A Compositional Calculus for Semantic Synergy in Language Model Embeddings

Abel Jansma [1 2 3]

Abstract

We introduce *semantic synergy*: a training-free measure of non-compositional representation in language models, obtained by taking the discrete derivative of a phrase embedding over its sub-span structure. Formally, semantic synergy is the Möbius inverse of the embedding function on the partial order of contiguous sub-spans. Across two embedding models (Qwen3-Embedding-0.6B & all-mpnet-base-v2) and 107 manually curated short English idiom/literal pairs, semantic synergy strongly separates idiomatic from literal phrases (Cohen's $d \approx 1.80$–1.81, $p < 10^{-28}$). The measure outperforms alternative residuals, distinguishes non-compositional proper names in a supporting experiment, and yields steering directions that move held-out phrase embeddings towards, or away from, idiomatic interpretations. Layer-wise extraction in Qwen3-0.6B and Pythia-1B models shows that the non-compositional structure emerges mainly in intermediate layers, and becomes strong only late in training. An appendix replication on a larger, non-manually-curated 836-pair list generated by an LLM confirms the same findings, with somewhat smaller absolute effect sizes but stronger statistical significance. Span-Möbius residuals therefore provide a lightweight algebraic probe of compositional structure in embedding spaces and a bridge toward hidden-state mechanistic analysis.

1. Introduction

Mechanistic interpretability asks which structural features of a model's representations correspond to identifiable semantic properties. A long-standing hypothesis in distributional semantics is that sentence meaning is *compositional*: the meaning of a phrase is determined by the meanings of its parts and the way they are combined (Pagin & Westerståhl, 2010; Baroni et al., 2014). That is essentially a *mereological* statement, which we will make mathematically precise and useful in Section 2. Modern transformer-based language models encode semantics in a very contextual way through the attention mechanism, which can introduce nontrivial compositional effects (Reimers & Gurevych, 2019). This motivates our question: *where does compositional structure break down in language model embedding spaces, and how can we measure this precisely?*

Our approach is deliberately narrower than full circuits-style mechanistic interpretability work, and sits between compositionality analysis, representation probing, concept/steering vectors, and mechanistic interpretability of internal states (Kim et al., 2018; Elhage et al., 2021). Two recent studies used scalar-valued measures of compositionality to measure violations of compositionality in LLMs (Jansma, 2025b; Wold et al., 2026). Our contribution is complementary: we compute an explicitly *vector-valued* residual that can be localized across layers, inspected by nearest neighbours in embedding space, and used as a steering direction. We therefore frame our measure of semantic synergy as a *representation-level diagnostic*: an algebraic decomposition that reveals where additive composition fails, while leaving token-level causal tracing and circuit identification to future work.

In some static word embeddings, it has been observed that `king−man+woman` $\approx$ `queen` (Mikolov et al., 2013), a phenomenon known as latent space arithmetic. This, however, does not transfer cleanly to modern contextual phrase embeddings (see Appendix A). Our work can also be seen as developing a latent space *calculus*, grounded in the Möbius inversion theorem. This lifts the Linear Representation Hypothesis (LRH) (Elhage et al., 2022) to a "homomorphic" representation hypothesis, where the additive structure of the representations reflects the algebraic, or *mereological* structure of the input semantics, rather than simply a conjunction of concepts. We will see that this nonlinearity, however, does not hinder interpretable decompositions of the representations.

[1]Dutch Institute for Emergent Phenomena, the Netherlands [2]Institute for Theoretical Physics, University of Amsterdam, the Netherlands [3]Institute for Logic, Language and Computation, University of Amsterdam, the Netherlands. Correspondence to: Abel Jansma <a.a.a.jansma@uva.nl>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

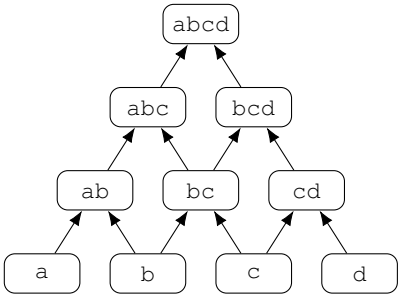

Figure 1. Span mereology for the 4-token string `abcd`. Each node is a contiguous span; arrows point from part to whole.

