for these categories.

4.5.3 Local Maxima Analysis

Examining local maxima in cosine distance reveals the processing signatures for each slang type.

Category	Early	Middle	Deep	Global
Semantic Shift	7, 11	24, 29	35, 37, 41, 43, 46	46 (deep)
Neologism	2, 8, 12, 15	18, 24, 26, 28	35, 38, 41, 43, 46	24 (middle)
Construction	2, 7, 13	18, 24, 28	35, 38, 40, 43, 46	24 (middle)

Table 6: Local maxima in cosine-distance separation across early, middle, and deep layer regions.

Semantic shift shows a clear progression with the global maximum in deep layers. Neologisms and constructions show a shared peak at Layer 24.

4.6 Discussion

4.6.1 Hypothesis Evaluation

The original hypothesis of this study predicted:

1. Neologisms: Early layers (lexical access)
2. Constructions: Middle layers (syntax-pragmatics)
3. Semantic Shift: Deep layers (semantic disambiguation)

Through the experiment, **constructions and semantic shift aligned with predictions**. Constructional slang shows peak divergence in middle layers where syntactic templates are parsed, while semantic shift peaks in deep layers consistent with the inhibition of literal meaning and contextual disambiguation.

However, neologisms deviated from the original hypothesis. Instead of early-layer processing, neologisms peaked in middle layers alongside constructions. This may reflect that neologisms in context often require pragmatic integration beyond simple lexical lookup. Hence, the model must infer meaning from surrounding context when encountering a previously unknown term.

4.6.2 Interpretation of Metrics

The dissociation between L2 and cosine distance is theoretically meaningful:

- **L2 distance** captures how much representations differ in total magnitude. Its monotonic increase through layers reflects the cumulative processing pipeline, not the locus of category-specific computation.
- **Cosine distance** captures which features are activated differently, independent of magnitude. Its category-specific peaks reveal where the model's representational geometry diverges for slang versus literal language.

The convergence of neologisms and constructions at Layer 24 suggests a shared processing stage—possibly the point where non-standard linguistic forms (whether lexical or structural) are recognized as requiring special interpretation. The subsequent divergence, with only semantic shift continuing to differentiate through deep layers, reflects the additional computational demand of inhibiting dominant literal meanings.

4.6.3 Magnitude Differences

Beyond peak location, slang types differ systematically in the overall magnitude of divergence. Neologisms show the highest cosine distance across all layers, while constructional slang shows the lowest. This pattern suggests that each slang type poses a qualitatively different processing challenge.

Neologisms, as entirely unfamiliar lexical forms, lack pre-existing representations; the model must construct representations from contextual cues alone, resulting in high divergence from literal paraphrases throughout the network. Semantic shift slang involves familiar lexical items whose conventional meanings are well-established; the challenge lies in sense selection rather than lexical access, producing moderate divergence that accumulates toward deeper layers. Constructional slang occupies an intermediate position: although the overall construction carries non-compositional meaning, its constituent words are individually familiar, allowing slang and literal representations to share substantial surface-level structure.

These patterns align with a view of slang processing as engaging qualitatively different mechanisms: lexical novelty detection for neologisms, sense competition and inhibition for semantic shifts, and compositional integration for constructions. From a practical standpoint, these findings suggest that neologisms may be most readily detectable as non-standard language given their high representational distinctiveness, whereas constructional slang—with its low divergence from literal paraphrases—may pose the greatest challenge for automatic slang identification systems.

5 Conclusion

This paper presented a layer-wise analysis of slang processing in large language models using sparse autoencoders. Our findings reveal two complementary patterns in how different slang types are represented.

First, slang types differ in where representational divergence peaks. Constructional slang peaks in middle layers, consistent with Jing et al.'s (2025) hypothesis that middle layers handle syntax-pragmatics processing. Semantic shift slang peaks in deep layers, reflecting the computational demand of meaning disambiguation and inhibition of dominant literal senses. Neologisms, contrary to our initial prediction of early-layer processing, also peak in middle layers alongside constructions, suggesting that novel lexical items require pragmatic inference beyond surface-level lexical access.

