Polysemanticity or Polysemy? Lexical Identity Confounds Superposition Metrics

Anonymous ACL submission

Abstract

A neuron firing for two unrelated meanings of the same word is typically read as forced feature compression: the model has packed multiple concepts into one neuron, a phenomenon called *superposition*. The standard measurement conflates this with a simpler signal. The two inputs share a hidden variable the metric does not control for: the word form itself. The neuron may be tracking that one shared variable rather than compressing multiple concepts. A direct synonym-substitution test confirms it: replacing a polysemous word with a sense-preserving synonym (*deposit money in the bank* $\rightarrow$ *deposit money in the financial institution*) collapses the activations of flagged neurons to baseline ($\rho \approx 0$), while sense-selective neurons retain $\rho \approx 0.8$. A monosemous control on unambiguous words (*democracy*, *hospital*) shows the form-encoding regime is polysemy-specific ($\rho \approx 0.7$–0.85). The dissociation replicates across ten transformers (110M–70B), is strongest on Llama-3-8B, and is corroborated by a population-level 2×2 factorial decomposition (form $\times$ meaning) across four corpora. The signal is compact ($\leq 1\%$ of activation dimensions), removable (a 20-dimensional subspace projection cuts the gap by 37–51%), and propagates into sparse autoencoders. We provide three deployable tools (a diagnostic ratio R_{lex}, a neuron-level classifier SSI, and a subspace-removal correction LIS) that make knowledge edits more selective ($p = 0.002$) and improve word sense disambiguation. Genuine compression remains; what current metrics call polysemanticity is substantially lexical identity.

1 Introduction

Suppose a single neuron in a language model fires strongly on "the lender approved the loan" and again on "we hiked along the riverside." By definition, this is a *polysemantic* neuron: the model has more concepts to represent than it has dimensions, so it is forced to pack multiple unrelated meanings into a single neuron. Mechanistic interpretability calls this phenomenon *superposition* (Elhage et al., 2022; Olah et al., 2020). Sparse autoencoders (SAEs), the leading tool for inspecting such neurons, are designed to recover one feature per concept, so that each feature in the dictionary fires for a single thing (Bricken et al., 2023; Templeton et al., 2024).

This paper points out that the underlying measurement has a confound. Both example sentences also share a token: the word *bank*. If the neuron is responding to that word form, rather than to either of its meanings (financial or geographic), it will fire for both inputs without representing either concept. From the outside this looks identical to compression. Two inputs with two unrelated meanings hit the same neuron. But nothing is being compressed: the neuron is doing exactly one thing, namely, firing in the presence of the word *bank*, before the model has separated the senses. Standard polysemanticity metrics only check whether the activating inputs differ in meaning, so they cannot tell which of the two stories is true (Olah et al., 2020; Bricken et al., 2023).

The distinction matters because every downstream tool that builds on this measurement inherits the confound. Run an SAE on a layer with a *bank*-detecting neuron, and the dictionary will produce a feature that fires for *bank* in all of its senses, looking monosemantic in isolation while in fact tracking a word form. Use such a feature to edit or steer model behavior (Meng et al., 2022; Arditi et al., 2024), and the change spreads across both senses. The same applies to any procedure that calls a representation "shared between concepts" on the basis of input-set overlap alone.

We isolate the confound by varying the two factors separately. For each polysemous word w, we build sentence pairs that cross *same/different word form* with *same/different sense*, giving four condi-

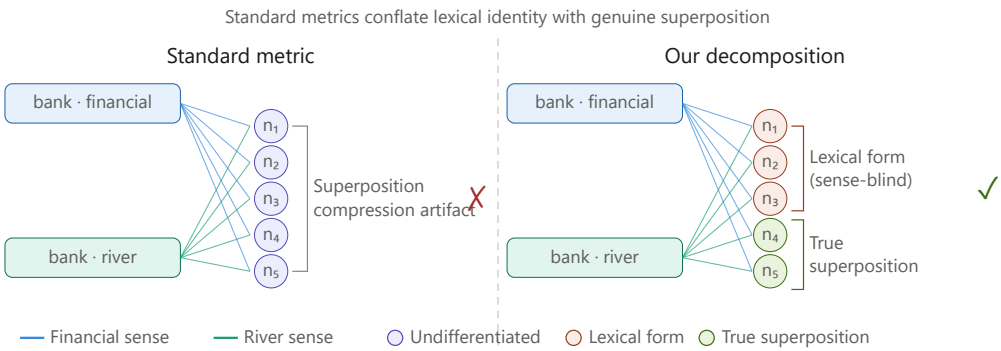

Figure 1: **Overview.** Standard polysemanticity metrics (top) see that neurons n_1–n_5 fire for both senses of *bank* and call all of it forced compression (superposition). Our decomposition (bottom) separates two populations: a large *sense-blind* group tracking one variable (word form; 85%/42% form/sense probe accuracy) and a small *sense-selective* group encoding both form and sense (85%/91%), the candidate for true superposition (multiple concepts in one neuron).

tions (Figure 2). The off-diagonal cells are diagnostic. PS (same word, different sense) measures overlap from shared word form alone; SYN (different word, same sense) measures overlap from shared meaning alone. Whichever produces more neuron overlap is the dominant driver of what the standard metric is picking up. A token-embedding baseline ($R_{\text{lex}}^{\text{emb}}$) controls for the fact that two sentences containing the same word already share their input embeddings; only MLP-level overlap that exceeds this baseline counts as evidence of a confound beyond trivial input reuse (Table 7).

This is a claim about measurement, not about a particular model. Across ten transformers (110M–70B parameters, four architecture families including GPT-2, Pythia, LLaMA-2, BERT, and ELECTRA), PS exceeds SYN at every layer in every model (Wilcoxon $p < .001$, Holm–Bonferroni corrected); the ordering persists above the embedding baseline. The same pattern carries into pretrained SAEs: 18–36% of features active on polysemous words are sense-blind. A probing dissociation makes the distinction concrete: a *sense-selective* subpopulation (<1% of neurons) reaches 91% sense-classification accuracy; *sense-blind* neurons reach 42% (near chance), though both groups encode word identity at ∼85%. Ablating the selective group produces sense-specific perplexity changes 6.6× more targeted than matched sense-blind groups, and the lexical signal occupies a 20-D subspace whose removal eliminates the PS–SYN gap.

A direct synonym-substitution test (Section 5.6) pins this down at the neuron level: when the polysemous word is replaced by a sense-preserving synonym, sense-blind neurons collapse to baseline ($\rho \approx 0$, sometimes negative) while sense-selective neurons retain ≈ 60–110% of their activation. A monosemous control on unambiguous content words (*democracy*, *hospital*, *electricity*) rules out the obvious counter-explanation: top-K neurons on those words retain $\rho \approx 0.7$–0.85 under the same substitution, so the form-encoding regime is specific to polysemy rather than a generic property of the model. The dissociation replicates across ten transformers (BERT-base through Llama-2-13B and Llama-3-8B) and is strongest on the most recent model (Llama-3), addressing the concern that it might be a small- or old-model artifact.

We do not claim superposition is illusory. Genuine compression surely exists. The contribution is more specific: a measurable component of what current polysemanticity metrics flag is lexical identity, this component propagates into the SAE-feature and model-editing pipelines that depend on these metrics, and a sparse sense-selective subpopulation carries the actual sense information once the lexical component is removed. We provide three tools that practitioners can apply before declaring a feature monosemantic: a diagnostic (R_{lex}), a per-neuron classification (Sense Selectivity Index, SSI), and a subspace-removal correction (Lexical Identity Subspace, LIS).

2 Related work

Polysemanticity and superposition. The superposition hypothesis (Elhage et al., 2022) frames polysemanticity as a compression artifact, and SAEs

are designed to recover monosemantic features from it (Bricken et al., 2023; Templeton et al., 2024; Sharkey et al., 2022). Recent work has shown that polysemanticity can arise even without capacity pressure (Lecomte et al., 2025) and has tracked its emergence during training (Wu et al., 2025). None of these works examine whether the *inputs* triggering multi-concept activation share a lexical form, which is the confound we identify.

Word sense representations. Ethayarajh (2019) and Wiedemann et al. (2019) established that contextual embeddings cluster by word identity before sense. This is a geometric observation about full hidden states; it was not connected to polysemanticity metrics or to individual neurons. We operate at the level of *individual MLP neurons* (the unit at which polysemanticity is defined), provide a *quantitative metric* (R_{lex}), and trace the confound through *causal and downstream consequences*. Lyu et al. (2025) and Minegishi et al. (2025) study polysemous words in neural representations but do not measure how the confound affects interpretability tools.

Causal methods and model editing. Probing classifiers (Conneau et al., 2018) test what information a representation encodes; activation patching (Vig et al., 2020; Meng et al., 2022) and mean-ablation (Chan et al., 2022) test whether that information plays a causal role. We use both to validate the sense-selective / sense-blind distinction. Model editing methods such as ROME (Meng et al., 2022) and embedding-level interventions (Ait Hou and Hwa, 2026) modify specific representations to change model behavior; none account for lexical identity as a source of collateral damage across word senses.

3 Methodology

Our approach has three parts: (1) a factorial decomposition that separates lexical from semantic contributions to neuron overlap, (2) a neuron classification that identifies which neurons encode word form vs. word meaning, and (3) causal and downstream experiments that test whether the distinction matters. All analyses target MLP intermediate activations, the level at which polysemanticity is typically analyzed (Olah et al., 2020; Elhage et al., 2022; Bricken et al., 2023).