2. Semantic Synergy via Möbius Inversion

Mereology. We follow Jansma (2025a), and define a *mereology* to be a partial order $(P, \leq)$ of parts and wholes, where $x \leq y$ means that x is a part of y. For a phrase $w_1 w_2 \cdots w_n$, we in particular define the *span mereology*, where elements are all contiguous sub-spans, ordered by sub-span containment (Figure 1).

Möbius inversion for vector-valued functions. Why is Möbius inversion the right calculus here? Ordinary finite differences are the discrete derivative on a total order: they subtract what is already explained by earlier points on a line. Möbius inversion is the generalisation of this idea from total orders to more intricate algebraic structures, namely arbitrary locally finite partial orders (Rota, 1964). When the partial order is the powerset lattice ordered by inclusion, Möbius inversion reduces to the familiar inclusion-exclusion principle. On the span mereology, it gives a precise correction for the fact that sub-spans overlap and therefore cannot be treated as independent contributors to the whole phrase embedding.

Let $T : P \to \mathbb{R}^d$ be the sentence embedding function. The mereological assumption is that *a priori* any part can contribute independently to the embedding:

$$T(x) = \sum_{y \leq x} q(y), \tag{1}$$

where $q : P \to \mathbb{R}^d$ is the *intrinsic* contribution of each span. The following theorem was first proved by Rota (1964) (and extended to the vector-valued case in (Forré & Jansma, 2025)):

Theorem 1 (Möbius inversion on $\mathbb{R}$-modules)**.** *Let $(P, \leq)$ be a locally finite poset and $T, q : P \to \mathbb{R}^d$. Then $T(x) = \sum_{y \leq x} q(y)$ if and only if $q(x) = \sum_{y \leq x} \mu(y, x)\, T(y)$, where μ is the Möbius function of P.*

It is easy to see that on the span mereology, the Möbius function $\mu(a, b)$ is ± 1 only when a is part of the square below b in the lattice, and zero elsewhere. That is, for any phrase $w_1 w_2 \cdots w_n$ with $n \geq 2$, the Möbius inverse of the embedding function T at that phrase (the local *derivative* of T) is given by

$$\begin{aligned} q(w_1 \cdots w_n) = T(w_1 \cdots w_n) - T(w_1 \cdots w_{n-1}) \\ - T(w_2 \cdots w_n) + T(w_2 \cdots w_{n-1}), \end{aligned} \tag{2}$$

with the final term interpreted as 0 in the 2-word case. For a 3-word phrase $w_1 w_2 w_3$ this reduces to

$$\begin{aligned} q(w_1 w_2 w_3) = T(w_1 w_2 w_3) - T(w_1 w_2) \\ - T(w_2 w_3) + T(w_2). \end{aligned} \tag{3}$$

This is the semantic synergy in the 3-word phrase: it measures how much the phrase embedding $T(w_1 w_2 w_3)$ deviates from what is predicted by the bigram embeddings $T(w_1 w_2)$ and $T(w_2 w_3)$, after correcting for the fact that they both include w_2. Note that when the embedding function T is fully compositional, i.e. when $T(w_1 \ldots w_n) = \sum_{i=1}^{n} T(w_i)$, the semantic synergy q is the zero vector for all phrases of length $n \geq 2$. In that sense, q measures the *non-compositional* meaning of the phrase: the part of the phrase embedding that cannot be reduced to the sum of its parts.

Interpretation. A large q signals that much of the phrase's meaning exists only at the top level, and is not compositionally encoded in the embeddings of sub-spans. Projecting q onto the embedding of a candidate meaning phrase $T(m)$ via the inner product $\langle q, T(m) \rangle$ measures how much of a phrase's meaning is non-compositional, or synergistic. One situation in which one might expect a large synergy is in idiomatic phrases, where the meaning of the whole cannot be derived from the meanings of the parts. For example, in the idiom *kick the bucket*, the meaning of the whole phrase (to die) is not related to the meanings of the individual words (`kick`, `the`, `bucket`). Therefore, we would expect q(*kick the bucket*) to have a large inner product with T(*to die*), indicating that the idiomatic meaning is strongly non-compositional. By contrast, in a literal phrase like *kick the ball*, the meaning of the whole can be derived from the meanings of the parts, so we would expect q(*kick the ball*) to have a small inner product with T(*to kick a ball*), indicating that the literal meaning is mostly compositional. This intuition motivates our first experiment, where we test whether the Möbius residual can reliably separate idiomatic from literal phrases.

