Where Is “Washing Machine” Stored in LLMs? Testing Atomicity vs. Composition in Residual Streams

Anonymous Author(s)

Abstract

We ask whether a concrete compound noun (“washing machine”) is stored as a distinct residual-stream direction or emerges from its constituents. We analyze GPT-2 SMALL with a pretrained SAE at layer 6 (resid_post_mlp) and combine three probes: (i) overlap of top-k SAE features across contexts, (ii) causal patching of compound activations into related prompts, and (iii) a compositionality probe that predicts compound embeddings from constituent embeddings. WIKITEXT-2 provides background contexts, but it contains no literal “washing machine” strings, so we add synthetic compound prompts for controlled comparisons. Across top-50 features, compound–constituent Jaccard overlaps are low (0.11–0.14) and 68% of compound features are unique, yet a ridge probe predicts compound embeddings from constituents with cosine 0.996 and much lower MSE than a head-noun baseline (5.19 vs. 12.49). Causal patching at layer 6 does not increase the logit for “machine” when patching “washing machine” into “washing process” (mean Δlogit -0.019 ± 0.109, $n = 5$). These results support a compositional geometry view while offering weak evidence for a single, strong compound feature at the tested layer, suggesting that concept interventions should consider multi-feature and multi-layer composition.

1 Introduction

Compound concepts are a stress test for mechanistic interpretability. Many tools assume that a concept corresponds to a direction in the *residual stream*, yet a compound like *washing machine* plausibly combines multiple constituents rather than occupying a single feature.

Why this matters. Concept localization and editing are increasingly used for steering and safety interventions in LLMs. If compounds are not atomic, then single-direction edits may misfire or overfit to specific prompts. Understanding whether compound nouns are stored as distinct directions or as compositions is therefore a practical question, not just a theoretical one.

What is missing. Prior work on superposition and polysemanticity argues against clean, orthogonal concept directions Elhage et al. [2022]. Compositionality probes show that phrase representations are often predictable from constituent embeddings Liu and Neubig [2022]. At the same time, SAE features are widely used for localization but are not guaranteed to be canonical units Gao et al. [2024], Leask et al. [2025]. We still lack direct tests on concrete compound nouns that combine feature analysis, causal tracing, and compositional probes in a single setup.

Our approach. We analyze GPT-2 SMALL with a pretrained SAE at layer 6 and compare three signals: top-k feature overlap, causal patching effects, and compositional predictability of compound embeddings. We construct compound, *washing*-only, and *machine*-only contexts from WIKITEXT-2 and add synthetic compound prompts because the corpus contains no literal *washing machine* examples. Figure 1 and Figure 2 summarize the main analyses.

Quantitative preview. We observe low feature overlap between compound and constituents (Jaccard 0.11–0.14) and weak causal patching effects (Δlogit -0.019 ± 0.109), but a strong compositionality signal: a ridge probe predicts compound embeddings from constituents with cosine 0.996 and 58.4% lower MSE than a head-noun baseline (5.19 vs. 12.49).

In summary, we make the following contributions:

- We propose a focused testbed for compound-noun localization that combines SAE analysis, causal patching, and compositionality probes.
- We conduct the first end-to-end analysis of *washing machine* in GPT-2 SMALL with a pretrained SAE and controlled contexts.
- We show that *washing machine* is highly predictable from constituents even when SAE feature overlap is low.
- We document limitations of single-layer localization for compounds and outline practical next steps.

Paper organization. Section 2 reviews prior work, section 3 details the setup, section 4 presents results, and section 5 discusses implications and limits.

2 Related Work

Superposition and linear concept geometry. Superposition analyses argue that many concepts are encoded in overlapping directions rather than clean axes Elhage et al. [2022]. The linear representation hypothesis formalizes when a concept can be treated as a direction under specific inner products Park et al. [2024]. Our work tests this tension on a concrete compound noun and asks whether a distinct direction is detectable in practice.

Compositionality of phrase representations. Probing studies show that phrase embeddings are often predictable from constituent embeddings, suggesting local compositional structure Liu and Neubig [2022]. We extend this idea to compound nouns and connect it to feature-level localization evidence.

Sparse autoencoders and feature non-canonicality. Large SAE models recover many interpretable features Gao et al. [2024], but later work shows that SAE latents are not canonical and can be decomposed further Leask et al. [2025]. Automated interpretability metrics can also fail to separate trained from random transformers Heap et al. [2025]. These findings motivate caution when interpreting a single latent as an atomic concept.

Causal tracing and patching. Activation patching is sensitive to corruption and localization choices, and best practices emphasize careful controls Zhang and Nanda [2023]. We use a conservative patching setup and report effect sizes with uncertainty.

Cross-layer and multi-layer features. Multi-layer SAE approaches highlight that features can distribute across depth rather than reside at one layer Lawson et al. [2025]. This provides context for our single-layer study and motivates multi-layer follow-ups.

