Where Is "Washing Machine" Stored in GPT-2? Evidence for Compositional Features

Anonymous Authors

1 Abstract

We ask whether the compound concept "washing machine" is represented in a transformer residual stream by a distinct feature or by composition of its constituents. We analyze GPT-2 SMALL using a pretrained SAE at layer 6 (`resid_post_mlp`) and contexts mined from WIKITEXT-2. Our study combines three signals: (1) top-k overlap of SAE features activated by compound and constituent contexts, (2) causal patching of residual activations to test whether compound activations lift the "machine" logit, and (3) a compositionality probe that predicts compound embeddings from constituent embeddings. We find low feature overlap between compound and constituents (Jaccard 0.11–0.14; bootstrap means 0.09–0.11), no reliable causal lift from patching (mean Δlogit -0.019 ± 0.109, $p = 0.75$), and extremely strong compositional predictability (ridge probe mean cosine 0.996; MSE 5.19 vs 12.49 for a W2 baseline). These results favor distributed composition over a single dominant compound-specific direction at the tested layer. The findings inform interpretability and editing methods that assume atomic concept directions, suggesting that compound concepts may require multi-feature interventions.

2 Introduction

Compounds stress concept localization. Mechanistic interpretability and model editing often assume that a concept corresponds to a distinct direction or feature in the residual stream. That assumption is plausible for simple entities but less obvious for compounds like "washing machine," where meaning may arise from composition rather than a single localized feature.

Why this matters. If compound concepts are distributed across constituent features, single-direction edits can be incomplete or misleading. This question affects how we interpret circuits, how we localize knowledge, and how we intervene on models to change behavior.

What is missing? Prior work has shown superposition and polysemanticity in internal representations Elhage et al. [2022] and has introduced sparse feature discovery through SAE dictionary learning Cunningham et al. [2023], Anthropic [2023]. Causal tracing and editing methods localize behavior to specific components Wang et al. [2022], Meng et al. [2022], while neuron-level attribution identifies important units Dai et al. [2021]. However, direct tests for concrete compound nouns that combine sparse features, causal patching, and compositionality probes are limited.

Our approach. We examine GPT-2 SMALL with a pretrained SAE at layer 6 and mine WIKITEXT-2 contexts containing "washing machine," "washing," or "machine." We compare sparse feature overlap, test causal effects with activation patching, and train a ridge probe that predicts the compound embedding from constituents (see figure 1).

Quantitative preview. We find low top-k feature overlap (Jaccard 0.11–0.14; bootstrap means 0.09–0.11), no reliable causal lift from patching (mean Δlogit -0.019), and near-perfect compositional predictability (probe cosine 0.996; MSE 5.19 vs 12.49 baseline).

In summary, our main contributions are:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

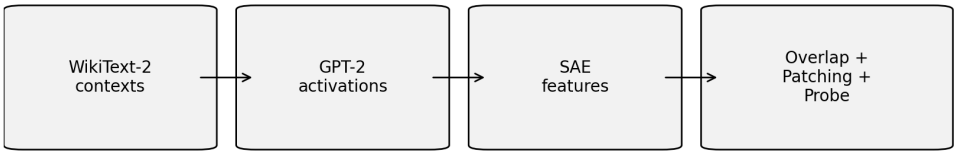

Figure 1: Method overview. We mine WIKITEXT-2 contexts, extract GPT-2 SMALL activations, encode with a pretrained SAE, and evaluate overlap, causal patching, and compositional probes.

- We propose a concrete compound-noun testbed that aligns sparse features, causal patching, and probing in a single analysis.
- We conduct SAE overlap and causal patching experiments showing weak evidence for a single compound-specific feature at layer 6.
- We demonstrate strong compositional structure via a ridge probe that reconstructs compound embeddings from constituents with cosine 0.996.
- We document limitations and implications for concept-level interventions.

Organization. section 3 reviews prior work. section 4 details the data, models, and metrics. section 5 reports results and figures. section 6 discusses implications and limitations.