3. Results

3.1. Data

The main experiments use a manually curated benchmark of 107 matched short English idiom/literal pairs. Each row contains an idiom phrase, a matched literal control phrase, a figurative meaning gloss for the idiom, and a literal paraphrase for the control. Candidate idioms and meanings were

scraped from the English-language idiom list on Wikipedia (Wikipedia, 2026), filtered to short phrases, augmented with literal controls and glosses by a large language model, and then manually checked and edited by the author. During curation, we tried to keep lexical overlap between each phrase and its meaning gloss low, so that $\langle q, T(\text{meaning})\rangle$ is not explained by shared surface words. Both the original 107-pair list and the auxiliary extended list described below are available in the GitHub data repository (Abel Jansma, 2026).

The main text below uses only these 107 manually curated pairs. This is deliberately conservative: the benchmark is small, but the examples are checked by hand. Because $n = 107$ is limited, Appendix B repeats the same analyses on an auxiliary list of 836 3-word idiom/literal pairs generated by an LLM (ChatGPT 5.5 Pro) from the original 107 examples. That list is not manually curated, and is therefore treated as a robustness appendix rather than as the primary evidence. Across the appendix replication, the extended list confirms the main findings; absolute effect sizes sometimes decrease, but the larger sample increases the statistical significance of the reported effects. For random-triplet controls, we sample 107 random word triplets from the pooled idiom+literal vocabulary. The main embedding models are Qwen3-Embedding-0.6B (Qwen Team, 2025) and all-mpnet-base-v2 (Reimers & Gurevych, 2019).

3.2. Experiment 1: Idiomaticity Detection

Figure 2 shows the distribution of $\langle q(\text{phrase}), T(\text{meaning})\rangle$ for idioms, literal controls, and random-triplet controls on the 107-pair benchmark. Table 1 reports means and effect sizes.

The effect is large and highly significant in both main models. Random phrases sit near zero, confirming that the signal is not a generic property of arbitrary word triplets. The same construction also transfers beyond the 3-word special case: on 27 matched 4-word idiom/literal pairs, the idiom-vs.-literal contrast remains positive in both main models (Qwen3-Embedding-0.6B $d = 1.56$, all-mpnet-base-v2 $d = 0.92$). Pooling all available 4+ word phrases ($n = 47$) still yields positive contrasts in both main models, but the 5–7 word bins are too small and noisy for strong length-specific claims.

To verify that the Möbius residual specifically–rather than any deviation from the phrase embedding–drives the separation, we compute five alternative residuals for each phrase and repeat the idiom/literal contrast. For a 3-word phrase the alternatives are: (1) $r_{\text{uni-sum}} = T(w_1w_2w_3) - T(w_1) - T(w_2) - T(w_3)$; (2) $r_{\text{uni-mean}} = T(w_1w_2w_3) - \frac{1}{3}(T(w_1) + T(w_2) + T(w_3))$; (3) $r_{\text{bigram}} = T(w_1w_2w_3) - T(w_1w_2) - T(w_2w_3)$ (Möbius without the overlap correction); (4) r_{random}: a random unit vector of matched norm, redrawn per phrase with a fixed seed; (5) $r_{\text{phrase}} = T(w_1w_2w_3)$: the raw phrase embedding.

Several patterns emerge (see Table 2). The random direction shows no useful correct-direction separation. The raw phrase embedding and unigram-mean residual are reversed: literal phrases score higher than idioms. The unigram-sum residual is unstable across models. Bigram-only separation is positive, especially in all-mpnet-base-v2, but both idiom and literal means are negative; without the span-overlap correction, the score is less interpretable as a zero-centred idiomaticity diagnostic. The exact Möbius residual is large in both models, has the cleanest sign interpretation, and achieves ROC-AUC 0.90–0.92 with zero-threshold accuracies 0.80–0.83.

