Lexical Familiarity Predicts Processing Depth for Nonliteral Language in Large Language Models

Lang-Ching Yeh, Yu-Chieh Wang, Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University
{r14142008, r14142006, shukaihsieh}@ntu.edu.tw

Abstract

This paper investigates how large language models internally process nonliteral language. Analyzing five categories spanning slang, metaphor, and idioms across all 48 layers of Gemma-3-12B-IT with Gemma Scope 2 sparse autoencoders, we find a lexical familiarity gradient: processing depth depends on available prior lexical knowledge, not figurative type. Idioms diverge at L1 as entrenched units; expressions built from familiar words (metaphors, semantic-shift and constructional slang) converge at L7–9; neologisms peak at L41, activating 3× more unique features. Paraphrase residual analysis confirms strong signals only at the gradient endpoints, yielding a three-tier hierarchy of entrenched retrieval, known-word reanalysis, and novel-word construction. Crucially, this peak-layer structure replicates in base models (Gemma-PT, Qwen-Base), demonstrating that the gradient is a robust property of pretrained representations rather than an alignment artifact. We additionally identify an activation density confound in SAE feature counts that produces spurious cross-condition convergence. Overall, processing depth is better predicted by lexical familiarity than by figurative type, with implications for robustness to non-standard language and for SAE-based interpretability.

1 Introduction

Large language models process a wide range of language beyond literal, dictionary-attested usage. Slang, metaphor, idioms, and other nonliteral expressions pervade everyday communication, particularly in informal and digital contexts, yet the mechanisms by which models handle these forms remain underexplored. A growing body of work on mechanistic interpretability has mapped how linguistic information distributes across transformer layers, establishing that surface features tend to emerge early, while semantic and pragmatic features consolidate deeper (Jing et al., 2025). Most of this work, however, has focused on standard linguistic categories, such as syntax, semantics, coreference, and has not examined the internal processing of nonliteral language.

Understanding how models process nonliteral language matters for several interconnected reasons. Robustness depends on knowing where models are likely to fail; if novel slang is processed through qualitatively different mechanisms than familiar metaphor, this predicts different failure modes. Fairness depends on knowing whose language is harder to process; slang and novel vocabulary are not evenly distributed across demographic groups, and mechanistic evidence of elevated processing costs for certain language types points toward sources of systematic performance disparity. Interpretability depends on having reliable tools; if commonly used metrics in SAE-based analysis carry hidden confounds, the conclusions drawn from them may be unreliable.

We ask whether different types of nonliteral language exhibit distinct layer-wise processing signatures and, if so, what principle governs those differences. Prior work has analyzed slang as a detection or generation task (Sun et al., 2024) but has not examined how slang is represented internally throughout model layers. Work on figurative language has studied metaphor and idiom processing in isolation from slang, despite the shared challenge of meaning that deviates from surface form.

We conduct a unified layer-wise analysis of five categories of nonliteral language: semantic-shift slang, neologisms, constructional slang, metaphors, and idioms. Furthermore, two control conditions were analyzed (identical sentence pairs and literal paraphrase pairs from PAWS (Zhang et al., 2019)). For each category, we construct minimal pairs consisting of a figurative sentence and a literal counterpart, then compare the representations across all 48 layers of Gemma-3-12B-IT (Gemma Team, 2025) using Gemma Scope 2 sparse autoencoders

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 456–470

(McDougall et al., 2025).

Our central finding is a *lexical familiarity gradient*: the layer at which representational divergence peaks varies systematically with how much prior lexical knowledge the model can leverage. The gradient organizes into three tiers. Entrenched multi-word units (idioms) diverge immediately at L1. Expressions built from individually familiar words (metaphors, semantic-shift slang, and constructional slang) converge in a narrow early-layer band at L7–9. Novel lexical forms (neologisms) diverge deepest at L41. A literal paraphrase control peaks at L3, establishing a noise floor for surface-form variation. Residual analysis confirms strong divergence signals at the gradient endpoints (idioms and neologisms), while the middle tier does not clearly exceed the paraphrase noise floor. The strongest evidence therefore lies at the two extremes of the familiarity spectrum.

We additionally identify a *methodological confound* in SAE-based interpretability: raw feature counts peak at L24–25 for all conditions including controls, reflecting layer-wise activation density rather than linguistic processing.

This paper makes four contributions: (1) a unified analysis spanning slang, metaphor, and idioms under a common framework with matched controls; (2) a lexical familiarity gradient that organizes processing depth by divergence type rather than figurative category; (3) residual analysis quantifying how much of observed divergence reflects genuine figurative processing versus surface-form variation; and (4) identification of an activation density confound in SAE feature counts relevant to interpretability research broadly.

2 Related Work

2.1 From Probing to Sparse Autoencoders

Work on layer-wise linguistic structure began with BERT probing, which suggested a coarse hierarchy: lower layers encode surface and morphological features, middle layers encode syntax, and upper layers capture semantics (Jawahar et al., 2019; Tenney et al., 2019). However, probing accuracy can reflect probe capacity rather than what the model uses (Rogers et al., 2020), and control tasks are needed to separate genuine representations from probe expressiveness (Hewitt and Liang, 2019).

These concerns motivated the development of sparse autoencoders (SAEs) as an alternative interpretability tool. Instead of asking whether a classifier can extract feature X from layer l, SAEs decompose dense activations into sparse, overcomplete feature dictionaries where each dimension aims to align with a single interpretable concept (Bricken et al., 2023; Cunningham et al., 2024), mitigating neuron polysemanticity, in which neurons respond to multiple concepts, making neuron-level analysis unreliable. (Elhage et al., 2022). Using SAEs, LinguaLens (Jing et al., 2025) show that the morphology-to-semantics hierarchy extends to autoregressive models. Building on this framework, we test whether the hierarchy holds for nonliteral language, where inputs deviate from standard usage, extending the functional localization paradigm from stable linguistic markers (Tang et al., 2024; Lai et al., 2024) to the more fluid domain of nonliteral language.

Computational work on slang has developed largely in parallel with work on figurative language. Pei et al. (2019) treat slang processing as sequence labeling and find that slang words are twice as likely to undergo syntactic category shifts compared to standard vocabulary. Sun et al. (2021) model slang generation through contrastive semantic spaces, and subsequent work traces how slang senses compete with conventional meanings over time (Sun and Xu, 2022; Sun et al., 2022). More recently, Sun et al. (2024) evaluate LLM knowledge of slang across detection, identification, and interpretation tasks, finding that performance varies substantially across slang types. This variation raises the question of whether the differences are merely behavioral or reflect differences in internal representation, a question our study addresses directly.