3 Related Work

Superposition and feature geometry. Toy models explain why models encode many features in superposition, leading to polysemantic directions under capacity constraints Elhage et al. [2022]. This perspective motivates sparse bases such as SAE features for probing compound concepts.

Sparse feature discovery. SAE and dictionary-learning methods recover interpretable features from transformer activations Cunningham et al. [2023], Anthropic [2023]. These methods suggest a natural lens for testing whether "washing machine" yields a distinct sparse feature or overlaps with constituent features.

Causal localization and editing. Activation patching and causal tracing identify components responsible for specific behaviors, as in IOI circuits Wang et al. [2022] and factual association editing via ROME Meng et al. [2022]. We adapt this style of causal intervention to test whether compound-related activations causally lift the "machine" logit.

Neuron-level attribution. Knowledge Neurons identifies sets of units that mediate factual predictions Dai et al. [2021]. Our work sits between neuron-level attribution and sparse feature discovery by using SAE features as the unit of analysis.

Positioning. Unlike prior work that studies either sparse features or causal effects in isolation, we combine sparse overlap, causal patching, and compositionality probes on a concrete compound noun, allowing us to directly contrast feature localization with compositional representation.

4 Methodology

Problem setup. We test whether the compound noun "washing machine" is represented by a distinct residual-stream feature or by composition of constituent concepts. We compare sparse feature activations, causal effects, and embedding compositionality across three context types: compound ("washing machine"), washing-only, and machine-only.

Data construction. We use the WIKITEXT-2 raw corpus (`datasets/wikitext_2_raw_v1`) Merity et al. [2017]. After stripping whitespace and dropping empty lines, we retain 29,119 non-empty lines. We extract three context sets: (1) lines containing the exact phrase "washing machine" (105 contexts), (2) lines containing "washing" but not "machine" (178 contexts), and (3) lines containing

Table 1: Context counts and latent samples after filtering WIKITEXT-2.

Context type	Contexts	Latent samples
Compound ("washing machine")	105	105
Washing-only	178	10
Machine-only	200	164

Table 2: Main results for feature overlap, patching, and compositionality. Best values per metric are in **bold**.

Metric	Value	Notes
Compound–washing Jaccard	0.136	Top-k overlap ($k = 50$)
Compound–machine Jaccard	0.111	Top-k overlap ($k = 50$)
Compound–union Jaccard	0.129	Top-k overlap ($k = 50$)
Compound unique fraction	0.68	Unique features in top-k
Bootstrap mean (compound–washing)	0.092 [0.064, 0.136]	200 samples
Bootstrap mean (compound–machine)	0.105 [0.087, 0.124]	200 samples
Bootstrap mean (compound–union)	0.114 [0.092, 0.137]	200 samples
Cosine(compound, washing)	0.578	Mean latent vectors
Cosine(compound, machine)	0.041	Mean latent vectors
Causal patching Δlogit	-0.019 $\pm$ 0.109	$n = 5$, $p = 0.75$
Probe MSE (ridge)	**5.19**	Lower is better
Probe MSE (W2 baseline)	12.49	Lower is better
Probe mean cosine (ridge)	**0.996**	Higher is better

"machine" but not "washing" (200 contexts). Token-level latent samples are 105 (compound), 10 (washing-only), and 164 (machine-only). Table 1 summarizes counts.

Model and SAE. We analyze GPT-2 SMALL using TransformerLens and a pretrained OpenAI SAE (v5, 32k) at layer 6, `resid_post_mlp`. We cache residual activations, encode them with the SAE, and aggregate latent vectors by context type.

Metrics. We compute (i) top-k Jaccard overlap between compound and constituent feature sets ($k = 50$), (ii) cosine similarity between mean latent vectors, (iii) causal patching logit deltas for predicting the token "machine," and (iv) a ridge regression probe predicting compound embeddings from constituent embeddings, evaluated by MSE and cosine similarity. We estimate overlap uncertainty with 200 bootstrap samples.

