A Compositional Calculus for Semantic Synergy in Language Model Embeddings

Abel Jansma — ICML 2026 — Mechanistic Interpretability Workshop

TL;DR

How do language models represent “kick the ball” vs. “kick the bucket”? The latter has *semantic synergy.* We develop a calculus, based on the algebra of language, to quantify this embedding synergy in transformers. We find:

- Idioms have much larger synergy than literals.
- Synergy emerges across layers and training epochs.
- Synergy vectors can steer transformers toward figurative or literal representation.
- **Non-compositional embeddings are measurable and interpretable.**

1. Background

Latent space arithmetic, like the famous “`king−man+woman≈queen`” pattern, doesn’t work in transformers. The attention mechanism of transformers makes their embedded semantics contextual and often *non-compositional*.

We develop latent space *calculus*: a more precise way to disentangle representations that recovers interpretability when meaning is not compositionally encoded.

2. Latent Space Calculus

We aim to capture how meaning changes when words are composed. Mathematically, this corresponds to the derivative of the embedding function T over the partial order that describes word-composition (the *mereology, Fig. 1*). This calculus was recently developed in [1] as a vector-valued extension of Möbius inversion [2,3,4]. For a 3-word phrase “*abc*”, the derivative of the full phrase is a residual vector q given by

$$q = \partial T(abc) = T(abc) - T(ab) - T(bc) + T(b)$$

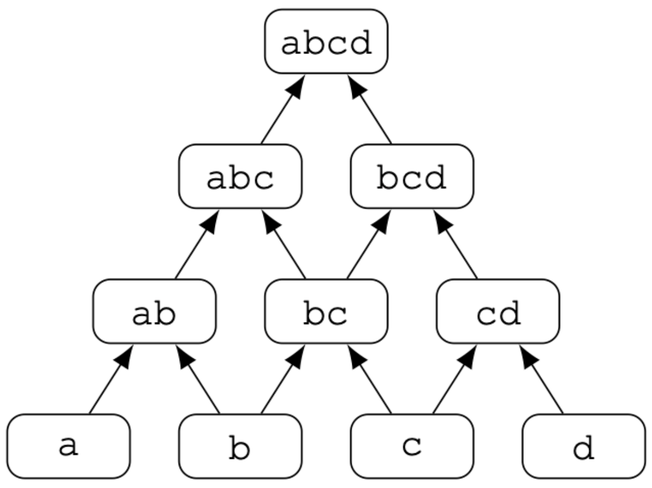

Figure 1. The span-mereology of the 4-word string “abcd”.

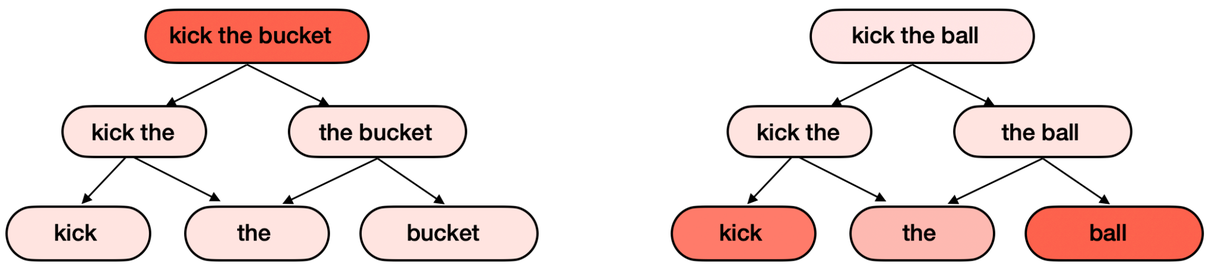

Figure 2. The mereology of two 3-word phrases: one idiom, and one literal. Coloured by $|\partial T|$.

3. Idioms display synergistic embeddings

The meaning of an idiom resides only at the top node of its mereology. Intuitively, one would expect the synergy $q = \partial T(abc)$ to be large for the idiom, and small for the literal (see Fig. 2). Furthermore, one would expect q to carry the meaning of the idioms, but not of the literals. Over two corpora of 107 & 836 English 3-word idioms, this is indeed the case (Fig. 3).

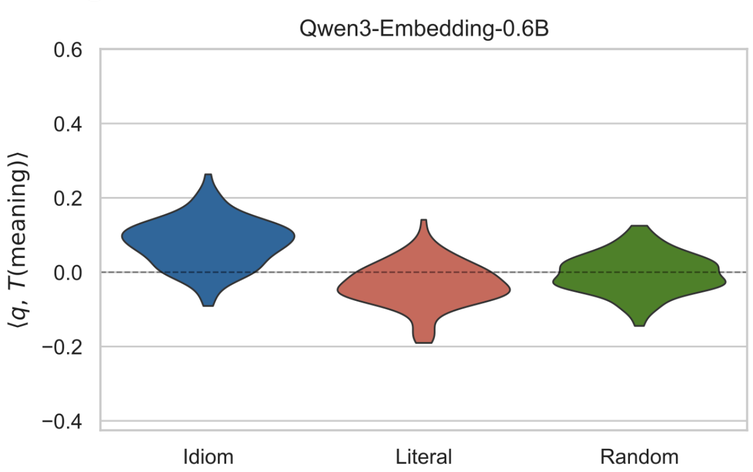

Figure 3. The inner product between the synergy vector q and the phrase meaning is positive for idioms, and negative for literals (Cohen’s $d = 1.80,\ \ p < 10^{-28}$). Random phrases have inner product ≈ 0.

4. Semantic synergy across layers and training

In sentence embedding models like Qwen3-Embedding-0.6B, the synergy vectors of idioms and literals mostly separate at the output layer, whereas next-token predictors like Qwen3-0.6B and Pythia-1B show strong separation both in intermediate and final layers, mostly emerging late in training (Fig. 4).