3.1 Decomposition framework

We construct sentence pairs crossing two binary factors (same/different word $\times$ same/different sense), yielding the four conditions in Figure 2. The two off-diagonal cells are diagnostic: PS isolates the contribution of shared word form, SYN isolates the contribution of shared meaning.

For each sentence, we extract MLP intermediate activations at the last subword of the target word (Ethayarajh, 2019) and compute pairwise overlap via three metrics (cosine similarity, Jaccard overlap of active-neuron sets, magnitude divergence), aggregated per word with bootstrap 95% CIs.

Lexical contribution ratio. Let $M(\mathbf{a}, \mathbf{b})$ denote a pairwise overlap measure between MLP activations $\mathbf{a}$ and $\mathbf{b}$ at the target-word position (cosine similarity, Jaccard overlap of above-median active-neuron sets, or magnitude divergence; defined formally in Appendix F). For condition $c \in \{\text{SL}, \text{PS}, \text{SYN}, \text{CL}\}$, $\bar{M}_c$ is the mean of $M(\mathbf{a}, \mathbf{b})$ over all sentence pairs $(\mathbf{a}, \mathbf{b})$ assigned to c, averaged first within each polysemous word and then across words. If lexical and semantic contributions do not interact (validated in Section 5.7), their relative sizes can be read off directly. Normalizing by the total range gives the *lexical contribution ratio*:

$$R_{\text{lex}} = \frac{\bar{M}_{\text{PS}} - \bar{M}_{\text{SYN}}}{\bar{M}_{\text{SL}} - \bar{M}_{\text{CL}}} \tag{1}$$

The numerator is the excess overlap due to shared word form ($\bar{M}_{\text{PS}}$, same word/different sense) over the excess due to shared meaning ($\bar{M}_{\text{SYN}}$, different word/same sense). The denominator normalises by the full range from maximum overlap (SL) to minimum (CL). $R_{\text{lex}} > 0$ means word form contributes more overlap than meaning; $R_{\text{lex}} = 1$ means word form accounts for all of it. As a control, we also compute $R_{\text{lex}}^{\text{emb}}$ from raw token embeddings (before any transformer computation); MLP-level R_{lex} exceeding $R_{\text{lex}}^{\text{emb}}$ indicates the confound is not simply inherited from the input encoding.

Numeric ranges (e.g., $R_{\text{lex}} = 0.74$–0.29) denote layer 0 followed by the final layer; per-layer tables in Appendix B.2.

3.2 Neuron classification

The decomposition above measures overlap at the population level. To test whether individual neurons behave differently depending on whether they

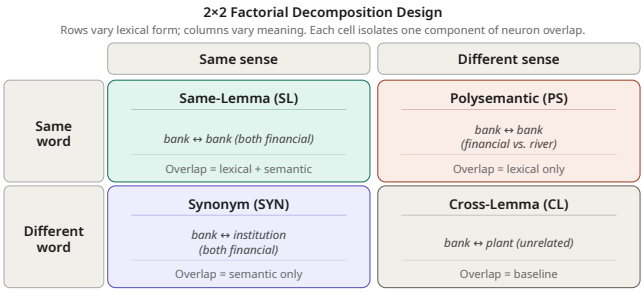

Figure 2: The 2×2 factorial decomposition. Rows = same vs. different word form; columns = same vs. different sense. **SL** (same lemma): same word, same sense; both word form and meaning shared. **PS** (polysemy): same word, different sense; only word form shared, the diagnostic condition for lexical-identity encoding. **SYN** (synonym): different word, same sense; only meaning shared. **CL** (cross-lemma): different word, different sense; neither shared, the floor for activation overlap. The acronym CL stands for "cross-lemma"; the cell does *not* aim to isolate compression-induced collisions, only to bound the lower end of overlap when neither factor is shared.

encode word form or word meaning, we classify them into two groups.

Sense Selectivity Index (SSI). Each polysemous word w has two WordNet sense labels separated by the Wu-Palmer threshold of Section 4; we call these *sense-A* and *sense-B* for w (the assignment is arbitrary and word-specific, e.g., *bank*: A = financial institution, B = geographic land; *light*: A = illumination, B = light-in-weight). For neuron j at the target-word position, let μ_A^j and μ_B^j be the mean activations on sense-A sentences and sense-B sentences respectively, with sample standard deviations s_A^j, s_B^j. We compute Cohen's d:

$$\text{SSI}_j = \frac{|\mu_A^j - \mu_B^j|}{\sigma_{\text{pooled}}^j} \tag{2}$$

with

$$\sigma_{\text{pooled}}^j = \sqrt{\frac{(n_A-1)(s_A^j)^2 + (n_B-1)(s_B^j)^2}{n_A + n_B - 2}},$$

where n_A, n_B are the sentence counts per sense. SSI is invariant to the A/B labeling and held fixed per word across all experiments. Neurons with SSI > 2 are *sense-selective*; those with SSI < 0.5 (above median activation) are *sense-blind*, following Cohen's d conventions for large vs. small effect. The condition ordering does not depend on the threshold choice (Appendix B.4).

Form detectors (sense-label-free). To avoid circularity, we independently identify form detectors using only word identity: for each word w and neuron j, we compute *consistency* ($1 -$ CV of activations across all w-sentences) and *specificity* (Cohen's d vs. other words). Neurons ranking in the top K by consistency $\times$ specificity are form detectors for w: they fire reliably for w regardless of sense.

3.3 Causal intervention protocol

If sense-selective and sense-blind neurons truly encode different information, removing one group should affect model behavior differently from removing the other. We test this via mean-ablation (Chan et al., 2022): replacing selected neurons' activations with their dataset mean, simultaneously across all layers at the target-word position. Neuron counts are matched across groups (sense-A-selective, sense-B-selective, sense-blind, random) at each layer. We evaluate via: **(1)** KL divergence of the full output distribution, **(2)** sense accuracy (diagnostic token probabilities), and **(3)** sense-specific perplexity change, with specificity $= |\Delta\text{ppl}_{\text{target}} - \Delta\text{ppl}_{\text{other}}|$.

4 Experimental setup

Dataset. We use SemCor (Miller et al., 1993), the only large-scale corpus with human-annotated WordNet sense tags (37,176 sentences). From it we select 407 content words (nouns/verbs, ≥ 2 senses, ≥ 5 sentences per sense, Wu-Palmer similarity < 0.50 between senses; varying this threshold from 0.35 to 0.65 does not change the condition ordering). The Wu-Palmer cutoff selects predominantly for *homonymy* (etymologically unrelated senses sharing a token, e.g., *bank/finance* vs. *bank/river*) rather than fine-grained polysemy (*mouth/face* vs. *mouth/river*); the lexical-identity mechanism we study is sharpest in the homonymy case (we revisit this in Limitations). SYN pairs use WordNet

Model	Params	Layers	d_{mlp}	Type
GPT-2 (2019)	117M	12	3072	Auto
GPT-2-Med (2019)	345M	24	4096	Auto
BERT-base (2019)	110M	12	3072	Bidir
ELECTRA-base (2020)	110M	12	3072	Bidir
Pythia-1B (2023)	1.0B	16	8192	Auto
Pythia-6.9B (2023)	6.9B	32	16384	Auto
Pythia-12B (2023)	12B	36	20480	Auto
LLaMA-2-13B (2023)	13B	40	13824	Auto
LLaMA-2-70B (2023)	70B	80	28672	Auto
LLaMA-3-8B (2024)	8B	32	14336	Auto

Table 1: Models evaluated. Auto = autoregressive LM; Bidir = bidirectional (masked LM / replaced-token detection). LLaMA-2-70B uses NF4 quantization.

synonyms attested in SemCor; CL pairs sample unrelated words. Results replicate across ten models trained on four corpora and validate on modern Wikipedia text (Appendix F.1, B.6, A).

Models. The claim concerns measurement, so the relevant model set is the one on which polysemanticity metrics are actually applied in the literature. We use the standard set: GPT-2 and GPT-2 Medium, Pythia (1B, 6.9B, 12B), LLaMA-2 (13B, 70B), LLaMA-3-8B, plus BERT-base and ELECTRA-base. The two bidirectional encoders rule out a left-to-right causal-masking artifact; the set spans four architecture families and 110M–70B parameters (Table 1). Published baselines, open sparse autoencoders (Bloom, 2024; Templeton et al., 2024), and the polysemanticity metrics we critique (Elhage et al., 2022; Bricken et al., 2023; Lecomte et al., 2025; Minegishi et al., 2025) are all defined on these architectures. Models up to 13B are loaded in `float32`; LLaMA-2-70B uses NF4 quantization.

5 Results

We present the evidence in stages: the confound appears consistently across layers (Section 5.1), replicates across model scale (Section 5.2) and into SAEs (Section 5.3), has identifiable mechanistic structure (Sections 5.4–5.9), and affects downstream tasks (Section 5.10). Primary results use GPT-2; replication across models follows in Section 5.2.

5.1 The lexical confound across layers

Cosine similarity follows SL > PS > SYN > CL across layers and models (Figure 3); the same ordering holds for Jaccard overlap and magnitude divergence (Appendix C). PS exceeds SYN at every layer (Wilcoxon $p < 0.001$, Holm-Bonferroni corrected): sharing a word form produces more overlap than sharing a meaning. A token-embedding baseline ($R_{\text{lex}}^{\text{emb}} = 0.71$–$0.88$) confirms the confound is not merely inherited from token embeddings (Table 7).

R_{lex} decreases with depth (0.74 to 0.29 across GPT-2 layers; Figure 11). In all models tested, the bootstrap CI remains above zero at every layer. This is an empirical observation, not a guaranteed property: a model with stronger disambiguation could in principle drive R_{lex} to zero or below. Sense-selective neurons (SSI > 2) stay below 1% of the population, confirming sense information is encoded by a very sparse subpopulation.