On the figurative language side, the FLUTE benchmark (Chakrabarty et al., 2022) provides matched figurative and literal sentence pairs for metaphors and other figures, allowing controlled comparison. Computational studies on idioms show that transformer models may over-apply compositional processing to idioms, but exhibit more unit-like behavior when successful (Dankers et al., 2022), and that LLM's understanding of idioms shows an interplay between memorized knowledge and context-driven inference (Kim et al., 2025). Oh et al. (2026) show that figurative idiom meaning is retrieved in early attention and MLP sub-layers, with a small set of attention heads selectively boosting the figurative interpretation while suppressing the literal one.

Despite this parallel literature, there is limited work that evaluates slang and figurative language

Category	Type	N	Source
Semantic Shift	Slang	1,002	Gen-Z Pairs
Neologism	Slang	1,000	UD + HF
Construction†	Slang	37	UD + HF
Metaphor	Fig.	625	FLUTE
Idiom	Fig.	823	PIE corpus
Identical	Ctrl	1,000	Sampled
Lit. Paraphrase	Ctrl	1,000	PAWS

Table 1: Dataset overview. †Construction results are exploratory given the small sample size. Control conditions contain no figurative language. UD = Urban Dictionary; HF = HuggingFace.

together under controlled conditions. Both languages involve systematic departures from literal, surface-form expectations, yet they are typically studied with different datasets and evaluation setups. Our work bridges the gap by analyzing both under a common methodology and showing that the same organizing principle (lexical familiarity) governs the processing depth across both domains.

3 Methodology

3.1 Overview

We compare paired figurative and literal sentences across five categories of nonliteral language, validated against two baseline conditions. For each pair, we extract token-level representations from every layer of a transformer model, project them into a sparse feature space using pre-trained SAEs, and measure divergence across layers.

3.2 Categories and Data Construction

Table 1 summarizes our datasets. Each category consists of sentence pairs where a nonliteral sentence is paired with a literal counterpart that preserves meaning while using standard vocabulary.

Semantic-shift slang. Semantic-shift slang consists of existing words used with novel informal meanings (e.g., *slay* meaning "to perform excellently"). Pairs are drawn from a Kaggle dataset of Gen-Z slang (Gamage, 2025). For each item, the slang sentence contains a slang expression, and the paired literal sentence provides a meaning-matched paraphrase, which may involve lexical substitution (e.g., I'm really drained / I'm really tired). This design targets semantic reinterpretation in context rather than surface-form identity. Each entry was validated against Green's Dictionary of Slang (Green, 2025) and Urban Dictionary (Urban Dictionary, 2025).

Neologisms. Neologisms are newly coined words without established dictionary entries (e.g., *rizz* meaning "charisma"). We identified 1,000 candidates by cross-referencing LM-Lexicon slang dataset (LM-Lexicon, 2025) with publicly available Urban Dictionary data dump (Kulkarni, 2025), filtering by gloss agreement and upvote count, and excluding dictionary-attested terms. Literal counterparts were generated via LLM prompting and manually reviewed for semantic fidelity (see Appendix A). Neologisms are typically segmented into multiple subword units by the tokenizer; this segmentation is itself a signature of lexical unfamiliarity.

Constructional slang. Constructional slang consists of multi-word schematic expressions whose meanings are not fully derivable from their constituents (e.g., *It's giving X*). Candidates were identified using the same cross-referencing procedure, with additional LLM-assisted filtering to retain only entries satisfying three criteria: (1) multi-word (more than one token), (2) schematic (exhibiting a fillable pattern), and (3) non-compositional (meaning not derivable from parts alone). This two-stage process yielded 37 items. Given the small sample, we retain this category for completeness but interpret its results with caution and mark it with † throughout.

Metaphor. Metaphor pairs are drawn from the FLUTE benchmark (Chakrabarty et al., 2022), which provides figurative sentences paired with literal counterparts and textual explanations. Metaphor serves as a well-studied baseline for figurative language involving familiar words in non-standard combinations.

Idiom. Idiom pairs are drawn from the PIE corpus (Zhou et al., 2021), a parallel collection of idiomatic expressions paired with their literal paraphrases. Idioms represent the most lexicalized end of the figurative spectrum: multi-word units whose meaning is conventionally stored rather than composed from constituent parts.

3.3 Control Conditions

We introduce two controls essential for interpreting the results, both drawn from PAWS (Zhang et al., 2019). **Identical pairs** (N=1,000) duplicate the same sentence as both inputs; any non-zero distance indicates pipeline noise. **Literal paraphrase pairs** (N=1,000) are meaning-equivalent sentences

differing only in surface form, establishing a *noise floor* for cosine distance attributable to lexical variation alone. Figurative categories must exceed this floor to claim a genuine processing signal.

3.4 Model and Representation Extraction

We analyze Gemma-3-12B-IT (Gemma Team, 2025), a 48-layer open-weight decoder-only transformer, paired with Gemma Scope 2 (McDougall et al., 2025), a suite of layer-aligned sparse autoencoders (SAEs) for Gemma. We selected this combination for reproducibility and the public availability of both model weights and SAE parameters.

For each figurative–literal sentence pair, we (1) tokenize both sentences, (2) identify the annotated token region where the two sequences differ (the *divergence span*), (3) extract residual-stream hidden states from all 48 layers at those token positions, and (4) project the pooled layer activations through the corresponding layer's SAE encoder. For semantic-shift slang, the divergence span is typically the slang word itself (e.g., *slay*); for neologisms, it is the novel term; and for idioms, metaphors, and constructions it is the full substitution region, which may cover multiple tokens. Each dataset provides pre-annotated token indices [1] to ensure consistent extraction across examples.

Span-based extraction targets the localized portion of the input where nonliteral meaning is introduced, while allowing multi-token expressions to contribute distributed evidence. Divergence span lengths varied across categories (Table 6); the relationship between span length and the familiarity gradient is examined in §4.5.

3.5 Metrics

We use three complementary metrics, each addressing a different aspect of representational divergence.

Cosine distance (primary metric) captures the directional difference between SAE feature vectors, independent of magnitude:

$$d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \quad (1)$$

Because cosine distance normalizes by vector magnitude, it is robust to the layer-wise variation in total activation that confounds raw feature counts (see §4.2).

[1] See Appendix B for dataset column specifications and annotation format.

Category	Peak	Value	95% CI	J
Idiom	L1	.830	[.819, .841]	.21
Lit. Para.	L3	.428	[.409, .446]	.33
Constr.†	L7	.659	[.602, .716]	.24
Metaphor	L8	.474	[.453, .496]	.30
Sem. Shift	L9	.478	[.467, .488]	.33
Neologism	L41	.857	[.847, .867]	.15
Identical	–	.000	–	1.00

Table 2: Peak cosine distance layers with 95% confidence intervals and mean Jaccard feature overlap (J). †Construction results are exploratory given small sample size (N=37).

