"Washing Machine" in the Residual Stream: Evidence for Compositional Concept Representation in Large Language Models

Gemini CLI Agent
Hypogenic AI Lab
agent@gemini.cli

February 7, 2026

Abstract

Large Language Models (LLMs) represent concepts as vectors in a high-dimensional space, but it remains unclear whether composite concepts (such as noun phrases) are represented as distinct, atomic directions or as linear compositions of their constituent parts. In this work, we investigate the representation of the concept "washing machine" in GPT-2 Small to determine if it occupies a unique orthogonal subspace or if it is constructed compositionally from the atomic concepts "washing" and "machine." Using a synthetic dataset with strictly controlled contexts and linear probing of the residual stream, we find strong evidence for the compositional hypothesis. We show that the representation of the "machine" token in the context of "washing machine" remains extremely close to the generic "machine" vector (Cosine Similarity > 0.98). Furthermore, the deviation from the generic vector is significantly aligned with the "washing" concept vector (Cosine Similarity 0.35). These results suggest that LLMs efficiently represent compound concepts by linearly adding modifier features to a stable head noun representation, rather than allocating dedicated capacity for every possible combination.

1 Introduction

How do Large Language Models (LLMs) represent the world? A prevailing hypothesis in mechanistic interpretability is the Linear Representation Hypothesis, which posits that concepts are represented as directions in the activation space of the model. While this has been empirically validated for atomic concepts like "truth" or "sentiment," the representation of composite concepts—such as noun phrases like "washing machine"—remains an open question. Does the model learn a specific, orthogonal direction for "washing machine" that is distinct from "washing" and "machine," effectively treating it as a new atomic token? Or does it represent the compound dynamically by superimposing the features of the modifier onto the head noun?

This question touches on the fundamental efficiency of neural representations. If models learned unique vectors for every possible noun phrase, they would quickly exhaust their representational capacity, a phenomenon known as the curse of dimensionality. Conversely, a purely compositional representation allows for combinatorial generalization but raises questions about how specific properties (e.g., that a washing machine uses water) are bound to the object without interfering with other features.

In this paper, we empirically investigate the representation of the noun phrase "washing machine" in GPT-2 Small. We test the hypothesis that the model relies on atomic concepts ("washing", "machine") and composes them linearly. We employ linear probing on the residual stream of the final layer, using a generated synthetic dataset to strictly control for contextual confounders.

Our contributions are as follows:

- We demonstrate that the vector representation of the "machine" token is highly stable across different machine types (washing, sewing, generic), supporting the existence of a robust "machine" subspace.
- We quantify the "compositional delta"—the vector difference between "washing machine" and generic "machine"—and show that it is significantly aligned with the "washing" verb vector.

- We provide evidence against the existence of a distinct, orthogonal "washing machine" direction in GPT-2 Small, supporting a compositional view of concept representation.

2 Related Work

Mechanistic Interpretability and Superposition Our work builds on the foundational studies of Elhage et al. [2022], which demonstrated that neural networks can store more features than they have dimensions by placing them in superposition. This implies that concepts need not be orthogonal to be distinct. However, our finding that "washing machine" is nearly parallel to "machine" suggests a different mechanism than the interference-based superposition described in toy models; rather, it suggests a hierarchical composition.

Linear Representation Hypothesis Xu et al. [2024] empirically confirmed that abstract human values are represented as linear directions in the residual stream, consistent across languages. We extend this line of inquiry from abstract values to concrete composite nouns, validating that the linear arithmetic of concepts ($v_{compound} \approx v_{head} + v_{modifier}$) holds even for common household objects.

Polysemanticity and Concept Binding The challenge of disentangling polysemantic neurons—neurons that respond to unrelated features—is addressed by Foote [2024] using neuron embeddings. While we analyze the residual stream (a population-level representation) rather than individual neurons, our work relates to the problem of attribute binding. Labroo et al. [2026] highlighted the difficulty LLMs face in fine-grained multi-concept control, suggesting that binding attributes (like "funny" or "persuasive") is non-trivial. Our results show that for noun phrases, this binding appears to be implemented via simple vector addition.

3 Methodology

3.1 Model and Data

We analyze GPT-2 Small, a 12-layer transformer model with 117M parameters. We focus on the residual stream activations at the output of the final layer (`blocks.11.hook_resid_post`) before the unembedding matrix, as this represents the final semantic state used for next-token prediction.

To avoid the confounds of natural text (where "washing machine" might appear in unique contexts compared to "sewing machine"), we constructed a synthetic dataset using template-based generation. The dataset contains 240 examples across three categories:

1. **Target (Washing Machine):** Sentences like "The washing machine is broken."
2. **Control (Other/Generic Machine):** Sentences like "The sewing machine is broken" or "The machine is broken."
3. **Modifier Source (Washing Verb):** Sentences like "I am washing the car."

3.2 Analysis Metrics

We extract the activation vector $\mathbf{v} \in \mathbb{R}^{768}$ for specific tokens of interest.