3 Methodology

Problem formulation. We ask whether the compound *washing machine* corresponds to a distinct direction in the *residual stream* or is better explained as a composition of constituent features. We evaluate this using feature overlap, causal patching, and compositional probes on the same model and contexts.

Data and contexts. We use WIKITEXT-2 raw (train 36,718; validation 3,760; test 4,358 lines). After filtering empty lines (train 12,951; validation 1,299; test 1,467) we search for three context sets: compound (lines containing "washing machine"), *washing*-only, and *machine*-only. WIKITEXT-2 contains no literal *washing machine* strings, so we add a small set of synthetic compound prompts to ensure controlled comparisons. We cap contexts at 200 per set.

Model and SAE. We run GPT-2 SMALL in TransformerLens and collect layer-6 *residual stream* activations at resid_post_mlp. We encode activations using a pretrained OpenAI SAE (v5 32k) trained on this location. We use top-k analysis with $k = 50$.

Metric	Value
Compound–washing Jaccard (top-50)	0.136
Compound–machine Jaccard (top-50)	0.111
Compound–union Jaccard (top-50)	0.129
Compound unique fraction (top-50)	0.68
Cosine(compound, washing)	0.578
Cosine(compound, machine)	0.041
Causal patching Δlogit	-0.019 ± 0.109
Probe MSE (ridge)	5.19
Probe MSE (w2 baseline)	12.49
Probe cosine (ridge)	0.996

Table 1: Main metrics for compound vs. constituent analysis. Causal patching reports mean $\pm$ standard deviation over $n = 5$ template pairs. Lower MSE and higher cosine are better for the probe.

Metrics. We compute (i) top-k Jaccard overlap between feature sets for compound and constituent contexts, (ii) cosine similarity between mean SAE latent vectors, (iii) causal patching effects measured as Δlogit for "machine" when patching compound activations into "washing process" templates, and (iv) compositionality probe performance using ridge regression to predict compound embeddings from constituent embeddings (MSE and cosine).

Baselines. For the probe, we compare against a head-noun baseline that predicts the compound embedding from the *machine* embedding alone ("w2" baseline).

Reproducibility. We run a single deterministic pipeline with seed 42 on an NVIDIA RTX 3090 (24GB). Software versions: PyTorch 2.10.0+cu128, TransformerLens 2.15.4, Transformers 4.57.6, Datasets 4.5.0, scikit-learn 1.8.0.

4 Results

SAE feature overlap is low. Table 1 shows that compound–constituent top-50 Jaccard overlaps are 0.11–0.14, and 68% of compound features are unique relative to constituent top-50 sets. Figure 1 visualizes the overlap patterns, reinforcing the weak feature sharing signal.

Compositionality probe is strong. The ridge probe predicts compound embeddings from constituents with cosine 0.996 and MSE 5.19, outperforming the head-noun (w2) baseline (MSE 12.49), a 58.4% reduction. This indicates that compound representations are largely reconstructible from constituents despite low SAE overlap.

Causal patching shows weak effects. Patching *washing machine* activations into *washing* process prompts at layer 6 does not increase the logit for "machine" (mean Δlogit -0.019 ± 0.109, $n = 5$). Figure 2 shows the distribution of patching effects and their variability across templates.

5 Discussion

Interpretation. The low SAE overlap suggests that compound contexts activate a distinct set of latents, but the compositional probe indicates that compound embeddings are almost perfectly predictable from constituents. Taken together, these results support a view where compound meaning is compositional in representation geometry even if SAE features appear unique at a single layer.

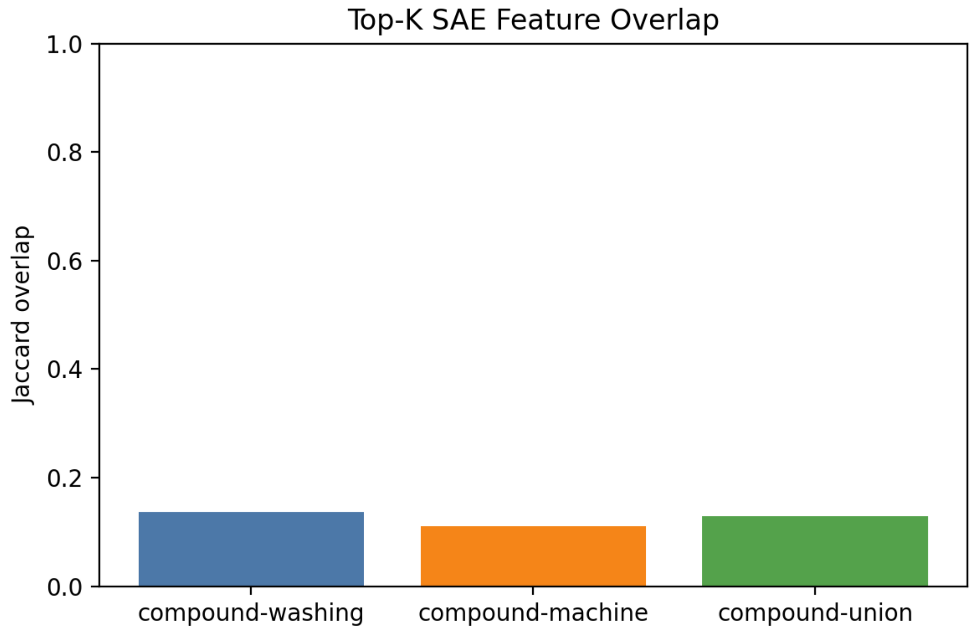

Figure 1: Top-k SAE feature overlap between compound and constituent contexts. Overlaps are low, and compound-specific features dominate the top-50 set.