5.2 Cross-model consistency

The PS > SYN gap replicates across all ten models tested, from 110M to 70B parameters (Figure 3; Appendix B). Bidirectional models (BERT, ELECTRA) show steeper R_{lex} declines with depth; autoregressive models maintain the pattern through 12–13B. LLaMA-3-8B (layer-averaged $R_{\text{lex}} = 0.31$; PS>SYN significant in $28/32$ layers, $p_{\text{corr}} < 0.001$ in most) replicates the family pattern on the most recent open-weights LM. LLaMA-2-70B (NF4, 80 layers) confirms the ordering at scale (PS $= .117$ vs. SYN $= .076$; $R_{\text{lex}} = 0.45$); per-neuron experiments run up to 13B in full precision.

5.3 Impact on SAEs and standard metrics

Pre-trained GPT-2 SAEs (Bloom, 2024) (32k and 128k features) inherit the confound: 18–32% of active features per layer are sense-blind (Cohen's $d < 0.5$ between two senses of the same word in SemCor), and the 128k SAE ($4\times$ capacity) produces nearly identical ratios (Figure 4). On Pythia-410M (65k SAEs), the range is 19–36%. Increasing SAE capacity by $4\times$ does not change the collision rate, suggesting these features are not artifacts of under-trained dictionaries.

An illustrative case. SAE feature #5132 in GPT-2 layer 8 (32k OAI v5)[1] fires sense-blind for **232 of our 407 polysemous words** across unrelated domains, while its auto-interp label is "intensifying adjectives describing sound or scale." This is an existence proof of mislabeling. The aggregate 18–36% rate and the per-neuron dissociation (§5.4) stand independently.

[1] https://www.neuronpedia.org/gpt2-small/8-res_post_32k-oai/5132

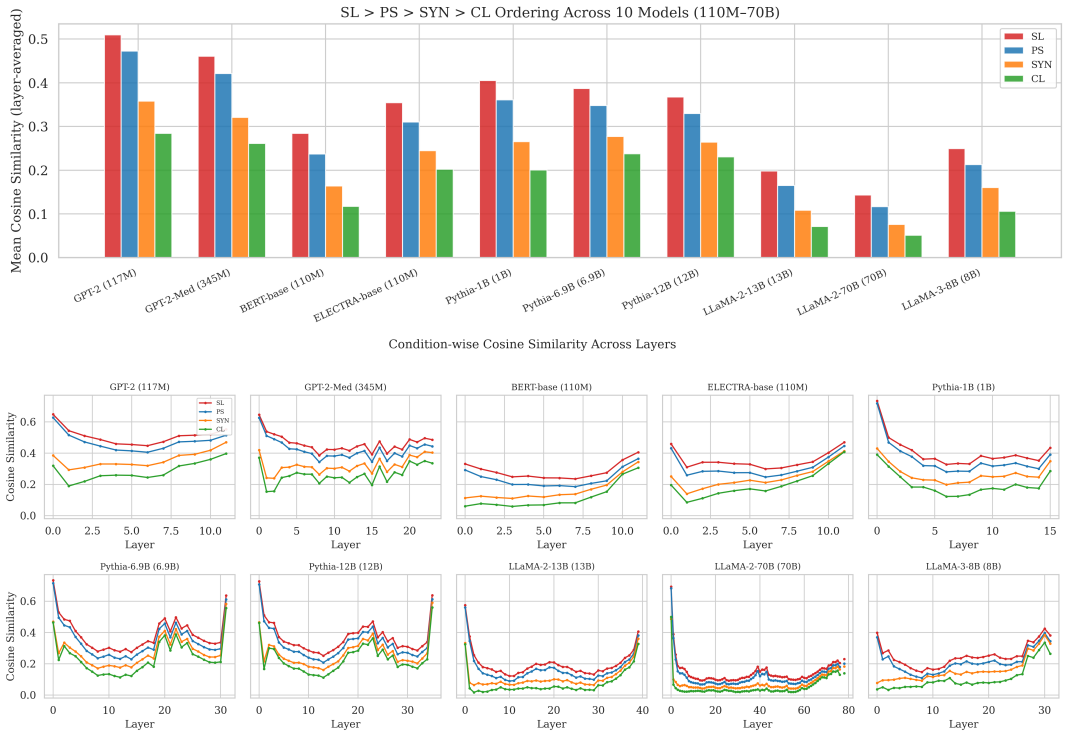

Figure 3: **Top**: Layer-averaged cosine similarity by condition across ten models (110M–70B); conditions defined in Section 3.1. PS consistently exceeds SYN: word form drives more neuron overlap than shared meaning. An embedding baseline $R_{\text{lex}}^{\text{emb}}$ (Table 7) controls for residual-stream propagation; the per-neuron dissociation in Section 5.4 carries the load-bearing evidence. **Bottom**: Per-layer breakdown; R_{lex} trends in Appendix B.

Applying standard polysemanticity scores to our dataset, $\approx 57\%$ of activating input pairs for flagged neurons (conditional on polysemous inputs) share a word form rather than a genuinely different concept. This 57% figure is conditional on polysemous inputs and is *not* a claim about the share of all superposition. Polysemous words are common (>80% of frequent English vocabulary; Rodd et al. 2002), so the practical impact is substantial even if the global share of superposition we explain is necessarily smaller.

5.4 Neuron-level validation

Two non-circular tests confirm the confound at the neuron level.

Form detectors are sense-blind. Identifying form detectors using only word identity (Section 3.2) and classifying them by SSI: on GPT-2, 67.3% are sense-blind (SSI < 0.5); only 0.1% are sense-selective.

Probing dissociates form from sense. Logistic-regression probes (LOO CV, $C = 0.01$) yield 91.1% sense accuracy on selective neurons, 41.9% (near chance) on blind, 68.9% random; word-form accuracy is ∼85% for all groups. Blind neurons encode form but not sense; selective neurons encode both.

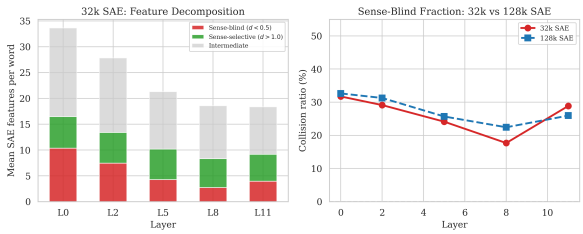

Figure 4: SAE collision analysis (GPT-2). A feature is *sense-blind* if it activates on both senses of the same word with Cohen's $d < 0.5$ (small effect). **Left**: mean number of features per word broken down by sense discriminability (sense-blind vs. sense-selective) in the 32k and 128k OAI v5 SAEs. **Right**: per-layer collision ratio = sense-blind / active. 18–32% of features active on polysemous words conflate senses of the same word (18–36% with Pythia-410M included); doubling SAE capacity from 32k to 128k does not reduce the ratio. See Section 5.3.

5.5 Causal intervention: sense-specific editing

Mean-ablating sense-selective neurons produces 3–5× larger KL divergence than matched blind or random groups (up to 345M; 1.2–2.1× at 12–13B): they carry disproportionate causal weight de-

spite small absolute effects (10^{-3}–10^{-4} KL, as expected when ablating ∼50 of 3k–20k neurons). On GPT-2 (53 words), sense-selective ablation is $6.6\times$ more specific than blind (sense-A: $+0.55/-0.10$; sense-B: $+0.10/+2.62$; blind: $-0.07/+0.17$; full table in Appendix B.1). Per-word at layer 6, the specificity gap reaches $5.5\times$ for *match* and $12\times$ for *light*: sense-blind neurons cannot target one meaning without equally affecting the other.

5.6 Anchor persistence: a direct token-vs-concept test

Probing and ablation are both indirect: one correlates activations with labels, the other perturbs the *neuron*. We now perturb the *input*. Replace each polysemous word w with a sense-preserving synonym (*deposit money in the bank* → *deposit money in the financial institution*); the form is gone, the meaning preserved. For the top-K=50 most-active sense-blind neurons (SSI<0.5) compute $\rho = (\bar{a}_{\text{SYN}} - \bar{a}_{\text{CL}})/(\bar{a}_{\text{PS}} - \bar{a}_{\text{CL}})$, with $\bar{a}_c$ the mean activation of the selected neurons in condition c. Concept encoding predicts $\rho \to 1$; literal-token encoding predicts $\rho \to 0$. We restrict to 13 polysemous words whose WordNet synonyms cleanly carry sense-specific information, audited via activation cosine (Appendix D); two parallel controls (sense-selective top-K, random-active top-K) use the same rule on different pools.

Sense-blind firing tracks word form. Across ten transformers (GPT-2, BERT, ELECTRA, Pythia, Llama-2, Llama-3; 110M–13B), sense-blind top-K collapses to $\rho_{\text{median}} \in [-0.53, 0.18]$ (Table 2); synonym substitution kills the firing despite preserving meaning. Sense-selective top-K retains $\rho \in [0.49, 1.11]$. The effect is strongest in Llama-3-8B (ρ_{blind}=-0.53), so it is not a small- or old-model artifact.

Monosemous control: form-encoding is polysemy-specific. On 15 unambiguous content words (*democracy*, *hospital*, *telephone*, ...) with sense-preserving synonyms (*self-government*, *medical center*, *phone*), top-K neurons retain $\rho \approx 0.69$–0.79 (Table 2): the form-encoding regime is polysemy-specific.