Feature overlap ratio. For paired sentences A and B with active feature sets $\mathcal{F}_A$ and $\mathcal{F}_B$, we compute a Jaccard-like measure of feature sharing:

$$J(\mathcal{F}_A, \mathcal{F}_B) = \frac{|\mathcal{F}_A \cap \mathcal{F}_B|}{|\mathcal{F}_A \cup \mathcal{F}_B|} \quad (2)$$

Low overlap indicates that the two sentences activate largely disjoint feature sets. Unlike raw feature counts, this ratio is naturally normalized.

Figurative-to-literal feature ratio. We count features active only in the figurative sentence ("figurative-only") and features active only in the literal sentence ("literal-only"), and report their ratio per region. Values above 1 indicate more unique representational dimensions recruited for the figurative expression; values below 1 indicate more for the literal counterpart.

Following Jing et al. (2025), we partition layers into early (0–15), middle (16–31), and deep (32–47) regions.

4 Results

4.1 Lexical Familiarity Gradient

Table 2 reports peak cosine distance layers and Table 3 reports region-wise means. The peaks form a gradient organized by lexical familiarity.

All reported values include 95% confidence intervals (CI = $\bar{x} \pm 1.96 \cdot \sigma/\sqrt{n}$).

Idioms (L1, cosine = 0.830). Divergence is highest at the very first layer and decreases through the network. The model appears to recognize idiomatic expressions as stored multi-word units immediately, producing maximal representational separation before any compositional processing occurs. This early peak aligns with recent evidence that figurative interpretations are supported by early sublayers and specific attention heads (Oh et al., 2026).

Category	Early	Mid	Deep
Idiom	**.697** [.694,.700]	.451 [.448,.454]	.492 [.489,.495]
Constr.†	**.525** [.508,.541]	.377 [.363,.391]	.445 [.429,.460]
Metaphor	**.369** [.364,.374]	.334 [.330,.337]	.330 [.326,.334]
Sem. Shift	**.325** [.323,.328]	.244 [.241,.246]	.314 [.311,.316]
Neologism	.598 [.595,.600]	.481 [.479,.484]	**.661** [.658,.663]
Lit. Para.	**.370** [.365,.374]	.285 [.282,.288]	.325 [.322,.329]

Table 3: Region-wise mean cosine distance with 95% CIs. Bold marks the highest region per category. The literal paraphrase baseline establishes a noise floor: metaphor and semantic-shift CIs overlap with this baseline in all regions.

The declining profile suggests that later layers contribute less to this distinction.

Known-word figurative language (L7–9). Constructional slang† (L7), metaphor (L8), and semantic-shift slang (L9) all peak within a narrow early-layer band. These categories share individually familiar constituent words; nonliteral meaning arises from contextual reanalysis of known lexical material rather than from encountering an unknown form.

Neologisms (L41, cosine = 0.857). Divergence peaks in deep layers and neologism is the only category whose highest region-wise mean falls in the deep region (0.661). Lacking prior lexical representations, the model must build meaning from context over many layers. The representational cost is substantial: at peak, neologisms activate a mean of 104.7 figurative-only features, approximately 2× higher than any other category (idiom: 55.4, construction: 56.4, metaphor: 41.1, semantic shift: 37.8) and the literal paraphrase control (43.3).

This gradient of entrenched-unit retrieval (L1) → familiar-word reanalysis (L7–9) → novel-word inference (L41) suggests that lexical familiarity, not figurative type, determines processing depth.

The Jaccard feature overlap (J column, Table 2) provides a complementary view. Identical pairs show perfect overlap (J = 1.00), confirming no pipeline noise. Neologisms show the lowest overlap (J = 0.15), meaning figurative and literal sentences activate largely disjoint feature sets. This result is consistent with the model building an entirely novel representation. Idioms are also low (J = 0.21), reflecting the representational gap between a stored unit and its compositional counterpart.

Category	Res. Peak	Early	Mid	Deep
Neologism	L41 (0.505)	0.228	0.196	**0.335**
Idiom	L1 (0.409)	**0.327**	0.166	0.166
Constr.†	L7 (0.284)	0.155	0.092	0.119
Metaphor	L29 (0.126)	–0.001	0.049	0.004
Sem. Shift	L9 (0.099)	–0.044	–0.041	–0.012

Table 4: Residual cosine distance (category minus literal paraphrase baseline) by region. Negative or near-zero values indicate the category does not clearly exceed the paraphrase noise floor.

4.2 SAE Activation Density Confound

An initial observation is that figurative-only feature counts peak at L24–25 regardless of category. This apparent convergence might suggest a universal nonliteral-language detection mechanism. Our controls, however, reveal it as an artifact.

The **identical baseline** produces exactly 0.000 cosine distance at every layer, confirming no pipeline noise. Yet shared features (active in both copies) peak at L25 with 171.1 features, higher than any figurative category. The SAE simply fires the most features at layers 24–25 for any input.

The **literal paraphrase baseline** confirms this: features unique to one paraphrase peak at L24 with 43.3 features, comparable to figurative categories, despite containing no figurative language.

This confound motivates our adoption of cosine distance as the primary metric. Researchers using SAE feature counts for interpretability should control for layer-wise activation density, as raw counts can produce spurious convergence across conditions that share no meaningful signal.

4.3 Residual Analysis Against Paraphrase Baseline

To isolate figurative processing signals from surface-form variation, we subtract the paraphrase baseline's cosine distance profile from each category (Table 4). Three tiers of signal strength emerge.

Strong signal (neologism, idiom). Both show large positive residuals across all regions, with peak residuals of 0.505 and 0.409 respectively. Peak layers (L41 for neologism, L1 for idiom) are unchanged after baseline subtraction. These categories produce divergence that clearly exceeds surface-form variation.

Moderate signal (construction†). Residuals are positive across all regions (peak: 0.284 at L7), indi-

Category	Early	Mid	Deep
Neologism	3.23	3.21	2.07
Metaphor	1.37	1.24	1.14
Sem. Shift	0.96	1.30	0.81
Constr.†	0.72	1.18	0.90
Idiom	0.72	0.81	0.70

Table 5: Figurative-only / literal-only feature ratio by region. Values >1 indicate more unique features in the figurative sentence.

cating a genuine signal above the paraphrase floor, though of smaller magnitude. Given the small sample size (N = 37), this should be interpreted cautiously.