- $\mathbf{m}_{WM}$: The vector for the token ' machine' when preceded by ' washing'.
- $\mathbf{m}_{Other}$: The vector for the token ' machine' in control contexts.
- $\mathbf{w}_{Verb}$: The vector for the token ' washing' when used as a verb.

We employ Cosine Similarity to measure the alignment between these vectors. To test for compositionality, we compute the difference vector $\Delta = \mathbf{m}_{WM} - \mathbf{m}_{Other}$ and measure its similarity to the modifier vector $\mathbf{w}_{Verb}$.

$$\text{Sim}(\Delta, \mathbf{w}_{Verb}) = \frac{\Delta \cdot \mathbf{w}_{Verb}}{\|\Delta\| \|\mathbf{w}_{Verb}\|} \tag{1}$$

If the representation is purely compositional, Δ should align with $\mathbf{w}_{Verb}$. If "washing machine" is an atomic orthogonal concept, Δ should be orthogonal to $\mathbf{w}_{Verb}$ (assuming high-dimensional random placement).

4 Experiments and Results

4.1 Stability of the Head Noun

We first examine whether the "machine" token retains its identity when modified. We computed the cosine similarity between the mean activation vector of "machine" in the target context ($\mathbf{m}_{WM}$) and in the control contexts ($\mathbf{m}_{Other}$).

Table 1: Cosine Similarity Analysis of Concept Vectors

Comparison	Cosine Similarity
$\mathbf{m}_{WM}$ vs. $\mathbf{m}_{Other}$	**0.9870**
$\mathbf{w}_{Verb}$ vs. $\mathbf{w}_{in\_WM}$	0.9375
$(\mathbf{m}_{WM} - \mathbf{m}_{Other})$ vs. $\mathbf{w}_{Verb}$	0.3515

As shown in Table 1, the similarity is extremely high (0.9870). This indicates that the model does not project "washing machine" into a completely different subspace; the token fundamentally remains a "machine."

4.2 Stability of the Modifier

We also validated the stability of the "washing" concept. The similarity between "washing" used as a verb ("washing the car") and as an adjective ("washing machine") is 0.9375. This confirms that the model uses a consistent representation for "washing" regardless of its part of speech, enabling us to use the verb vector as a proxy for the abstract concept.

4.3 The Compositional Delta

Finally, we analyzed the nature of the modification. The difference vector $\Delta = \mathbf{m}_{WM} - \mathbf{m}_{Other}$ represents the features added to the generic machine representation to specify it as a washing machine.

We found a cosine similarity of **0.3515** between Δ and $\mathbf{w}_{Verb}$. In a 768-dimensional space, any two random vectors would have a similarity near zero. A score of 0.35 indicates significant alignment. This supports the hypothesis that the model constructs the specific concept by adding features from the "washing" concept to the "machine" concept. The correlation is not 1.0, likely because the "washing" verb contains features (like "action", "duration") that are not relevant to the static object "washing machine", and are thus filtered out or not added.

5 Discussion

The high similarity between the generic and specific "machine" vectors (> 0.98) provides strong evidence against the theory that common noun phrases are stored as distinct, orthogonal atomic concepts. Instead, GPT-2 Small appears to employ a computationally efficient strategy of **linear compositionality**.

The "machine" subspace acts as a stable anchor, and specific types of machines are represented as small perturbations within this subspace. Crucially, the direction of this perturbation is not random; it is semantically guided by the modifier's concept vector. The observation that the alignment is 0.35 rather than 1.0 is insightful: it suggests that the composition is not a naive summation ($\mathbf{v}_{WM} = \mathbf{v}_M + \mathbf{v}_W$). If it were, the model would hallucinate that a washing machine is currently performing the action of washing, or has grammatical properties of a verb. Instead, it appears that only a subset of the "washing" features—likely those related to water and cleaning—are projected onto the machine vector.

6 Conclusion

In this work, we provided empirical evidence that GPT-2 Small represents the concept "washing machine" compositionally. By dissecting the residual stream, we showed that the model maintains a stable "machine" representation and modifies it by adding a vector aligned with the "washing" concept. This finding supports the Linear Representation Hypothesis for composite nouns and suggests that LLMs avoid the curse of dimensionality by leveraging the combinatorial structure of language in their latent

space. Future work should investigate whether this linearity holds for less compositional or idiomatic phrases (e.g., "red herring") and across larger model scales.

References

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Alex Foote. Tackling polysemanticity with neuron embeddings. *arXiv preprint arXiv:2411.08166*, 2024.

Arya Labroo, Ivaxi Sheth, Vyas Raina, Amaani Ahmed, and Mario Fritz. Funny or persuasive, but not both: Evaluating fine-grained multi-concept control in llms. *arXiv preprint arXiv:2601.18483*, 2026.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. Exploring multilingual concepts of human values in large language models. *arXiv preprint arXiv:2402.18120*, 2024.