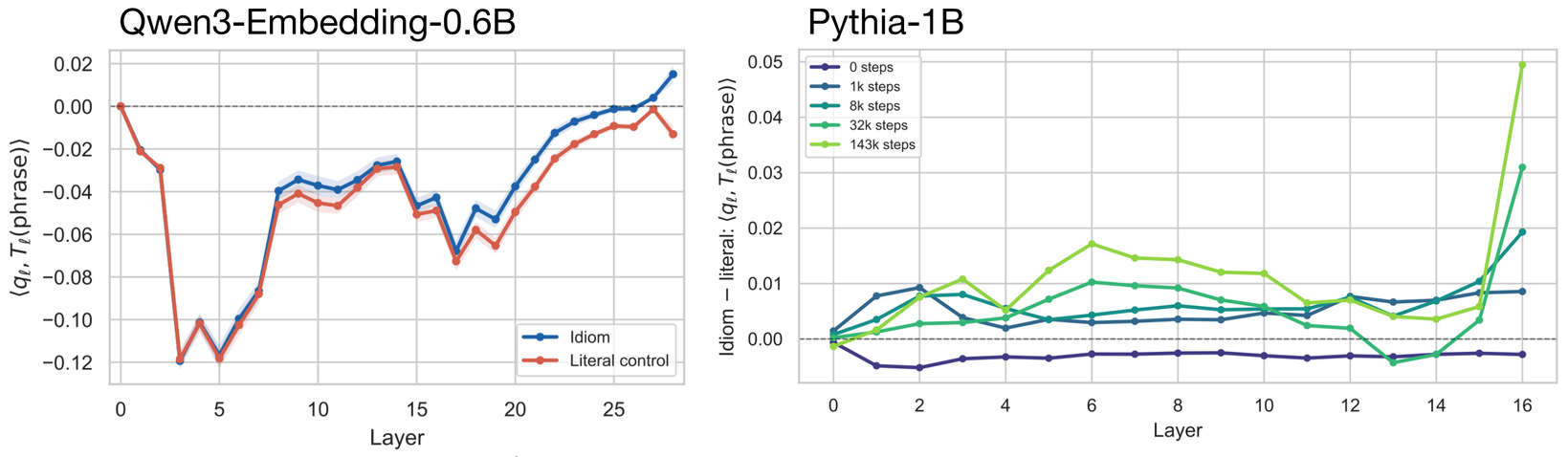

Figure 4. Layer-wise idiom/literal separation, with different checkpoints shown for Pythia-1B.

5. Synergistic Steering Vectors

The average idiom synergy vector $\overline{q}$ can steer embeddings towards figurative or literal representations of a withheld set of phrases. Steering with $\overline{q}$ works as well as using supervised approaches.

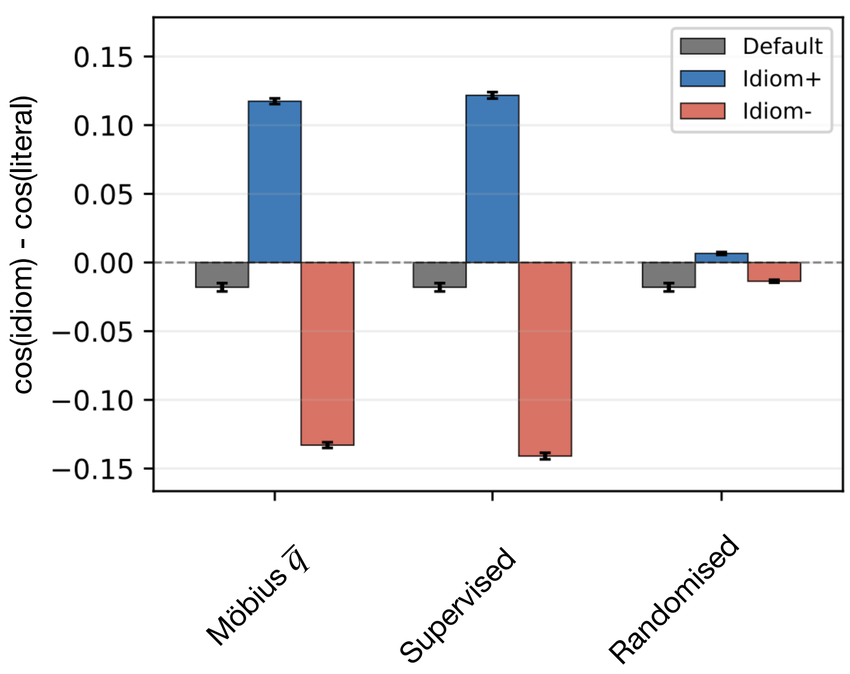

Figure 5. Adding $\overline{q}$ to the embedding z of a withheld idiom moves it closer to the idiomatic meaning, whereas subtracting it brings it closer to the literal meaning. The effect is comparable with supervised methods.

References

[1] Forré and Jansma. "Möbius transforms and Shapley values for vector-valued functions on weighted directed acyclic multigraphs.”arXiv:2510.05786 (2025).
[2] Rota. ”On the foundations of combinatorial theory I. Theory of Möbius Functions." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 2.4 (1964): 340-368.
[3] Jansma. "Mereological approach to higher-order structure in complex systems: From macro to micro with Möbius." *Physical Review Research* 7.2 (2025): 023016.
[4] Jansma. "Decomposing interventional causality into synergistic, redundant, and unique components." *Advances in Neural Information Processing Systems* 38 (2026): 57581-57601.

Paper

Abel Jansma
a.a.a.jansma@uva.nl

Dutch Institute for Emergent Phenomena, University of Amsterdam
Institute for Logic, Language and Computation
Institute of Physics