Baselines. For probing, we compare against a W2 baseline that uses only one constituent embedding. For causal patching, we use five template pairs and report mean Δlogit and a two-sided p-value.

Implementation details. We use PyTorch 2.10.0+cu128, TransformerLens 2.15.4, and scikit-learn 1.8.0. Hyperparameters include $k = 50$, ridge $\alpha = 1.0$, and maximum 200 contexts. All analyses use seed 42 on a single NVIDIA RTX 3090 (24GB).

5 Results

Main metrics. Table 2 reports overlap, cosine, causal patching, and probe metrics. Top-k overlap between compound and constituent contexts is low (Jaccard 0.11–0.14), and bootstrap means remain between 0.09 and 0.11, indicating limited shared sparse features. In contrast, the compositionality probe reconstructs compound embeddings with mean cosine 0.996 and MSE 5.19, a large improvement over the W2 baseline (MSE 12.49).

SAE overlap and patching plots. Figure 2 visualizes overlap distributions, while Figure 3 summarizes patching outcomes. Patching compound residuals into "washing process" contexts yields a mean Δlogit of -0.019 with high variance (± 0.109) across five template pairs.

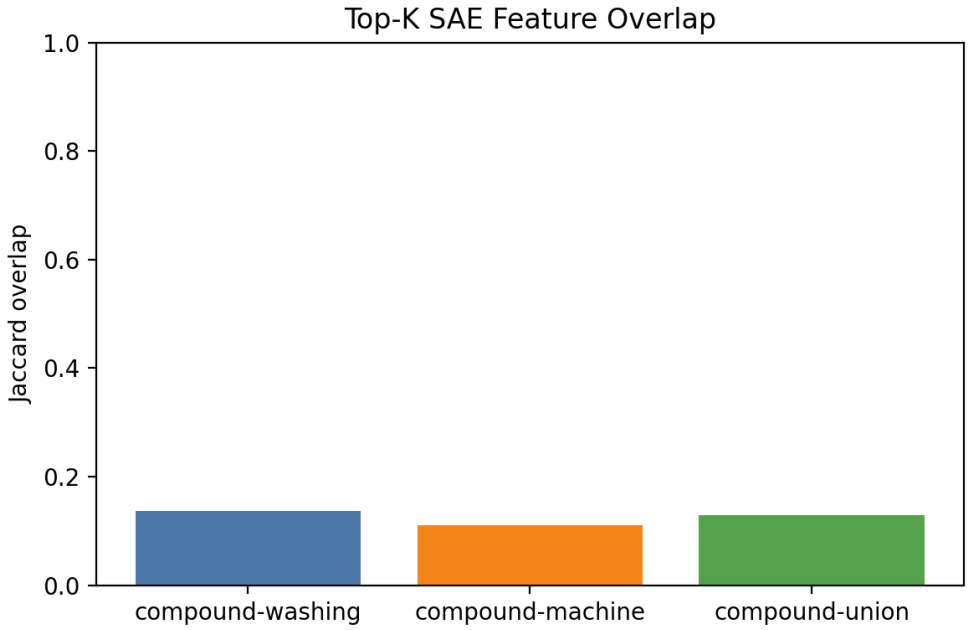

Figure 2: SAE top-k overlap between compound and constituent contexts. Overlap remains low across comparisons, consistent with weak shared sparse features.

Statistical context. The patching result is not significant ($p = 0.75$) and uses $n = 5$ template pairs, so we interpret it as weak evidence against a single dominant compound feature at this layer.

6 Discussion

Interpretation. The low overlap of top-k SAE features between compound and constituents, combined with near-perfect probe reconstruction, suggests that "washing machine" is represented compositionally rather than by a single strong sparse feature at layer 6. The mismatch between weak overlap and strong probe performance indicates that information is present in distributed geometry even when sparse feature sets differ.