As an additional domain check, Appendix D applies the same 2-word residual to proper names. The Möbius residual reproduces the intuition that common-name combinations which refer to famous individuals (e.g. *Michael Jackson*) are less compositionally reducible than names whose individual parts already strongly identify the referent (e.g. *Albert Einstein*).

Appendix F reports an additional 2×2 meaning-embedding control crossing phrase type with meaning type. On the 107-pair benchmark, this interaction is significant for all-mpnet-base-v2 ($p = 4.9 \times 10^{-9}$), while Qwen3-Embedding-0.6B is directionally consistent but underpowered ($p = 0.11$). However, the extended-list replication in Appendix B makes the interaction significant in both models.

Nearest-Token Interpretability As a qualitative interpretability check, we retrieved nearest single-token vocabulary strings to each embedding model's vectors, embedding each candidate token string with the same sentence encoder. Candidates were restricted to clean, common English single tokens after dictionary and frequency filtering (2,870 candidates for Qwen3-Embedding-0.6B and 4,952 for all-mpnet-base-v2 in the 107-pair run). For each idiom we compared the Möbius residual q against compositional controls: the mean of the three unigram embeddings, the full phrase embedding, and the bigram-only residual.

The full-phrase embedding can also retrieve idiomatic meaning, as expected, but the synergy vector seems to extract the idiomatic meaning more robustly (see, for example, *on cloud nine* and *under the weather* in Table 3). No clear nearest-neighbour signal was found for the averaged steering vectors in either of the models.

Finally, some idioms have similar figurative meanings, like *kick the bucket* and *bite the dust*. One would expect the synergy vectors of such pairs to align. Across a range of similar idioms, Appendix B reveals that their synergy vectors indeed align more strongly than those of unrelated idioms.

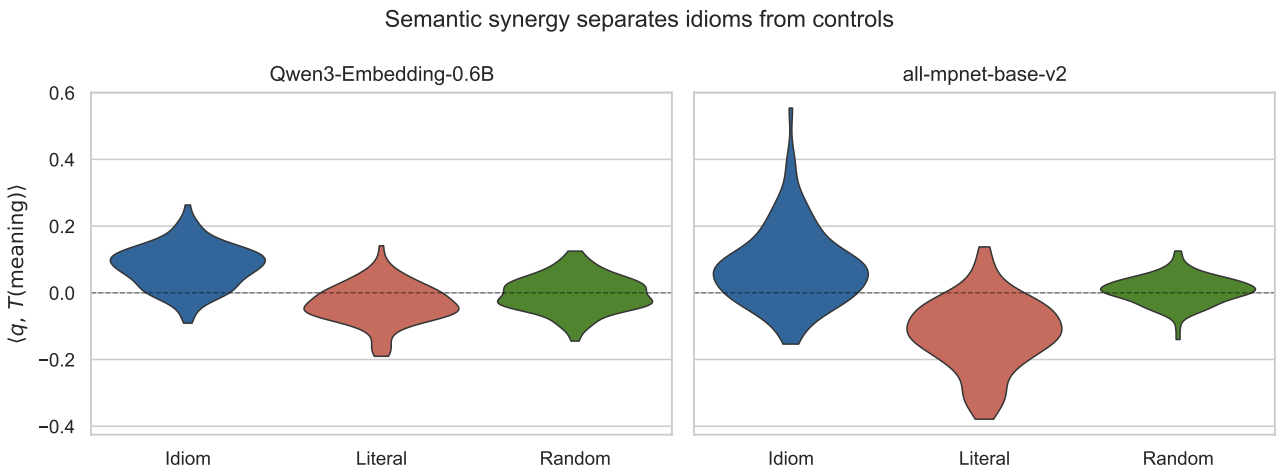

Figure 2. Distribution of $\langle q, T(\text{meaning})\rangle$ across the two main embedding models for 107 manually curated idiom phrases (blue), matched literal controls (red), and random word triplets (green).