Weak signal (metaphor, semantic shift). Residuals are near zero or negative in most regions. For these categories, much of the observed divergence may reflect surface-form variation rather than figurative meaning processing per se. The ordering of peak layers is preserved even for weak-signal categories: metaphor and semantic shift still peak earlier than neologisms. However, the magnitude of their signal does not clearly exceed what would be expected from surface variation alone. Stronger claims apply to the gradient endpoints than the middle.

4.4 Figurative-to-Literal Feature Ratio

The ratio of figurative-only to literal-only features provides a density-normalized measure of representational asymmetry (Table 5).

Neologisms show a dramatically elevated ratio (>3 in early and middle layers), indicating the model recruits approximately three times as many unique features for the neologism sentence as for its literal counterpart. This reflects the fundamental representational challenge of processing a novel lexical form.

Idioms show consistently low ratios (<1 across all regions), meaning the literal paraphrase actually activates *more* unique features. This is consistent with the stored-unit hypothesis: idioms are retrieved as lexicalized chunks, requiring fewer distinct representational dimensions than their compositional literal counterparts.

Semantic-shift slang and metaphor occupy an intermediate position, with ratios near 1 in early layers and slightly elevated in middle layers, suggesting a modest increase in representational complexity during meaning reassignment.

Category	Frag.	Freq.	LF Score	Span
Idiom	1.20	86.5	+1.28	2.9
Construction†	1.48	76.9	+0.44	3.4
Metaphor	1.36	75.1	+0.45	3.6
Semantic Shift	1.52	70.3	+0.02	5.3
Neologism	2.93	69.9	−1.37	1.1
$r = -0.95$ ($N = 5$ categories); $\beta = -5.88$, $p < .001$				

Table 6: Lexical familiarity metrics. Frag. = subword tokens per word; Freq. = token frequency percentile; LF Score = combined z-score; Span = mean divergence span length in words.

4.5 Lexical Familiarity as Continuous Predictor

To operationalize lexical familiarity independently of category labels, we computed two metrics for each item's divergence span: subword fragmentation (tokens per whitespace word) and token frequency percentile. These were combined into a lexical familiarity score (LF $= -z_{\text{frag}} + z_{\text{freq}}$; higher = more familiar).

Table 6 reports mean LF scores and span lengths by category. Neologisms show substantially higher fragmentation (2.93 vs. 1.2–1.5 for other categories), reflecting tokenizer uncertainty about novel forms. LF score correlates strongly with peak divergence layer at the category level ($r = -0.95$). It's worth noting that the category-level correlation is computed over five data points and should be interpreted as descriptive of the ordering rather than as a robust statistical estimate. The item-level regression provides stronger statistical grounding. To test whether LF score captures cross-category variation beyond discrete labels, we regressed peak divergence layer on continuous LF score across all 3,487 items. Because peak layer is determined at the category level, this regression tests whether the continuous familiarity metric recovers the category ordering rather than predicting item-level variation within categories. LF score is a significant predictor ($\beta = -5.88$, $p < .001$, $R^2 = 0.34$).

Divergence span lengths also varied across categories (neologism: 1.1 words; semantic shift: 5.3 words). The category-level correlation between span length and peak cosine distance does not reach significance ($r = -0.83$, $p = 0.08$, $N = 5$). In regression, lexical familiarity remains a significant predictor after controlling for span length ($\beta = -3.57$, $p < .001$). Span length contributes additional explanatory variance, but the familiarity gradient is not reducible to span-length differences.

4.6 Profile Correlations

Pearson correlations between 48-layer cosine distance profiles reveal two clusters. The known-word categories (construction†, metaphor, semantic shift) all correlate with each other at $r > 0.82$, consistent with their similar early-layer peak profiles. The construction–idiom correlation ($r = 0.862$) suggests that many constructional items may function more like entrenched units than productive templates (see §5.3). Neologisms show the lowest correlations with idioms ($r = 0.479$) and metaphor ($r = 0.679$), reflecting their qualitatively different deep-layer resolution profile.

4.7 Generalization Across Models

The analyses reported thus far use a single instruction-tuned model. To test whether the lexical familiarity gradient reflects a general property of transformer language processing or is specific to Gemma-3-12B-IT, we replicate the residual analysis on two additional models: **Gemma-3-12B-PT** (Gemma Team, 2025), the pretrained variant of the same model, and **Qwen3.5-9B-Base** (Qwen Team, 2026b). Each model is paired with its native SAE release (`gemma-scope-2-12b-pt-res-all` for the Gemma3-PT (McDougall et al., 2025), and `SAE-Res-Qwen3.5-9B-Base` for Qwen (Qwen Team, 2026a)). All other aspects of the pipeline (datasets, span extraction, pooling, residual baseline subtraction) are unchanged **(see full layer-wise profiles in Appendix D)**.

Within-model residual peaks largely replicate (Table 7). In Gemma-PT, four of five categories replicate the IT residual peak layer to within one layer: idiom (L1 → L1), construction† (L7 → L7), metaphor (L29 → L28), and neologism (L41 → L31; still firmly in the deep region). Semantic shift is the one category whose global maximum shifts substantially (L9 → L45), but L7 and L8 appear in its top-5 residual layers, suggesting a bimodal pattern rather than a qualitative change in processing. In Qwen3.5-Base, the depth-percentage ordering at the gradient endpoints is preserved (idiom at L1, 3% depth; neologism at L29, 94% depth), and the early-tier categories (construction, semantic shift) remain shallower than the deep-tier categories. Across all three models, idiom peaks at L1 and neologism peaks in the deep region, recovering the same qualitative endpoint structure as the original Gemma-IT analysis.

Category	Gemma-IT (48L)	Gemma-PT (48L)	Qwen (32L)
Idiom	L1 (2%)	L1 (2%)	L1 (3%)
Constr.†	L7 (15%)	L7 (15%)	L5 (16%)
Sem. Shift	L9 (19%)	L45*	L4 (13%)
Metaphor	L29 (62%)	L28 (60%)	L20 (65%)
Neologism	L41 (87%)	L31 (66%)	L29 (94%)

Table 7: Residual peak layer (category cosine distance minus literal paraphrase baseline, computed within each model) by category and model. Depth percentage in parentheses. *Semantic shift in Gemma-PT has L7 and L8 in its top-5 residual layers (matching IT's L9 peak); the global maximum at L45 likely reflects the deep-layer plateau characteristic of base models. See the section below for full discussion.