5.7 Controls

The interaction term I is small in all models ($|I| < 0.04$), supporting additivity. R_{lex} correlates weakly with frequency ($r = 0.15$) and non-significantly with position. A synonym-free variant and subset invariance tests confirm robustness (Appendix B, B.6; embedding baseline and interaction plots in Appendix C).

Model	Polysemous (top-50) blind	random	sel.	Monosemous
GPT-2 (117M)	0.01	0.14	0.89	0.79
GPT-2-Med (345M)	0.18	0.09	0.57	0.76
BERT-base	−0.17	0.10	0.49	0.78
ELECTRA-base	−0.46	0.17	0.65	0.84
Pythia-410M	0.06	0.09	0.58	0.69
Pythia-1B	−0.02	0.06	0.50	0.77
Pythia-6.9B	−0.06	0.09	0.53	0.71
Llama-2-7B	−0.21	0.16	0.75	0.76
Llama-2-13B	−0.17	0.05	0.80	0.70
Llama-3-8B	−0.53	0.06	1.11	0.69

Table 2: Anchor-persistence ρ (median across three layers per model at ≈ 25/50/75% depth). Across ten transformers and four architectures, sense-blind top-K neurons collapse to ρ near or below zero under sense-preserving synonym substitution while sense-selective and monosemous top-K neurons retain $\rho \geq 0.49$.

5.8 The confound is correctable

We define a *lexically-adjusted polysemanticity score*:

$$P_j^{\text{adj}} = P_j^{\text{raw}} - \hat{\lambda}_\ell \cdot F_j \tag{3}$$

where P_j^{raw} is the standard score, F_j indicates whether neuron j is a form detector, and $\hat{\lambda}_\ell = R_{\text{lex}}(\ell) \cdot \bar{P}_\ell$ is the layer-specific expected lexical inflation. On GPT-2, this reclassifies 6.5% (layer 6) to 10.4% (layer-averaged) of neurons flagged as polysemantic, with 93–94% of reclassified neurons confirmed as form detectors (SSI < 0.5). The correction targets the top-50 form detectors per word (≈1.6% of GPT-2's MLP neurons), a conservative lower bound.

5.9 Mechanistic analysis: lexical identity subspace

Does the confound have geometric structure? For each word w_i with synonym data ($n = 62$), $\mathbf{d}_i = \bar{\mathbf{a}}_{w_i} - \bar{\mathbf{a}}_{\text{syn}_i}$ isolates form from meaning. PCA on these vectors yields the **lexical identity subspace** (LIS), mirroring the difference-vector construction of Arditi et al. (2024) for refusal directions. Removing 20 LIS dimensions (≤1%) cuts the PS–SYN gap by 37–51% (Appendix D). An unsupervised approximation via embedding nearest neighbors recovers 0.19–0.25 overlap (7–10× random; Appendix D). The confound is compact and consistent across GPT-2 and Pythia-1B.

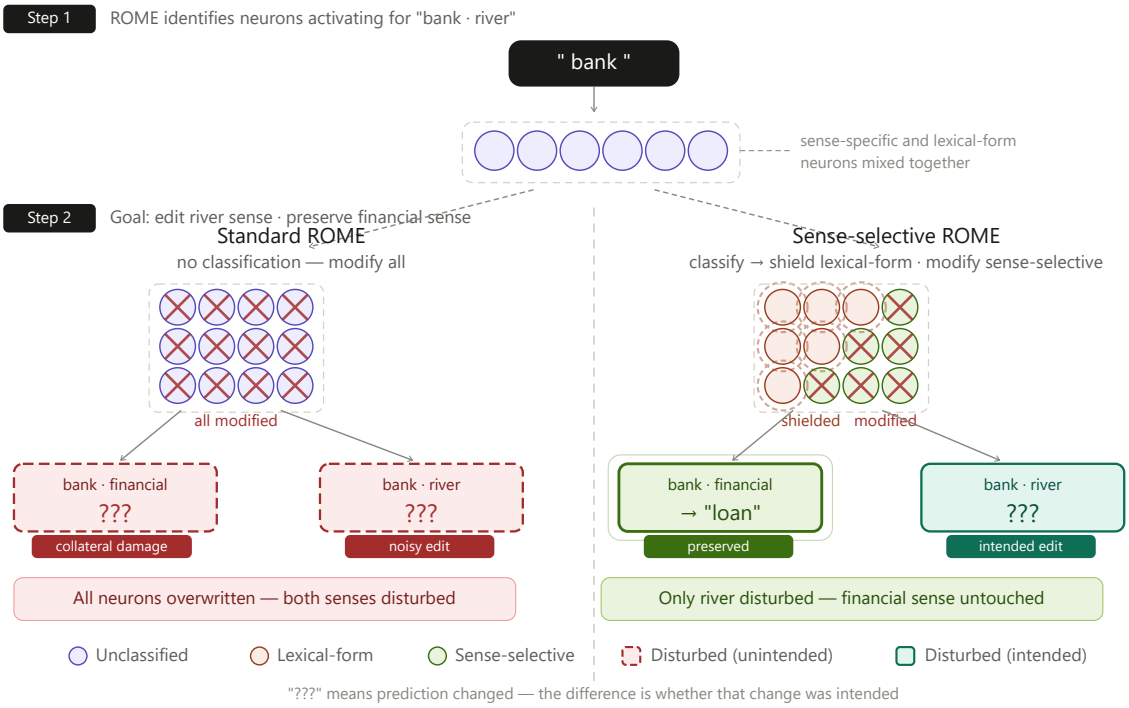

Figure 5: Sense-selective ROME editing (§5.10). The edit installs a counterfactual continuation for sense-A (*bank* = financial); the goal is that sense-B (*river bank*) continuations are unchanged. Standard ROME modifies all activated neurons, disturbing both senses; our variant classifies neurons by SSI, shields sense-blind (lexical-form) neurons, and edits only sense-selective ones.

5.10 Downstream applications

The sense-selective / sense-blind distinction is practically useful (Figure 5).

Word sense disambiguation. On CoarseWSD-20, sense-selective neurons (top-25% by SSI) reach **89.8%** accuracy (GPT-2; **89.3%** Pythia-1B), beating sense-blind (83.6%/84.6%) by 5–6 pp at matched neuron budget.

ROME knowledge editing. We adapt ROME (Meng et al., 2022) to overwrite one sense of a polysemous word (e.g., financial *bank*), scoring *selectivity* as a paired log-probability shift on target vs. held-out sense (Appendix D). Across 53 words, sense-selective restriction (SSI > 2) reaches 0.540; standard ROME 0.519 ($p = 0.002$, Wilcoxon); sense-blind restriction only 0.512. If sense-blind neurons encoded compressed sense, editing them would produce sense-targeted shifts; instead they yield the least targeted edits.

6 Conclusion

A substantial portion of what current metrics call polysemanticity reflects lexical identity rather than concept superposition. Across ten transformers (110M–70B; four architecture families), sense-blind neurons collapse under sense-preserving synonym substitution ($\rho \approx 0$), while sense-selective and monosemous-word neurons retain $\rho \approx 0.7$–0.9. A population-level 2×2 factorial decomposition replicates the finding at every layer: shared word form drives more neuron overlap than shared meaning. The confound is compact, localizable, and removable. Three tools separate lexical identity from genuine compression: a layer-wise diagnostic ratio (R_{lex}), a neuron-level Sense Selectivity Index (SSI), and a low-rank Lexical Identity Subspace (LIS). Applying them improves downstream tasks: word sense disambiguation gains 5–6 pp at matched neuron budget, and sense-selective ROME editing is more targeted than unrestricted ROME ($p = 0.002$). Genuine compression remains a real phenomenon; the standard metric simply overestimates it on polysemous inputs. Whether analogous confounds exist for polysemanticity not driven by polysemous-word inputs is an open question.

Limitations

This paper is a measurement critique: we identify a confound in the standard polysemanticity metric,

not in any specific model's behavior. Several consequences follow. Absolute ablation effects are small (<3 ppl points), as expected when ablating ∼150 of 3k–20k neurons; the primary evidence is the condition ordering itself, not the downstream magnitudes. SemCor's vintage text is validated on modern Wikipedia (Appendix F.1), and the ordering replicates across ten models on four corpora. The 57% confound ratio applies only to polysemous inputs. The corrected score requires sense labels; the LIS removal is more deployable but a fully unsupervised correction remains open (Appendix D). Our Wu-Palmer threshold (<0.5) selects primarily for homonymous senses (e.g., *bank*: financial vs. riverbank), so the metric is most informative for homonymy; whether the same magnitudes hold for finer-grained polysemy (e.g., *mouth*: face vs. river) is open. The neuron classification and causal experiments are run on models up to 13B; the 70B result confirms the population-level ordering but the per-neuron dissociation is not re-verified at that scale. The measurement claim itself is architecture-agnostic and was tested across four architecture families and 110M–70B parameters to make this point. We do not argue that sense-blind neurons are useless; they likely encode a pre-disambiguation representation. We argue only that they should not be counted as evidence of *superposition*: they encode word form rather than compressed concepts.

Ethics Statement

This paper is a measurement-methodology critique of an existing interpretability metric and introduces three diagnostic metrics (R_{lex}, SSI, LIS) computed on publicly released pre-trained transformers (GPT-2, GPT-2-Medium, BERT-base, ELECTRA-base, Pythia 410M–6.9B, Llama-2 7B–13B, and Llama-3-8B). We do not train new models, collect human-subject data, release new datasets, or evaluate fairness- or safety-sensitive applications. All linguistic resources (SemCor, WordNet, CoarseWSD-20) are standard publicly available corpora used in their released form. The ROME knowledge-editing experiments (Section 5.10) use the same procedure as Meng et al. (2022) as a diagnostic, not a deployment tool; our sense-selective variant inherits the dual-use considerations of the underlying ROME framework and adds no new editing capability. No ethics-specific concerns arise beyond those already documented for the base models and corpora cited above.