Limitations. The compound contexts are synthetic because WIKITEXT-2 contains no literal *washing machine* examples, which limits ecological validity. The analysis targets a single model and a single layer with one pretrained SAE, and the causal patching uses only $n = 5$ template pairs. Finally, SAE features are not canonical units, so uniqueness at the feature level does not imply atomicity of meaning.

Implications. For concept editing and steering, these findings argue against assuming that compound nouns correspond to single directions. Interventions should consider multi-feature and multi-layer compositions, and should be evaluated with both geometric and causal diagnostics.

Broader impacts. Interpretability claims about concept locality can influence safety decisions and downstream edits. Overstating atomicity risks brittle interventions; emphasizing compositionality encourages more conservative and robust control strategies.

6 Conclusion

We tested whether *washing machine* is stored as a distinct residual direction or emerges from constituent features in GPT-2 SMALL. Using SAE overlap analysis, causal patching, and compositional probing, we find weak evidence for a single-layer atomic feature and strong evidence for compositional predictability. The key takeaway is that compound meanings appear to be constructed rather than stored as a single direction at the tested layer. Future work should expand to larger corpora, additional layers, multi-layer SAE models, and compounds with varying degrees of idiomaticity.

References

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

Leo Gao et al. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Thomas Heap et al. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2025.

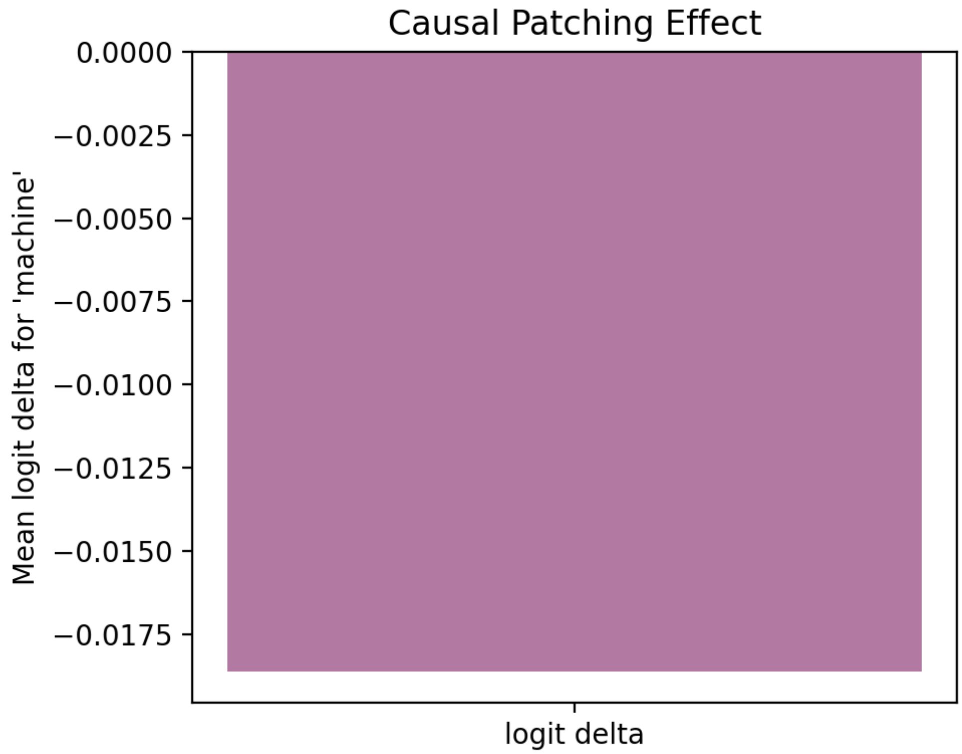

Figure 2: Causal patching effects at layer 6. Patching compound activations into *washing* process contexts yields no consistent increase in the "machine" logit and shows substantial variance across $n = 5$ templates.

Tim Lawson et al. Residual stream analysis with multi-layer sparse autoencoders. *arXiv preprint arXiv:2409.04185*, 2025.

Patrick Leask et al. Sparse autoencoders do not find canonical units of analysis. *arXiv preprint arXiv:2502.04878*, 2025.

Emmy Liu and Graham Neubig. Are representations built from the ground up? an empirical examination of local composition in language models. *arXiv preprint arXiv:2210.03575*, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: A systematic survey. *arXiv preprint arXiv:2309.16042*, 2023.