Table 1. Idiom vs. literal contrast for $\langle q, T(\text{meaning})\rangle$ on the 107-pair manually curated benchmark. All p-values are Welch's t-test.

Model	Idiom mean	Literal mean	Cohen's d	p-value
Qwen3-Embedding-0.6B	0.078	-0.033	1.81	1.6×10^{-29}
all-mpnet-base-v2	0.078	-0.122	1.80	2.5×10^{-29}

Table 2. Cohen's d for idiom-vs.-literal separation under alternative residuals on the 107-pair benchmark. Sign + means idiom > literal. Full means and p-values are in Appendix C.

Residual	Qwen3-Embedding	all-mpnet-base-v2
Möbius q	$+1.81$	$+1.80$
Bigram-only	$+0.58$	$+1.58$
Unigram sum	-0.75	$+1.00$
Unigram mean	-1.52	-0.74
Random direction	-0.28	$+0.10$
Phrase only	-1.29	-1.44

3.3. Experiment 2: Layer-wise Emergence

The previous analysis uses final sentence embeddings. To ask where the non-compositional phrase component appears inside the encoder, we repeat the span-Möbius calculation at every hidden layer. For the two sentence-embedding models, we apply the same SentenceTransformer pooling operation used by the final embedding; for a plain Qwen3-0.6B language-model control, which is not trained as an embedding model, we use last-token pooling to match the Qwen embedding model's pooling convention. All layer vectors are L2-normalised before computing

$$q_\ell(w_1 \cdots w_n) = T_\ell(w_1 \cdots w_n) - T_\ell(w_1 \cdots w_{n-1}) - T_\ell(w_2 \cdots w_n) + T_\ell(w_2 \cdots w_{n-1}). \tag{4}$$

We then score each idiom by the phrase-alignment inner product $s_\ell = \langle q_\ell, T_\ell(\text{phrase})\rangle$. This differs from the main idiom-detection score: it does not use an external meaning gloss, but asks how much of the same-layer phrase representation lies in the non-additive residual direction. Layer 0 denotes the input embedding layer before transformer blocks.

Figure 3 shows the results. The strongest localization is in all-mpnet-base-v2: the idiom mean separates from literal controls through the middle and late layers, and the final layer gives idiom 0.055, literal -0.068, $d = 0.95$, $p = 6.0 \times 10^{-11}$. Qwen3-Embedding-0.6B shows a later transition and peaks at the final layer, with idiom 0.020, literal -0.022, $d = 0.67$, $p = 1.7 \times 10^{-6}$. The plain Qwen3-0.6B language model behaves differently. It contains a weaker transient version of the same contrast from early (layer 2/28: $d = 0.46$, $p = 1.0 \times 10^{-3}$) to middle layers, while the final layer is null or reversed (idiom -0.413, literal -0.401, $d = -0.08$, $p = 0.55$). Note that it is well-documented that intermediate layers in transformers can contain richer representations than final output layers (Skean et al., 2025). The phrase-aligned synergy component is thus not simply inherited from static token embeddings; it can appear transiently in a related decoder transformer, but embedding-tuned models preserve and sharpen the signal in the final representation.

Table 3. Nearest content-token neighbours in all-mpnet-base-v2, where the lexical interpretability effect is clearest. The Möbius residual often retrieves figurative or pragmatic neighbours, while the token-mean control stays close to surface/component words.

Phrase	q neighbours	Token-mean neighbours	Full-phrase neighbours
break the ice	dating, friendly, meet, mutual	break, ice, breaking, freeze	conversation, approach, communicate, dating
hit the sack	sleep, tired, depression, sleeping	hit, beat, impact, attack	sleep, workout, slept, rest
under the weather	exhausted, homeless, retired, wounded	weather, underneath, beneath, rain	weather, cold, rain, outside
on cloud nine	happy, happier, happily, happiness	cloud, nine, ten, seven	cloud, flying, nine, heaven
break the news	told, ashamed, adopted, acknowledge	news, break, breaking, newspaper	announce, reveal, announcement, news