References

Iyad Ait Hou and Rebecca Hwa. 2026. Parameter-efficient token embedding editing for clinical class-level unlearning. *arXiv preprint arXiv:2603.19302*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Lee Sharkey, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purber, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Joseph Bloom. 2024. Understanding SAE features with the logit lens. *Alignment Forum*. Pre-trained GPT-2 SAEs.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, and 5 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing: A method for rigorously testing interpretability hypotheses. *AI Alignment Forum*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of EMNLP*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Victor Lecomte, Kushal Chandrasekaran, Tom Henighan, and Nelson Elhage. 2025. Incidental polysemanticity. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Daniel Loureiro, Jose Camacho-Collados, Kiamehr Rezaee, and Mohammad Taher Pilehvar. 2022. Language models and word sense disambiguation: An overview and analysis. *Computational Linguistics*, 48(4):789–823.

Xinyu Lyu and 1 others. 2025. Exploring layer-wise representations of English and Chinese homonymy in pre-trained language models. In *Findings of the Association for Computational Linguistics (ACL)*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*.

Go Minegishi, Hiroki Furuta, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Rethinking evaluation of sparse autoencoders through the representation of polysemous words. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

nostalgebraist. 2020. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. 2002. Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2):245–266.

Lee D. Sharkey, Dan Braun, and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders. AI Alignment Forum.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*.

Jiayi Wu and 1 others. 2025. From tokens to semantics: The emergence and stabilization of polysemanticity in language models. In *Proceedings of the Conference on Language Modeling (COLM)*.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 133–138.

Appendix: Table of Contents

App. A Dataset Details ..... p. 10
App. B Full Numerical Results ..... p. 11
App. C Additional Figures ..... p. 12
App. D Supplementary Analyses ..... p. 12
App. E Attention Representations ..... p. 18
App. F Reproducibility & Validation ..... p. 18

A Dataset details

Source corpus and selection. All sentences are drawn from SemCor (Miller et al., 1993), a sense-tagged corpus built on the Brown Corpus with human-annotated WordNet sense labels. From the full corpus (37,176 sentences, 88,334 sense-tagged tokens), we identified 494 content words (nouns and verbs) with at least two senses, each attested in ≥ 8 sentences.

Cleaning pipeline. Three cleaning steps were applied in order:

1. **Function verb exclusion.** We removed 31 high-frequency function verbs (*be, have, do, go, make, get, take, come, give, say*, etc.) whose multiple WordNet "senses" reflect grammatical rather than semantic distinctions.

2. **Cross-word deduplication.** 3,591 sentences that appear under multiple words in SemCor were kept only for their first word, preventing cross-word contamination.

3. **Length filter.** Sentences shorter than 30 characters were removed (typically sentence fragments or captions).

After cleaning, 407 words remain with ≥ 5 sentences per sense (mean: 10.3 sentences per sense; range: 5–15).

Sense distance verification. We computed Wu-Palmer similarity (Wu and Palmer, 1994) between the two selected WordNet synsets for each word. The mean similarity across all 407 word pairs is 0.16 (range: 0.10–0.22). We enforced a maximum threshold of 0.50 to exclude near-synonymous senses.

Condition pair counts. Table 3 shows the number of sentence pairs constructed for each condition. SL and PS pairs are available for all 407 words. The SYN condition requires a WordNet synonym attested in SemCor with ≥ 3 sentences; 172 words have such a synonym for sense A and 160 for sense B. CL pairs are formed by randomly sampling sentences from unrelated words, excluding the current word's own sentences.

POS distribution. Of the 407 words, 156 (38%) are nouns and 251 (62%) are verbs. 13 lemmas appear twice with different POS tags (e.g., *brush* as noun and verb), which we treat as separate entries since they have distinct sense pairs. The full word list, per-word SSI distributions, and SAE feature-level collision data (including the 232 words for which feature #5132 is sense-blind) are released with our code.

B Full numerical results

B.1 Sense-specific editing table

Full perplexity-shift numbers for the GPT-2 sense-specific ablation summarized in Section 5.5. All four groups have matched neuron counts ($n = 142$ each).

B.2 R_{lex} by layer

Table 5 reports R_{lex} computed from cosine similarity at each layer for the 12-layer models. R_{lex} decreases with depth in all models, with BERT and ELECTRA achieving lower minima than GPT-2, reflecting stronger sense separation in bidirectional models.

B.3 Token-embedding baseline

Table 7 reports the embedding-only $R_{\text{lex}}^{\text{emb}}$ for each model, computed from raw token embeddings before any transformer computation. MLP-level R_{lex} exceeds this baseline at layer 0 in all autoregressive models, confirming that the confound is amplified beyond trivial token reuse.

B.4 Sense selectivity index

Table 8 reports the fraction of MLP neurons with SSI > 2 (sense-selective) at each layer for the 12-layer models. In all models, fewer than 1% of neurons are sense-selective at any layer. The choice of threshold (SSI > 2) follows the standard Cohen's d convention for a large effect. Relaxing to SSI > 1.5 roughly doubles the count but does not change the qualitative pattern: sense-selective neurons remain a small minority (<2%) and the probing dissociation (sense accuracy for blind neurons near chance, selective neurons >90%) holds across threshold choices. The sense-blind threshold (SSI < 0.5) similarly follows the small-effect convention; using SSI < 0.3 or SSI < 0.8 shifts counts but preserves the functional split.

B.5 SAE collision: 32k vs. 128k dictionary comparison

Table 9 compares the two SAE dictionary sizes across all five probed layers. The collision ratios are strikingly similar despite a $4\times$ difference in dictionary capacity, confirming that the lexical confound is not resolved by increasing the number of SAE features.

B.6 Synonym subset validation

The full R_{lex} requires the SYN condition, which is available for 251 of 407 words (62%). We verify that this subset is unbiased by testing whether the core lexical confound (the per-word SL–PS cosine gap) differs between synonym-available and synonym-absent words.

Condition	Words	Pairs (total)	Pairs/word (mean)
Same-Lemma (SL)	407	∼40,700	100
Polysemantic (PS)	407	∼40,700	100
Synonym (SYN)	251	∼12,500	50
Cross-Lemma (CL)	407	∼40,700	100

Table 3: Sentence pair counts by condition. SYN coverage is lower because synonyms must be attested in SemCor. All pairs are capped at 200 per word.

Ablation	Δppl$_A$	Δppl$_B$	Specificity
Sense-A neurons	$+0.55$	-0.10	0.66
Sense-B neurons	$+0.10$	$+2.62$	2.52
Sense-blind	-0.07	$+0.17$	0.24
Random	-0.64	-0.32	0.32

Table 4: Sense-specific editing on GPT-2 (53 words, 142 neurons per group). Specificity $= |\Delta\text{ppl}_{\text{target}} - \Delta\text{ppl}_{\text{other}}|$. Sense-selective ablation targets one sense ($6.6\times$ specificity gap over blind); cf. Section 5.5.

Subset comparisons. Table 10 reports Mann-Whitney U tests at five representative layers in each model. No test reaches significance ($p < 0.05$) in any model at any layer; most p-values exceed 0.2. This confirms that the two subsets are statistically indistinguishable on the metric that drives R_{lex}.

POS stratification. The synonym subset is POS-imbalanced: 69% verbs vs. 50% in the non-synonym set ($\chi^2 = 13.8$, $p < 0.001$). Nouns exhibit a larger SL–PS gap than verbs ($p < 0.005$ at all layers in GPT-2), reflecting stronger lexical identity effects for nouns. However, both POS categories show a consistently positive SL–PS gap at every layer in every model, confirming that the confound is universal; the POS imbalance affects magnitude but not the direction or existence of the effect.

Other properties. Wu-Palmer similarity distributions are indistinguishable between subsets (KS $p = 0.98$, Mann-Whitney $p = 0.91$). Sentence counts per word are likewise non-significant (KS $p = 0.68$). Word length shows a small difference (6.3 vs. 5.9 characters; KS $p < 0.001$), but this does not correlate with the SL–PS gap at any layer.

C Additional figures

D Supplementary analyses

Sense accuracy after ablation. On GPT-2, clean (no ablation) sense accuracy averages 73% (measured via diagnostic token probabilities on held-out sentences). After ablating sense-selective neurons, accuracy drops by 8 percentage points; after ablating sense-blind neurons, it drops by <1 pp (indistinguishable from random variation), confirming that sense-selective neurons are causally necessary for the model's sense-discriminating behavior.

Concrete causal example. For the word *bank*, sense-A (financial institution) and sense-B (geographic land) yield distinct diagnostic tokens (*loan*, *account* for sense A; *river*, *shore* for sense B). Ablating sense-A-selective neurons increases perplexity on financial-institution sentences while leaving river-bank sentences unchanged; ablating sense-blind neurons has negligible impact on either.

Cross-model scaling details. Pythia-1B (>5× GPT-2's parameters) shows a qualitatively identical trajectory. Pythia-6.9B confirms the confound persists at the scale where modern interpretability work typically operates. Pythia-12B and LLaMA-2-13B (gated SiLU MLP, a fourth architecture family) maintain the strict condition ordering at every layer; LLaMA-2-13B achieves the lowest absolute similarities but the same pattern, with R_{lex} ranging from 0.91 (layer 0) to 0.25 (layer 38).