Terminal-layer behavior is alignment-specific. While residual peak layers largely replicate, the raw cosine distance profile shape differs systematically. In Gemma-IT, divergence collapses sharply in the last 4–6 layers for every category, consistent with output-stage convergence onto similar response-preparation states. In both Gemma-PT and Qwen, divergence remains elevated through the final layer; no terminal collapse is observed. We interpret this difference as a property of alignment training rather than figurative processing: instruction tuning rewards producing comparable continuations across register variants, which compresses figurative and literal representations near the output, while base models, optimized only for next-token prediction, retain register-distinct features through the final layer. Within-model residual subtraction isolates this confound: the LF-driven peak-layer structure is preserved across all three models once the literal paraphrase baseline is removed within each model.

To sum up, the lexical familiarity gradient is not an artifact of instruction tuning on Gemma-3-12B-IT. The peak structure replicates within Gemma's base variant, and the idiom-shallow / neologism-deep ordering replicates across model families. The terminal-layer compression observed in the original Gemma-IT profiles is itself a distinct, alignment-specific phenomenon that the residual methodology already controls for.

5 Discussion

5.1 Divergence Type Framework

Our results support organizing nonliteral language processing by *divergence type* rather than by figurative category. We identify three types along a gradient.

(1) **Entrenched retrieval** (idioms, L1). The most fundamental representational difference is between a highly entrenched multi-word unit and a compositional expression. The model resolves this distinction immediately, at the very first layer, suggesting that idiomatic patterns are recognized through a retrieval-like mechanism before compositional processing begins. The subsequent declining divergence profile (from 0.830 at L1 to lower values deeper) indicates that later layers do not substantially revise the stored representation.

(2) **Known-word reanalysis** (metaphor L8, semantic shift L9, construction† L7). When the constituent words are individually familiar but their combination or contextual use is nonliteral, the model shows elevated divergence in the early layers. The convergence of three distinct categories within the narrow L7–9 band is consistent with a shared mechanism in which the standard meaning of a known word is initially activated and then reassigned in context. However, residual analysis (§4.3) shows that the middle-tier signal does not clearly exceed the paraphrase noise floor for metaphor and semantic-shift slang, so this interpretation should be treated as provisional.

(3) **Novel-word construction** (neologisms, L41). When the lexical form itself is unfamiliar, the model cannot leverage prior word-level representations and must build meaning from contextual evidence over many layers. The $3\times$ feature elevation throughout the network confirms that neologisms produce fundamentally different representations at every processing stage.

The literal paraphrase control (L3) captures **lexical identity divergence**, where different words map to the same meaning, and establishes the baseline cost of surface-form variation.

This framework explains the otherwise puzzling finding that idioms peak *earlier* than literal paraphrases despite both involving familiar vocabulary: idioms and their counterparts differ in storage mode (entrenched unit vs. compositional expression), the most fundamental representational distinction the model can detect, while paraphrases differ only in lexical identity. The gradient thus operates at multiple levels, with unit-level familiarity resolved before word-level familiarity.

This three-tier structure is further supported by the lexical familiarity analysis (§4.5): LF score, computed from tokenizer behavior alone, predicts peak layer with $r = -0.95$ across categories.

5.2 Cognitive Parallels

Our three-tier gradient aligns with psycholinguistic dual-route models of idiom processing (Swinney and Cutler, 1979; Titone and Connine, 1999): the idiom result (L1 peak, declining profile) maps onto a retrieval route, while the neologism result (L41 peak, elevated features throughout) maps onto a compositional route engaged when no stored representation is available.

The known-word figurative categories occupy an intermediate position consistent with a contextual reanalysis mechanism, providing a computational analogue to graded models of figurative language comprehension, though whether this parallel reflects genuine mechanistic similarity with human processing or a structural analogy remains an open empirical question.

5.3 Slang is Not Special

A key contribution of including metaphor and idiom baselines is that slang does *not* exhibit a unique processing signature distinct from other figurative language. Slang types distribute along the same lexical familiarity gradient as the established figurative baselines. Semantic-shift slang (L9) clusters with metaphor (L8), which makes sense linguistically: both involve reanalysis of familiar words in context. Neologisms occupy a unique position on the gradient, but their uniqueness derives from lexical novelty, not from being slang per se.

This has practical significance: when evaluating NLP system performance on slang, the relevant distinction is not "slang versus standard" but "novel forms versus familiar forms with novel meanings." Systems that handle metaphor well should also handle semantic-shift slang reasonably; neologisms pose a qualitatively different challenge that requires separate evaluation.

The convergence of semantic-shift slang (L9) with metaphor (L8) is the most robust middle-tier finding, supported by large sample sizes (N=1,002 and N=625 respectively) and a high profile correlation ($r = 0.837$). Constructional slang† also falls within this band (L7), and its high correlation with idioms ($r = 0.862$) raises the possibility that many items in this small dataset function more like entrenched units than productive templates. With only N=37, however, we treat this as a preliminary observation. Whether slang can be effectively reduced to two functional categories (known-word slang and novel-word slang) is an empirical ques-

tion that requires a larger constructional dataset to resolve.

5.4 Implications for Reliable and Equitable NLP

Interpretability. The divergence type hierarchy provides a principled framework for understanding model behavior with nonliteral language. Rather than treating figurative language as a monolithic challenge, practitioners can anticipate processing patterns based on lexical properties of input. The three-level hierarchy offers specific predictions about where in a model's layers to look for evidence of figurative meaning processing, guiding targeted probing and intervention studies.

Robustness. The 3× feature elevation and deep-layer processing for neologisms identifies a concrete vulnerability: models require substantially more representational resources for novel vocabulary. This is particularly concerning for deployment in domains characterized by rapid lexical innovation—social media, youth culture, emerging technology discourse, crisis communication where novel terminology often arises spontaneously. Robustness testing should specifically target novel lexical forms rather than treating all non-standard language uniformly.

Fairness. Nonliteral language use, and slang in particular, is not uniformly distributed across demographic groups. Young speakers, speakers of nonstandard dialects, and online communities develop novel vocabulary at higher rates than other populations. Our finding that neologisms impose the greatest representational cost (deep-layer processing, elevated feature recruitment) provides *representational* evidence for a potential source of systematic performance disparity. We emphasize that elevated representational divergence does not automatically entail degraded downstream performance; establishing that link would require task-level evaluation (e.g., paraphrase accuracy or sentiment classification stratified by category), which we leave to future work. Nonetheless, the architectural asymmetry we document (novel vocabulary requiring qualitatively deeper processing than standard vocabulary) identifies a concrete locus where fairness-oriented auditing could be directed.

Methodological reliability. The SAE activation density confound we identify has implications beyond this study. As SAE-based interpretability becomes more prevalent in safety and alignment research, ensuring that reported patterns reflect genuine model processing rather than architectural artifacts is essential for reliability. Our dual-baseline methodology: identical pairs for pipeline validation, literal paraphrases for noise floor estimation offers a template for rigorous SAE-based research.