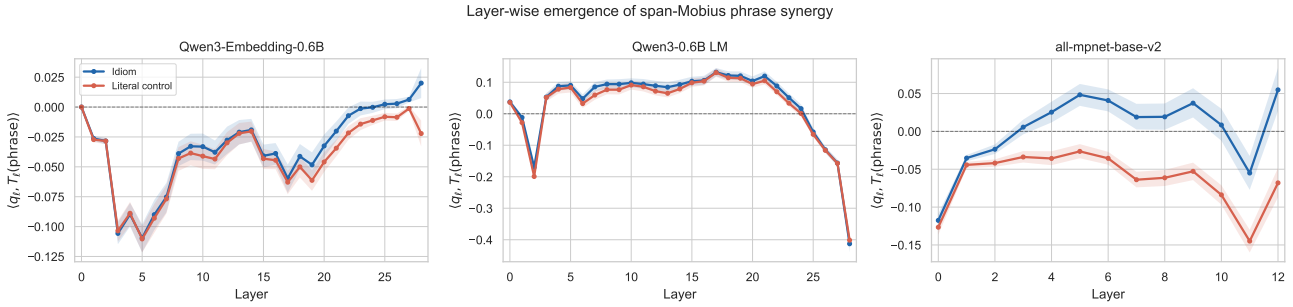

Figure 3. Layer-wise phrase-alignment score $\langle q_\ell, T_\ell(\text{phrase})\rangle$ for idioms and matched literal controls on the 107-pair benchmark. Bands are bootstrap 95% confidence intervals over phrase pairs. The signal emerges mainly in middle-to-late layers rather than at the input embedding layer.

Training-time emergence in Pythia-1B. The Pythia suite exposes intermediate checkpoints during language-model training (Biderman et al., 2023), which can reveal whether the same phrase-aligned structure appears early or late in training. We repeat the decoder-LM version of the layer-wise analysis on EleutherAI/pythia-1b at five checkpoints: step 0, 1k, 8k, 32k, and the final 143k checkpoint. As above, each hidden state is pooled at the final non-padding token, L2-normalised layer by layer, and scored by $\langle q_\ell, T_\ell(\text{phrase})\rangle$ on the 107-pair benchmark.

Figure 4 shows clear evidence for late emergence. The final-layer difference is near zero at initialization ($\Delta = -0.0056$, $d = -0.15$, $p = 0.28$) and after 1k steps ($\Delta = -0.0019$, $d = -0.03$, $p = 0.85$), becomes positive by 8k steps ($\Delta = 0.022$, $d = 0.30$, $p = 0.032$), rises at 32k steps ($\Delta = 0.030$, $d = 0.33$, $p = 0.016$), and is largest at the final checkpoint ($\Delta = 0.066$, $d = 0.53$, $p = 1.6 \times 10^{-4}$). Thus, in decoder-only language models, phrase-aligned non-compositional structure is not simply present at random initialization. It becomes a middle- to final-layer feature late in training. Appendix B shows that the same checkpoint trajectory becomes more statistically stable on the 836-pair list, and Appendix G compares Pythia-160M and Pythia-410M.

3.4. Experiment 3: Steering with Synergy Directions

If synergy residuals encode both specific idiomatic meaning and a general notion of 'idiomaticity', they should be usable as *steering vectors* in the embedding space. We construct four canonical directions for comparison:

$$
\begin{aligned}
c_{\text{möb}} &= \operatorname{norm}\big(\tfrac{1}{N}\textstyle\sum_i q_i\big)\,,\\
c_{\text{norm}} &= \operatorname{norm}\Big(\tfrac{1}{N}\textstyle\sum_i \tfrac{q_i}{\|q_i\|}\Big)\,,\\
c_{\text{concept}} &= \operatorname{norm}\big(\bar{T}_{\text{idiom}} - \bar{T}_{\text{literal}}\big)\,,\\
c_{\text{random}} &= \text{fixed random unit vector (seed 0).}
\end{aligned}
$$

$c_{\text{möb}}$ and c_{norm} are derived from the Möbius residuals (unsupervised); c_{concept} requires knowing the idiom/literal grouping (supervised); c_{random} is a null baseline.