Second, slang types differ in the magnitude of divergence. Neologisms show the highest cosine distance across all layers, reflecting the challenge of processing entirely unfamiliar lexical forms. Constructional slang shows the lowest divergence, likely because its constituent words share surface-level structure with literal paraphrases. These magnitude differences suggest that peak location and overall divergence capture orthogonal aspects of slang processing—where processing consolidates versus how challenging the processing demand is.

These results contribute to our understanding of how LLMs handle non-standard language and provide a framework for interpreting the internal processing of linguistic creativity and variation.

6 Limitations

6.1 Sentence-level Pooling

We used the final token as a representative of the sentence, which might smooth over fine-grained details. Analyzing the specific position where slang terms occur could reveal more localized processing signatures.

6.2 Imbalance in Dataset Size

The constructional slang subset ($N = 37$) is substantially smaller than other categories. This scarcity potentially limits the statistical robustness and generalizability of findings regarding this linguistic feature. The low cosine distance observed for constructional slang could reflect genuine processing characteristics or could be partially attributable to limited data.

6.3 Single Model Analysis

The findings in this research were tested on and are currently restricted to **Gemma-3-12B-IT**. Whether these mechanistic behaviors hold across different model families and sizes requires further cross-architectural investigation.

6.4 Lack of Causal Evidence

While SAE features highlight where representations differ, the evidence is correlational and does not establish a causal link, leaving the exact mechanism unproven. Causal validation through activation steering or ablation studies would strengthen these findings.

6.5 Methodological Differences Across Categories

The three slang categories differ not only in linguistic properties but also in how literal counterparts were constructed. Semantic shift pairs use dataset-provided literal sentences with the same lexical form, while neologism and construction pairs use LLM-generated paraphrases. This methodological difference may contribute to observed magnitude differences, and the high divergence for neologisms should be interpreted with this caveat in mind.

References

Yonatan Belinkov. 2022. *Probing Classifiers: Promises, Shortcomings, and Advances*. Volume 48. 1. MIT Press. pages 207–219.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and others. 2023. *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*. Anthropic.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. *Sparse Autoencoders Find Highly Interpretable Features in Language Models*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and others. 2022. *Toy Models of Superposition*. Anthropic.

Ranuga Disansa Gamage. 2025. genz-slang-pairs-1k (Revision 9728f69).

John Hewitt and Percy Liang. 2019. *Designing and Interpreting Probes with Control Tasks*. Hong Kong, China: Association for Computational Linguistics. pages 2733–2743.

Ganesh Jawahar, Benoit Sagot, and Djamé Seddah. 2019. *What Does BERT Learn about the Structure of Language?* Florence, Italy: Association for Computational Linguistics. pages 3651–3657.

Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. *LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder*. arXiv:2502.20344. Association for Computational Linguistics.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. *Style-Specific Neurons for Steering LLMs in Text Style Transfer*. Miami, Florida, USA: Association for Computational Linguistics. pages 13427–13443.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. 2025. *Gemma Scope 2 - Technical Paper*. technical report. Technical report. Google.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. *Slang Detection and Identification*. Hong Kong, China: Association for Computational Linguistics. pages 881–889.

Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. 2005. *Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access*. Volume 53. 2. Elsevier. pages 181–201.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. *A Primer in BERTology: What We Know About How BERT Works*. Volume 8. MIT Press. pages 842–866.

Zhewei Sun and Yang Xu. 2022. *Tracing Semantic Variation in Slang*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. pages 1299–1313.

Zhewei Sun and Yang Xu. 2024. *Toward Informal Language Processing: Knowledge of Slang in Large Language Models*.

Zhewei Sun, Richard Zemel, and Yang Xu. 2021. *A Computational Framework for Slang Generation*. Volume 9. MIT Press. pages 462–478.

Zhewei Sun, Richard Zemel, and Yang Xu. 2022. *Semantically Informed Slang Interpretation*. Seattle, United States: Association for Computational Linguistics. pages 5213–5231.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. *Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models*. Bangkok, Thailand: Association for Computational Linguistics. pages 5701–5715.

Gemma Team. 2025. Gemma 3.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. *BERT Rediscovers the Classical NLP Pipeline*. Florence, Italy: Association for Computational Linguistics. pages 4593–4601.