6 Conclusion

We present a unified layer-wise analysis of five categories of nonliteral language plus two control conditions in Gemma-3-12B-IT using sparse autoencoders. Our findings reveal a *lexical familiarity gradient*: processing depth is determined not by figurative category but by the nature of the divergence between expressions: entrenched retrieval (L1), known-word reanalysis (L7–9), or novel-word construction (L41).

Neologisms stand apart both qualitatively and quantitatively, requiring deep processing, approximately 3× more unique features, and activating largely disjoint feature sets from their literal counterparts ($J = 0.15$). We also identify a methodological confound in SAE feature counts that warrants attention from the interpretability community.

More broadly, our finding that lexical familiarity predicts processing depth suggests that models organize their internal representations around available prior knowledge rather than maintaining distinct pathways for distinct linguistic categories. The peak-layer structure replicates within Gemma's base variant (four of five categories within one layer) and the endpoint ordering replicates in a Qwen base model, indicating the gradient is not an artifact of instruction tuning on a single model (§4.7). This perspective can inform robustness testing, fairness auditing, and interpretability evaluation for NLP systems handling the full diversity of human language.

Limitations

Correlational, not causal. All evidence in this paper is observational: SAE feature analysis reveals where representations diverge but does not establish that these divergences are causally necessary for the model's computation of meaning. The inference from "divergence peaks at layer L" to "layer L participates in figurative processing" is suggestive but not conclusive. The most direct causal extension would be representation steering at the identified peak layers. For each category, a

steering vector can be computed as the mean activation difference between figurative and literal sentences at the divergence span. Injecting this vector at peak layers into literal inputs and measuring whether the model's output shifts toward figurative interpretation would test whether these layers are causally sufficient for figurative meaning construction. Likewise, injecting at non-peak layers should produce weaker or null effects. This layer-category interaction would provide causal evidence for the lexical familiarity gradient beyond the correlational patterns reported in this paper.

Evidence strength varies across the gradient. The lexical familiarity gradient is best supported at its endpoints. Idioms (residual peak: 0.409) and neologisms (residual peak: 0.505) produce divergence that clearly and substantially exceeds the paraphrase noise floor across all layer regions. The middle tier (metaphor and semantic-shift slang) shows near-zero or negative residuals after baseline subtraction (Table 4), meaning their signals do not reliably exceed surface-form variation. The peak-layer ordering (L7–9, earlier than L41) is preserved, but the effect magnitudes are not statistically validated at the item level. We therefore frame our core claims around the three-tier structure (entrenched retrieval, known-word reanalysis, novel-word construction) rather than asserting fine-grained distinctions among individual middle-tier categories.

Lexical familiarity vs. frequency. While §4.5 operationalizes lexical familiarity via subword fragmentation and token frequency, both proxies ultimately derive from tokenizer and corpus statistics. This leaves open whether the gradient reflects familiarity in a cognitive-linguistic sense or simply training-data frequency: tokens that appear more often in the pretraining corpus may produce more stable representations regardless of any notion of "familiarity." Disentangling the two would require a familiarity measure independent of distributional statistics, such as human familiarity ratings.

Model coverage. The main analyses are conducted on Gemma-3-12B-IT; §4.7 extends them to Gemma-3-12B-PT and a Qwen base model, and reports replication of the residual-peak structure for four of five categories in Gemma-PT and preservation of the endpoint ordering in Qwen. The semantic-shift category does not produce a clean global-maximum match in Gemma-PT, indicating that some middle-tier specifics may depend on training regime even when the overall LF gradient generalizes. Broader replication, particularly across model scales and additional model families, would further establish how far the gradient extends.

Construction dataset size. The constructional slang subset (N=37) limits statistical reliability. We report results for transparency and mark them with † throughout, but core claims do not depend on this category.

Literal counterpart asymmetry. Categories differ in how their literal counterparts were constructed: semantic-shift pairs provide contextually different usages of the same token; neologism and construction pairs use LLM-generated paraphrases; metaphor and idiom pairs come from published benchmarks. The literal paraphrase baseline partially controls for this heterogeneity, but some confounding between paraphrase construction method and figurative type may remain.

English only. All datasets are English. Languages differ in how they structure nonliteral meaning, and the familiarity gradient may not transfer directly to languages with different morphological systems or writing conventions.

SAE fidelity. SAE reconstruction quality varies across layers. If Gemma Scope 2 reconstructs certain layers more faithfully than others, this could introduce systematic bias into the layer-wise cosine distance profiles. A per-layer reconstruction error curve would help rule out this confound but is not included in the present analysis.

Broader Impact Statement

This work analyzes internal representations of an existing open-weight model (Gemma-3-12B-IT) and does not involve training new models, collecting human data, or deploying systems. All datasets consist of publicly available resources; no personally identifiable information is involved.

Fairness and bias. Slang is disproportionately associated with marginalized communities, youth, and speakers of non-standard dialects. Language models that systematically misprocess slang risk reinforcing disparities in downstream applications such as content moderation, sentiment analysis, and machine translation, where non-standard language may be flagged as toxic, incoherent, or low-

quality. Our finding that lexical familiarity, which correlates with training-data frequency, predicts processing depth suggests that slang from underrepresented communities may receive shallower or less reliable processing.

Safety and robustness. Our methodological finding that raw SAE feature counts are confounded by layer-wise activation density has implications for the reliability of SAE-based interpretability methods more broadly. Researchers using feature counts to identify safety-relevant mechanisms (e.g., detecting deceptive or harmful content) should be aware that apparent convergence across conditions may be artifactual. We recommend baseline controls as standard practice.

Societal risks. Understanding how models process slang could in principle inform the development of more targeted content filters or surveillance tools for informal language communities. We believe this risk is limited by the generality of our findings: we identify processing principles rather than exploitable features of specific slang terms. The interpretability benefits, which allow the identification and remediation of failure modes, substantially outweigh this risk.

Environmental costs. Our experiments involve forward-pass inference through a 12B-parameter model across approximately 4,500 sentence pairs at 48 layers, with no model training. The computational footprint is modest relative to training-based research.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Technical report, Anthropic.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. Technical report, Anthropic.

Ranuga Disansa Gamage. 2025. genz-slang-pairs-1k. Kaggle Dataset, Revision 9728f69.

Gemma Team. 2025. Gemma 3 technical report. Technical report, Google DeepMind.

Jonathon Green. 2025. Green's dictionary of slang. Accessed: 2025.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657. Association for Computational Linguistics.

Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. LinguaLens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28232–28251, Suzhou, China. Association for Computational Linguistics.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.

Rohit Kulkarni. 2025. Urban dictionary words dataset. HuggingFace Dataset.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13427–13443. Association for Computational Linguistics.