The U-shaped trajectory. All conditions exhibit a U-shaped similarity trajectory: high in early layers, lowest in mid-layers, partially recovering in final layers. This is consistent with prior observations that late-layer representations reconverge toward the output vocabulary space (nostalgebraist, 2020; Geva et al., 2022), possibly partially undoing sense-specific processing.

LIS dose-response. Table 11 shows the full dose-response from projecting out the top-k LIS components. The PS–SYN gap closes monotonically; PS similarity rises slightly (removing between-word variance makes within-word representations more uniform) while SYN drops (shared meaning partly encoded along word-form directions). The gap is the informative quantity.

Layer	GPT-2	BERT	ELECTRA	GPT-2-Med
0	.735	.662	.686	.751
1	.629	.566	.528	.700
2	.560	.552	.482	.690
3	.494	.471	.428	.612
4	.446	.398	.367	.549
5	.441	.409	.307	.524
6	.423	.367	.250	.521
7	.418	.313	.268	.498
8	.447	.275	.264	.443
9	.462	.225	.309	.454
10	.399	.360	.381	.440
11	.290	.209	.542	.429
Min	.290	.209	.250	.270

Table 5: R_{lex} (cosine-based) by layer for 12-layer models and GPT-2-Medium (first 12 of 24 layers shown). GPT-2-Medium's minimum of 0.27 occurs at layer 23. BERT achieves its minimum at layer 9 (0.23), ELECTRA at layer 6 (0.25). GPT-2's R_{lex} remains above 0.29 throughout.

Layer	Pythia-12B	LLaMA-2-13B	LLaMA-3-8B
0	.913	.911	.810
5	.516	.406	.278
10	.377	.262	.168
15	.360	.456	.393
20	.422	.476	.367
25	.391	.348	.271
30	.440	.335	.144
last	.288	.261	.129
Min	.288	.250	.129

Table 6: R_{lex} (cosine-based) at representative layers for the larger models. Pythia-12B has 36 layers (last = 35); LLaMA-2-13B has 40 (last = 39); LLaMA-3-8B has 32 (last = 31). All three start with $R_{\text{lex}} \geq 0.81$ at layer 0 and decline with depth. LLaMA-3-8B reaches the lowest minimum (0.13 at layer 31), consistent with stronger disambiguation in later layers of the most recent model. Note that deep-layer R_{lex} CIs in LLaMA-3 widen because the denominator (PS$-$CL) approaches zero; $R_{\text{lex}}^{\text{alt}}$ (without SYN) is more stable at the top of the network (layer-averaged $R_{\text{lex}}^{\text{alt}} = 0.72$).

Model	$R_{\text{lex}}^{\text{emb}}$	R_{lex} (L0)	Exceeds?
GPT-2	0.713	0.735	Yes
GPT-2-Med	0.720	0.751	Yes
BERT-base	0.778	0.662	No
ELECTRA-base	0.727	0.686	No
Pythia-1B	0.821	0.836	Yes
Pythia-6.9B	0.882	0.910	Yes
Pythia-12B	0.891	0.913	Yes
LLaMA-2-13B	0.815	0.911	Yes
LLaMA-3-8B	0.805	0.810	Yes

Table 7: Token-embedding baseline comparison. All autoregressive models show MLP-level R_{lex} exceeding the embedding baseline at layer 0, confirming non-trivial amplification. Bidirectional models (BERT, ELECTRA) show MLP R_{lex} below the embedding baseline at layer 0, suggesting early-layer bidirectional processing partially resolves lexical overlap; the confound nonetheless persists at all layers.

Per-word collateral damage. The collateral-damage experiment (Section 5.5) ablates sense-blind vs. sense-selective neurons for individual words at layer 6. Across 10 polysemous words tested, sense-blind ablation consistently produces low specificity (both senses equally affected), while sense-selective ablation is 5–72$\times$ more targeted. The full per-word results are included in the released code.

Layer	GPT-2	BERT	ELECTRA	GPT-2-Med
0	0.88	0.63	0.57	0.70
1	0.77	0.78	0.64	0.75
2	0.83	0.64	0.76	0.80
3	0.84	0.61	0.74	0.85
4	0.72	0.67	0.81	0.78
5	0.74	0.63	0.83	0.74
6	0.74	0.62	0.75	0.71
7	0.78	0.60	0.69	0.73
8	0.78	0.66	0.63	0.68
9	0.77	0.78	0.54	0.73
10	0.75	0.88	0.53	0.71
11	0.72	0.79	0.49	0.73

Table 8: Fraction (%) of neurons with SSI > 2 at each layer. The denominator is d_{mlp} (3072 for GPT-2/BERT/ELECTRA, 4096 for GPT-2-Med). Pythia-1B's layer 0 has the highest fraction (1.80%) due to its larger MLP producing more sense-differentiating neurons in absolute terms, but the fraction remains very small.

	32k SAE			**128k SAE**		
Layer	Active	Blind	Coll.%	Active	Blind	Coll.%
0	33.6	10.4	31.7	31.5	9.8	32.6
2	27.8	7.4	29.1	26.4	7.6	31.3
5	21.3	4.3	24.1	20.0	4.4	25.7
8	18.6	2.7	17.7	15.2	2.6	22.4
11	18.4	3.9	28.9	15.2	2.9	25.9

Table 9: SAE collision analysis: 32k vs. 128k features. *Active*: mean features active per word (>30% firing rate in either sense). *Blind*: mean sense-blind features (Cohen's $d < 0.5$, fires for both senses). *Coll.%*: collision ratio (blind/active). Both SAEs show similar collision ratios at every layer, confirming that increasing dictionary size does not resolve the confound.

Model	**L0**	$\mathbf{L}\frac{n}{4}$	$\mathbf{L}\frac{n}{2}$	$\mathbf{L}\frac{3n}{4}$	$\mathbf{L}_{n-1}$	**min** p
GPT-2	.43	.34	.45	.56	.37	.28
GPT-2-Med	.50	.28	.38	.42	.28	.28
BERT-base	.57	.31	.14	.18	.16	.14
ELECTRA	.48	.33	.50	.53	.26	.26
Pythia-1B	.62	.20	.33	.30	.16	.16
Pythia-6.9B	.80	.43	.26	.13	.25	.13

Table 10: Mann-Whitney U test p-values comparing the per-word SL–PS cosine gap between synonym-available ($n = 251$) and synonym-absent ($n = 156$) words. No test reaches significance; the minimum p-value across all models and layers is 0.13 (Pythia-6.9B, layer 24).

	PS–SYN gap	R_{lex}	**ΔPS**	**ΔSYN**
GPT-2 (d=3072, layer-averaged)				
Baseline	0.119	0.65	—	—
$k = 10$ (0.3%)	0.087	0.52	+0.008	−0.032
$k = 20$ (0.7%)	0.047	0.41	+0.019	−0.054
$k = 50$ (1.6%)	−0.028	<0	+0.031	−0.116
Pythia-1B (d=8192, layer-averaged)				
Baseline	0.097	0.59	—	—
$k = 10$ (0.1%)	0.062	0.47	+0.016	−0.019
$k = 20$ (0.2%)	0.033	0.29	+0.025	−0.040
$k = 50$ (0.6%)	−0.032	<0	+0.037	−0.092

Table 11: LIS dose-response ablation. Removing $k = 20$ components ($\leq 1\%$ of dimensions) reduces R_{lex} by 37–51%. At $k = 50$, the PS–SYN gap reverses, eliminating the confound.

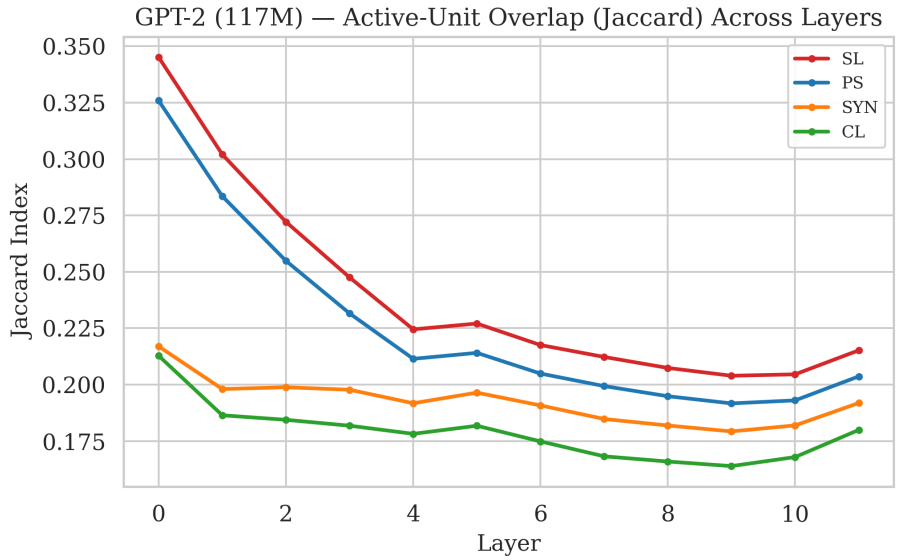

Figure 6: Jaccard overlap of active-neuron sets across layers (GPT-2). The condition ordering SL > PS > SYN > CL is consistent with the cosine similarity results, confirming that the lexical confound affects which neurons fire, not just how strongly they fire.