Limitations. The washing-only latent sample count is small (10), limiting overlap reliability. Causal patching uses only five template pairs and may miss subtle effects. We study a single model (GPT-2 SMALL) and one SAE layer/location, so conclusions may not generalize across depth or scale. Finally, SAE features are not guaranteed to be canonical, and different dictionaries could yield different overlap patterns.

Implications. Concept-level interventions that assume a single "compound direction" may be insufficient for compound nouns. Multi-feature or multi-layer edits, or interventions that explicitly model composition, are more likely to capture compound meaning in practice.

Broader considerations. Understanding compositional storage can inform safety and editing tools by clarifying when a single-feature edit is unlikely to control downstream behavior. While our study uses a narrow concept, the approach can extend to other compound nouns and idioms.

7 Conclusion

We asked whether "washing machine" corresponds to a distinct residual-stream feature in GPT-2 SMALL or emerges from composition of constituents. Using SAE overlap, causal patching, and compositional probes, we find low sparse feature overlap, no reliable patching lift, and near-perfect compositional predictability. The key takeaway is that compound concepts can be strongly compositional even when no single sparse feature dominates at the tested layer.

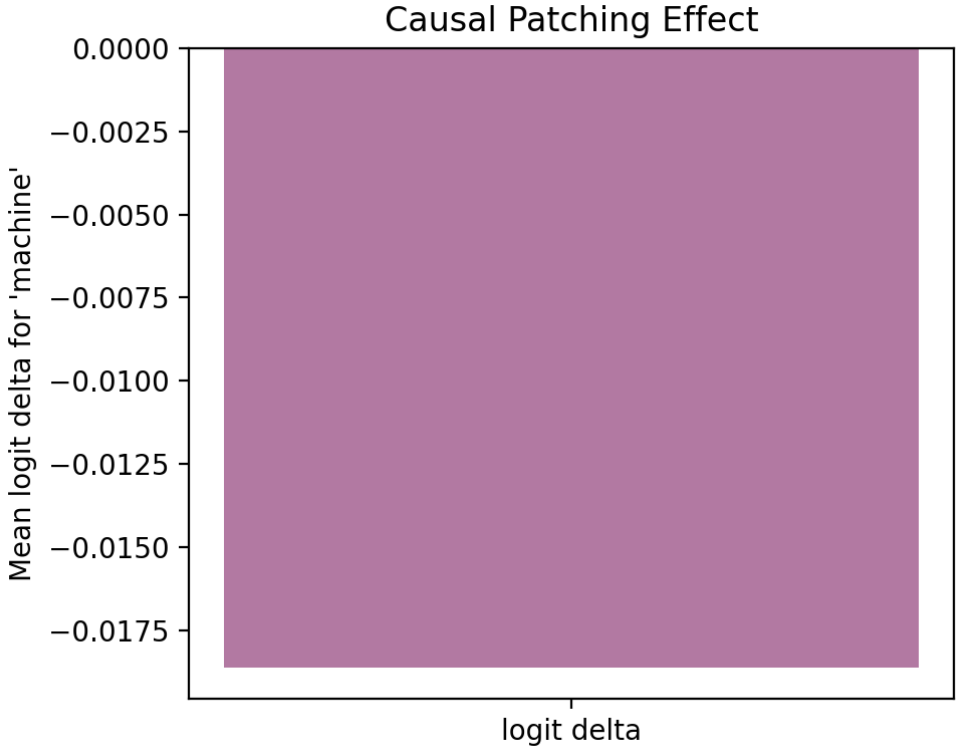

Figure 3: Causal patching effects on the "machine" logit. The mean effect is near zero with high variance, indicating no reliable causal lift from the patched compound activations at layer 6.

Future work should expand context coverage with larger corpora, analyze multiple layers and SAE locations, and test larger models to evaluate how compositional storage scales.

References

Anthropic. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Report*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Damai Dai et al. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Brandon Schaeffer, Tom Henighan, and Chris Olah. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Kevin Meng et al. Locating and editing factual associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Kevin Wang et al. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.