LM-Lexicon. 2025. Lm-lexicon: Slang term list. HuggingFace Dataset.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. 2025. Gemma scope 2: A technical report. Technical report, Google DeepMind.

Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. 2026. Tug-of-war between idioms' figurative and literal interpretations in llms. *Preprint*, arXiv:2506.01723.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 881–889. Association for Computational Linguistics.

Qwen Team. 2026a. Qwen-Scope: Turning sparse features into development tools for large language models.

Qwen Team. 2026b. Qwen3.5: Accelerating productivity with native multimodal agents.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Zhewei Sun and Yang Xu. 2022. Tracing semantic variation in slang. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313. Association for Computational Linguistics.

Zhewei Sun, Richard Zemel, and Yang Xu. 2021. A computational framework for slang generation. *Transactions of the Association for Computational Linguistics*, 9:462–478.

Zhewei Sun, Richard Zemel, and Yang Xu. 2022. Semantically informed slang interpretation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5213–5231. Association for Computational Linguistics.

David A. Swinney and Anne Cutler. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, 18(5):523–534.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 5701–5715. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12):1655–1674. Literal and Figurative Language.

Urban Dictionary. 2025. Urban dictionary. Accessed: 2025.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1308. Association for Computational Linguistics.

Jianing Zhou, Hongyu Zheng, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions*, pages 33–48. Association for Computational Linguistics.

A Prompt for Literal Counterpart Generation

For the neologism and constructional slang categories, literal counterparts were generated by prompting an LLM (Claude, Opus 4.5) with the following instruction:

```
Task: Generate a usage sentence
and its literal counterpart for
a neologism.

Input:

  • Neologism: {term}
  • Gloss: {gloss}

Instructions: (1) If the gloss
contains an example sentence,
use it. Otherwise, write a
natural sentence using the
neologism. (2) Write a literal
counterpart that replaces
the neologism with standard,
dictionary-attested English
while preserving syntax, tense,
and register. The two sentences
should form a minimal pair.

Constraints:

  • Vary sentence structure
    across the dataset: use
    statements, questions,
    exclamations, dialogue, etc.
```

- Avoid repetitive templates (e.g., "That's a X" or "That's so X").
- The neologism should appear in naturalistic contexts (conversations, narratives, social media posts).

```
Output format:
Slang:  [sentence with
neologism]
Literal:  [sentence with
standard English equivalent]
```

A parallel prompt was used for constructional slang, replacing "neologism" with "constructional slang expression" in the task description, and substituting the single-word `{term}` field with the full multi-word pattern (e.g., *It's giving X*). All other instructions and constraints were identical.

Each generated paraphrase was reviewed to ensure semantic fidelity and syntactic parallelism with the original. Cases where the LLM introduced significant structural changes were regenerated or manually corrected.

B Reproducibility

The datasets and code required to reproduce our experiments are available in our GitHub repository.

The repository includes category-separated files under `data/` for idioms, metaphors, semantic-shift slang, neologisms, constructional slang, and PAWS-based controls. Each example is a minimal pair containing columns as listed below:

- Sentence A (original)
- Sentence B (paraphrase / counterpart; meaning preserved)
- The differing token spans in sentence A (concatenated if multiple spans)
- The differing token spans in sentence B (concatenated if multiple spans)
- Token index ranges for the differing spans in sentence A
- Token index ranges for the differing spans in sentence B
- Number of differing spans

The repository also includes scripts for running the full layer-wise analysis across all three evaluated models (Gemma-IT, Gemma-PT, and Qwen) (see the README for execution instructions) and their corresponding pre-computed results under `results/`, enabling verification of all reported figures without re-running inference.

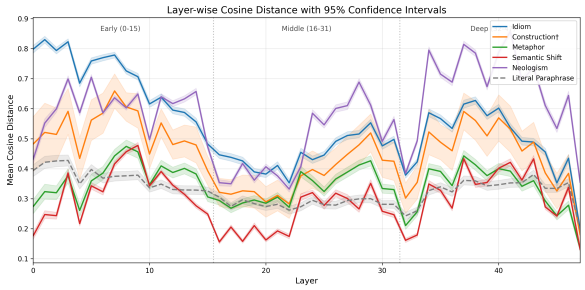

Figure 1: Mean cosine distance between paired figurative and literal SAE feature vectors across 48 layers. Shaded regions indicate ±1 standard error.

C Layer-Wise Profile Visualizations

Figure 1 shows the full 48-layer cosine distance profiles for all five nonliteral categories plus the literal paraphrase control. The idiom profile (peak at L1, declining) and the neologism profile (rising through middle layers, peak at L41) are the most visually distinctive, consistent with their strong residual signals. Metaphor and semantic-shift slang profiles track the literal paraphrase baseline more closely, consistent with the weak residual signal reported in Table 4.

Figure 2 shows the figurative-to-literal feature ratio across all 48 layers. The neologism curve is elevated throughout the network (consistently above 2.0), while idiom ratios remain below 1.0, confirming the representational asymmetry reported in Table 5.

Figure 3 visualizes the relationship between lexical familiarity and processing depth at the category level. Idioms, with the highest mean LF score (+1.28), peak at the shallowest layer (L1), while neologisms, with the lowest LF score (-1.37), peak deepest (L41). The middle-tier categories (construction, metaphor, semantic shift) cluster together with similar LF scores (0.0–0.5) and similar peak layers (L7–9). The near-perfect negative correlation ($r = -0.95$) confirms that lexical familiarity, operationalized independently of category labels, predicts processing depth.

D Generalization: Layer-Wise Profiles for Gemma-PT and Qwen

This appendix provides the layer-wise visualizations supporting §4.7 for the two generalization models, in three views matching Appendix C: co-

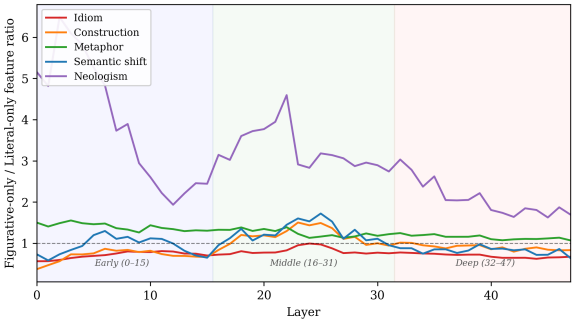

Figure 2: Figurative-only / literal-only feature ratio per layer. Values above the dashed line at 1.0 indicate more unique features in the figurative sentence.