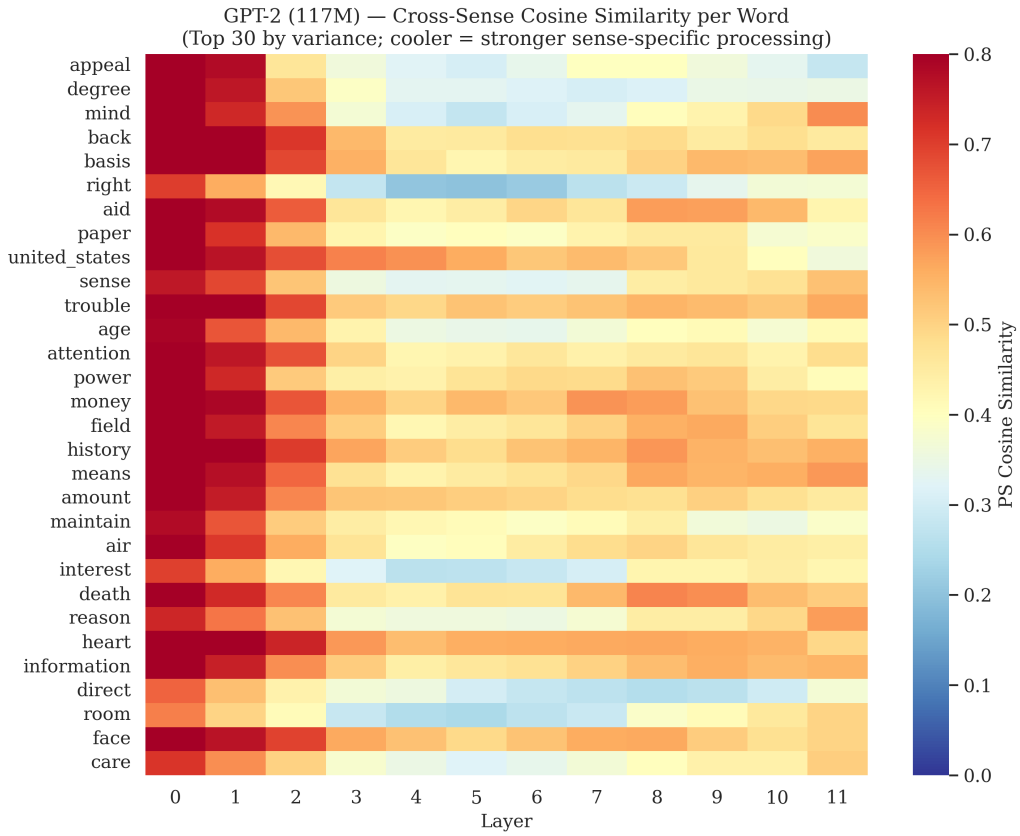

Figure 7: Cross-sense (PS) cosine similarity per word per layer in GPT-2 (top 30 words by variance). Cooler colors indicate lower similarity (stronger sense-specific processing). Most words exhibit a U-shaped trajectory with a dip in mid-layers, though the depth and timing of the minimum varies substantially across words. Sense separation is word-specific, not driven by a shared mechanism.

Unsupervised LIS detection. To test whether the LIS can be approximated without sense annotations, we replace WordNet synonyms with nearest neighbors from GPT-2's embedding layer ($n = 393$

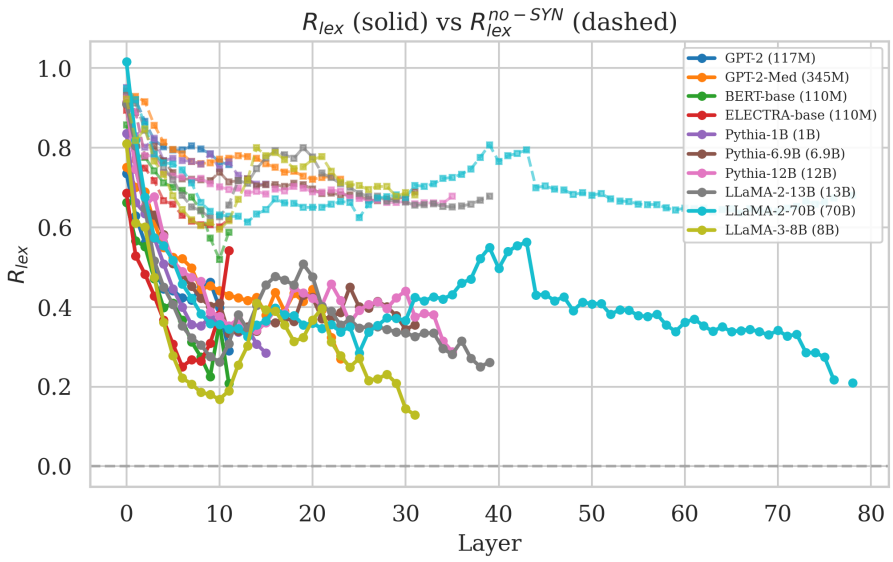

Figure 8: R_{lex} (solid) vs. $R_{\text{lex}}^{\text{no-syn}}$ (dashed) for all ten models. The alternative metric $R_{\text{lex}}^{\text{no-syn}} = (\bar{M}_{\text{PS}} - \bar{M}_{\text{CL}})/(\bar{M}_{\text{SL}} - \bar{M}_{\text{CL}})$ does not require synonym data and can be computed for all 407 words. The qualitatively identical trajectories confirm that our findings do not depend on synonym coverage (62% of words, validated by subset invariance test; see Appendix B.6).

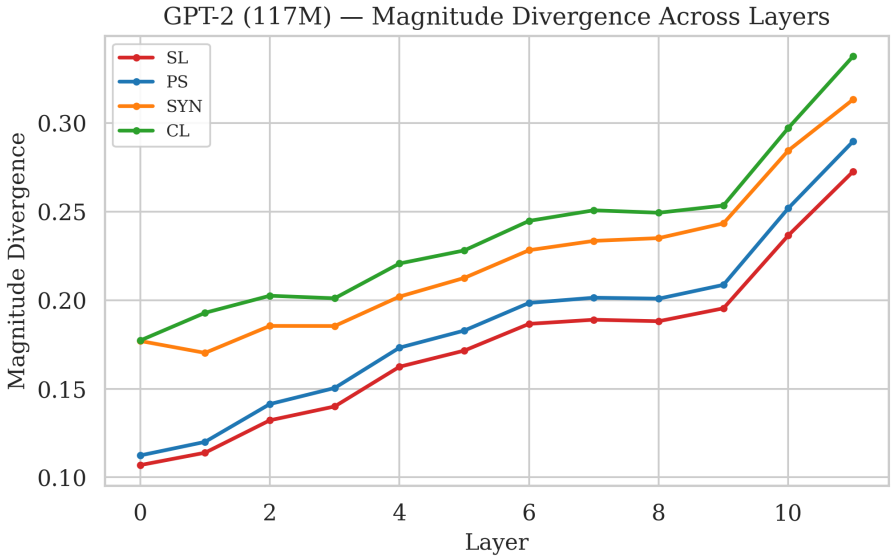

Figure 9: Mean magnitude divergence ($\Delta = \frac{1}{|A \cap B|} \sum_{i \in A \cap B} |a_i - b_i|$) by condition and layer. Unlike cosine similarity (which shows a U-shaped trajectory), magnitude divergence increases monotonically with depth, reflecting growing activation scale differences as representations become more specialized. The condition ordering is reversed (CL > SYN > PS > SL) because more dissimilar conditions produce larger activation differences among their shared neurons.

pairs). The resulting subspace overlaps with the WordNet-based LIS at 0.19–0.25 (7–10× above a random baseline of 0.026). This partial recovery suggests the subspace is identifiable without labels, though a fully unsupervised method remains open.

ROME selectivity: full setup. We apply ROME (Meng et al., 2022) to GPT-2-small, layer 5 (the locality layer ROME identifies for these models). For each of 53 polysemous target words, we construct two evaluation contexts: one anchoring sense-A (e.g., *"The lender approved the bank . . . "*) and one anchoring sense-B (*"We hiked along the riverside bank . . . "*), with held-out continuations that diagnose each sense (e.g., *loan, account* vs. *river, shore*). ROME computes a rank-one update that overwrites sense-A's continuation with

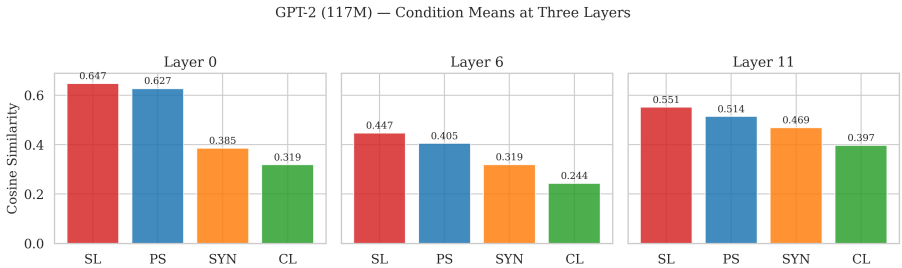

Figure 10: Condition means at three representative layers in GPT-2 (layers 0, 6, 11). At layer 0, PS and SL are close while SYN and CL are distant. By layer 6, the gap between PS and SL widens. At layer 11, all conditions retain the same ordering with moderate separation.