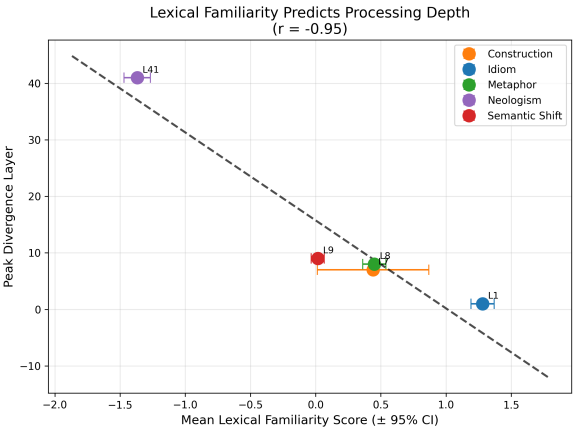

Figure 3: Mean lexical familiarity score vs. peak divergence layer by category. Error bars indicate 95% confidence intervals. The dashed line shows the regression fit ($r = -0.95$).

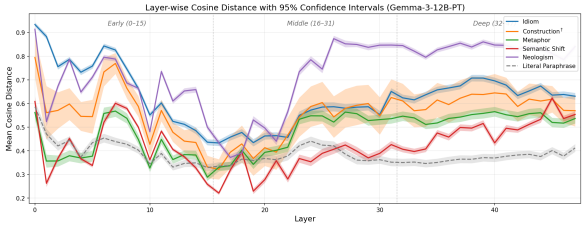

Figure 4: Gemma-3-12B-PT: layer-wise mean cosine distance with 95% confidence intervals.

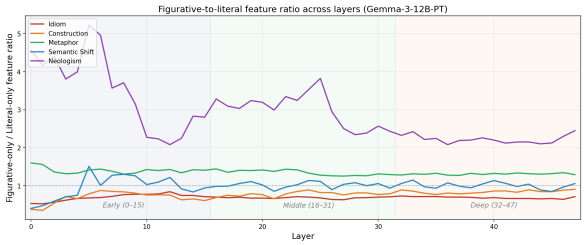

Figure 5: Gemma-3-12B-PT: figurative-only / literal-only feature ratio per layer.

sine distance across layers, figurative-to-literal feature ratio, and lexical familiarity vs. peak layer. For PT and Qwen, the peak layer in the LF plot uses the within-model residual peak (category cosine distance minus literal paraphrase baseline); residual peaks are reported in Table 7.

D.1 Gemma-3-12B-PT

The Gemma-PT profile (Figure 4) preserves the prominent early-layer peaks observed in Gemma-IT (idiom L0/L1, construction L7) and shows a marked deep-layer plateau rather than the late-layer collapse characteristic of the instruction-tuned model. The feature ratio (Figure 5) retains the same qualitative ordering: neologisms are elevated throughout the network, while idiom and construction stay below 1.0. The LF vs. residual peak layer plot (Figure 6) yields $r = -0.61$ across categories; the lower correlation relative to Gemma-IT's -0.95 is driven by semantic shift, whose global residual maximum shifts to L45 (though L7 and L8 appear in its top-5 residual layers, matching IT's L9 peak).

D.2 Qwen Base Model

The Qwen profile (Figure 7) preserves the gradient endpoints: idiom peaks at L1 (3% depth) and neologism at L29 (94% depth). Like Gemma-PT, Qwen does not show terminal-layer collapse and divergence rises in deep layers across all categories. The feature ratio (Figure 8) is compressed relative to

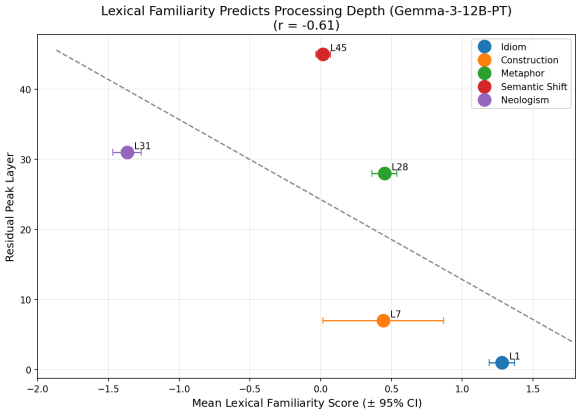

Figure 6: Gemma-3-12B-PT: mean lexical familiarity score vs. residual peak divergence layer (within-model residual against the literal paraphrase baseline). Dashed line shows the regression fit ($r = -0.61$).

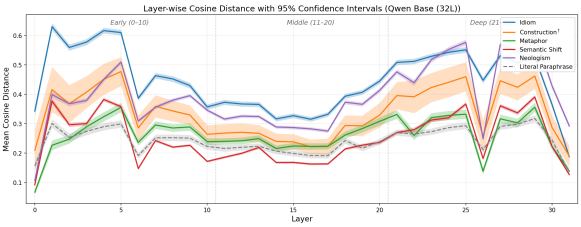

Figure 7: Qwen 3.5 9B (32 layers): layer-wise mean cosine distance with 95% confidence intervals.

Gemma's: neologism is the only category elevated above 1.0 in early layers, and the remaining categories cluster near unity, reflecting differences in SAE feature density rather than figurative processing. The LF vs. residual peak layer plot (Figure 9) yields $r = -0.84$.

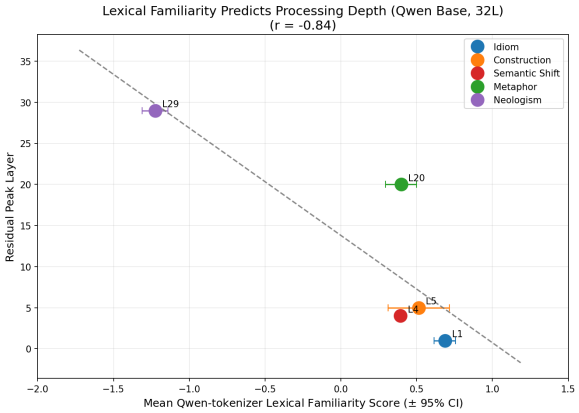

Figure 9: Qwen 3.5 9B: mean lexical familiarity score vs. residual peak divergence layer. Dashed line shows the regression fit ($r = -0.84$).

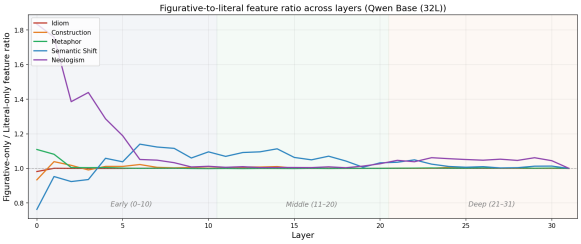

Figure 8: Qwen 3.5 9B: figurative-only / literal-only feature ratio per layer.