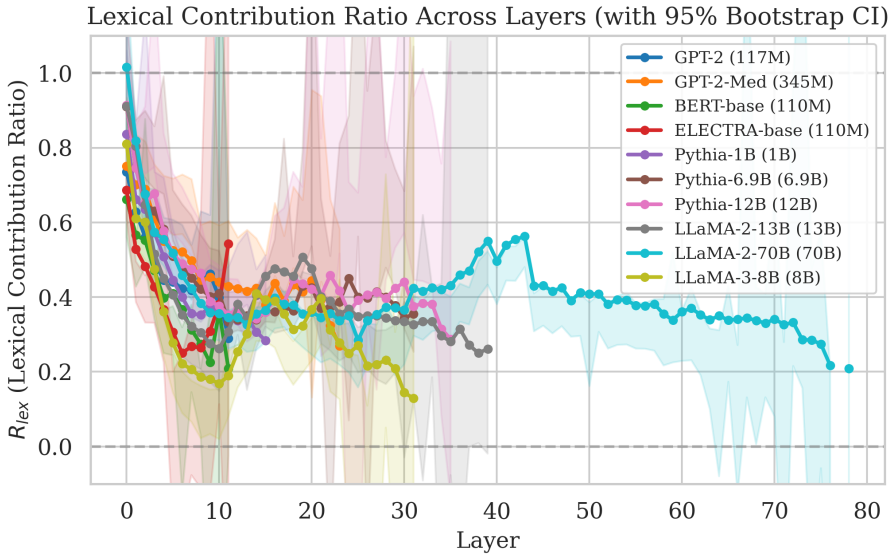

Figure 11: R_{lex} (lexical contribution ratio) across layers for all ten models with 95% bootstrap CIs. The lexical contribution decreases with depth but never reaches zero in any model. Larger models (LLaMA-2-70B, cyan) show sustained $R_{\text{lex}} \approx 0.35$–$0.45$ across 80 layers.

a counterfactual target token. We compare three neuron-restriction conditions on the same edit problem: (i) **standard ROME**, no restriction; (ii) **sense-selective**, only neurons with SSI > 2 are eligible for modification; (iii) **sense-blind**, only neurons below the SSI threshold are eligible (other neurons are mask-shielded). *Selectivity* is the paired log-probability shift score

$$s = \sigma\big(\Delta \log P^{A}_{\text{target}} - \Delta \log P^{B}_{\text{target}}\big),$$

mapped through a logistic to $[0, 1]$, where $\Delta \log P^{A}_{\text{target}}$ is the change in log-probability of the counterfactual target under the sense-A prompt and $\Delta \log P^{B}_{\text{target}}$ is the analogous shift on the sense-B prompt. Higher s means the edit moves sense-A toward the counterfactual without dragging sense-B along. Significance is the Wilcoxon signed-rank test on per-word paired differences between sense-selective and standard.

Anchor-persistence setup. For each polysemous word w with senses A and B we form five condition sets of token-position activations at a target MLP layer: PS_A, PS_B (the polysemous word in sense-A and sense-B sentences), SYN_A, SYN_B (the same sentences with w replaced by a sense-preserving WordNet synonym), and CL (a random sample of 20 sentences from non-overlapping polysemous words). To avoid noise from near-synonyms (e.g., *allow/let/permit* are roughly interchangeable in context), we audit synonym quality at the activation level: a word is admitted if (i) its two senses are sufficiently separated ($1 - \cos(\bar{\mathbf{a}}_{\text{PS}_A}, \bar{\mathbf{a}}_{\text{PS}_B}) > 0.15$), (ii) SYN_A patterns more with PS_A than with PS_B in activation cosine, and (iii) symmetrically for SYN_B. This filters from 66 candidate words in POLYSEMY-CLEAN down to 13 admitted words used in Section 5.6. The Sense Selectivity Index is computed on the PS_A/PS_B split

(Cohen's d; Section 3.2). Sense-blind = top-50 neurons by mean magnitude on PS, restricted to SSI < 0.5; sense-selective = top-50 with SSI > 2; random-active = top-50 by mean magnitude without the SSI filter. For each group we compute the shared activation $\bar{a}_c = \min_n\big(\overline{|a_n|}_{c_A}, \overline{|a_n|}_{c_B}\big)$ (the activation that survives across both senses or both synonyms), then $\rho = (\bar{a}_{\text{SYN}} - \bar{a}_{\text{CL}})/(\bar{a}_{\text{PS}} - \bar{a}_{\text{CL}})$. The monosemous control uses the same pipeline with no SSI filter: 15 unambiguous content words (*democracy*, *electricity*, *hospital*, *telephone*, *dictionary*, *refrigerator*, *elephant*, *ocean*, *mountain*, *forest*, *island*, *doctor*, *lawyer*, *garbage*, *rabbit*), each paired with one sense-preserving synonym (*self-government*, *electric power*, *medical center*, *phone*, *lexicon*, *fridge*, *pachyderm*, *sea*, *peak*, *woodland*, *isle*, *physician*, *attorney*, *trash*, *bunny*).

The ten models in Table 2 are evaluated at three layers each, chosen at approximately 25%/50%/75% depth: GPT-2 at L3/L6/L9; GPT-2-Med at L8/L16/L22; BERT and ELECTRA at L3/L6/L9; Pythia-410M at L6/L12/L18; Pythia-1B at L4/L8/L12; Pythia-6.9B at L8/L16/L24; Llama-2-7B at L8/L16/L24; Llama-3-8B at L8/L16/L24; Llama-2-13B at L10/L20/L30. dtype is float32 for the smaller models, float16 for Llama-2-7B and Llama-3-8B, bfloat16 for Llama-2-13B. Reported ρ in Table 2 is the median across the three layers per model.

E Attention representations

We apply the same four-condition decomposition to attention head outputs (concatenated per-head output vectors at the target word position). The ordering SL > PS > SYN > CL holds in attention representations at every layer in all models, with R_{lex} values comparable to MLP values in early layers. Since the residual stream sums attention and MLP outputs, the confound propagates to SAEs trained on `resid_post` (Section 5.3).

F Reproducibility details

Compute. All experiments were run on a university HPC cluster. Models up to 1B parameters used NVIDIA V100 GPUs (16 GB VRAM); Pythia-6.9B required an A100 GPU (40 GB VRAM); Pythia-12B and LLaMA-2-13B each required an NVIDIA GH200 GPU (96 GB VRAM). Decomposition experiments took approximately 2–4 hours per model on a single GPU (6–8 hours for the 12B and 13B models). The SAE collision experiments took approximately 20 minutes (32k SAE) and 30 minutes (128k SAE) on a V100.

Software. Models were loaded using HuggingFace Transformers (v4.36+) in `float32` precision. SAE experiments used TransformerLens (v1.6+) and SAE Lens (v3.0+). Random seed was fixed at 42 for all experiments. Bootstrap confidence intervals used 10,000 resamples stratified by word.

Code and data. All code (decomposition, SSI computation, anchor-persistence, LIS extraction, ROME variant, SAE collision analysis, plotting) and all per-word results files will be released under a permissive open-source license upon acceptance via an anonymized repository during the review period and a public repository thereafter. The sense-annotated dataset is derived from publicly available SemCor and CoarseWSD-20 corpora; our preprocessed splits will be released alongside the code.

Metric definitions (formal). For completeness, we provide the formal definitions of all metrics.

Cosine similarity: $\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|}$

Jaccard overlap: A neuron is active if its absolute activation exceeds the per-sentence median. $J = |A \cap B|/|A \cup B|$.

Magnitude divergence: $\Delta = \frac{1}{|A\cap B|}\sum_{i\in A\cap B}|a_i - b_i|$

Sense Selectivity Index: $\text{SSI}_i = |\mu_i^{(A)} - \mu_i^{(B)}|/s_i^{\text{pooled}}$ where $s_i^{\text{pooled}} = \sqrt{((n_A-1)s_A^2 + (n_B-1)s_B^2)/(n_A+n_B-2)}$. Neurons with SSI > 2 are classified as sense-selective.

F.1 Modern corpus validation (CoarseWSD-20)

A potential concern is that SemCor draws on the 1960s Brown Corpus. To rule out the possibility that our findings are artifacts of vintage text, we replicate the *full* four-condition decomposition on **CoarseWSD-20** (Loureiro et al., 2022), a dataset of 20 polysemous nouns with sense-annotated **Wikipedia** sentences. We construct all four conditions (SL, PS, SYN, and CL) by pairing CoarseWSD-20 sense labels with WordNet synonyms whose sentences are drawn from the same Wikipedia corpus (different words' entries), ensuring the SYN condition uses modern text throughout. We run the decomposition on GPT-2, BERT-base, and Pythia-1B.

The results replicate the SemCor findings on modern text. The critical PS > SYN gap is signif-

Model	SL	PS	SYN	CL	PS>SYN
GPT-2	.90–.60	.89–.47	.31–.45	.26–.45	$p < .001$
BERT-base	.70–.58	.54–.30	.09–.32	.05–.35	$p < .01$
Pythia-1B	.99–.46	.99–.27	.09–.28	.09–.26	$p < .001$

Table 12: Modern corpus validation on CoarseWSD-20 (Wikipedia sentences, 20 nouns). Ranges show cosine similarity across layers (layer 0 to final). The critical PS > SYN gap (indicating lexical form drives more overlap than shared meaning) is statistically significant at every layer in all three models (Wilcoxon signed-rank, Holm-Bonferroni corrected). R_{lex} bootstrap 95% CI remains above zero everywhere.

icant at every layer in all three models, confirming that lexical form is a stronger driver of MLP neuron overlap than semantic similarity even on Wikipedia text. GPT-2's R_{lex} ranges from 0.98 (layer 0) to 0.57 (layer 11), closely matching the SemCor trajectory (0.74–0.29). Pythia-1B shows an even stronger confound ($R_{\text{lex}} \geq 0.41$ at all layers). The full ordering SL $>$ PS $>$ SYN $>$ CL holds at the majority of layers; at a few late layers, SYN and CL converge (both approach 0.3–0.4), consistent with the SemCor pattern where all conditions converge at depth. The lexical confound is a property of how transformers represent polysemous words, not an artifact of the SemCor corpus.