For a phrase embedding z, the idiomaticity index is

$$
\begin{aligned}
\Delta = \;& \cos(z, T(\text{idiomatic target}))\\
& - \cos(z, T(\text{literal paraphrase})),
\end{aligned}
$$

where cos denotes cosine similarity. Steering adds or subtracts a scaled canonical vector: $z' = \operatorname{norm}(z \pm \alpha c)$, with $\alpha = 5.0$ for all reported evaluations (see Appendix E for a sweep across α values).

Over 200 random 80/20 splits of the 107-pair benchmark, each steering direction is constructed from the 86 training idioms and evaluated only on the 21 held-out idioms (Figure 5). The train-only Möbius direction $c_{\text{möb}}$ gives the expected signed movement on unseen phrases: in Qwen3-Embedding-0.6B, it moves the held-out idiomaticity index from $\Delta = -0.018$ by default to $+0.117$ under Idiom+ steering and -0.133 under Idiom− steering ($p = 1.7 \times 10^{-190}$ for Idiom+ vs. Idiom− across splits). This is close to the supervised concept vector, which reaches

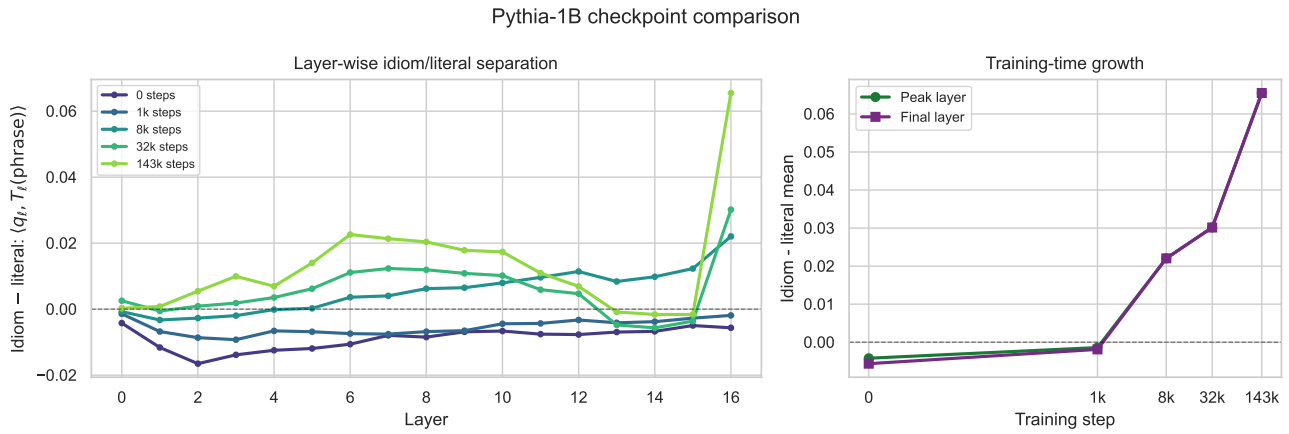

Figure 4. Training-time emergence of the layer-wise idiom/literal separation in Pythia-1B on the 107-pair benchmark. Left: difference between idiom and literal means for $\langle q_\ell, T_\ell(\text{phrase})\rangle$ at five checkpoints. Right: peak-layer and final-layer differences across training.

$+0.122/-0.141$. In all-mpnet-base-v2, $c_{\text{möb}}$ moves Δ from $+0.045$ to $+0.122/-0.102$ ($p = 2.7 \times 10^{-173}$), while c_{concept} reaches $+0.160/-0.148$. The averaged synergy direction therefore generalises to unseen idioms rather than merely fitting the phrases from which it was constructed.

4. Discussion and Conclusions

Taken together, the experiments support a simple interpretation: when a phrase has meaning that is not recoverable by additively composing its parts, that extra contribution—the semantic synergy—corresponds to a direction in embedding space that is captured by the Möbius residual on the span mereology. On the manually curated 107-pair benchmark, these residuals separate idioms from matched literal controls with large effects, beat random and alternative-residual baselines, and give nearest-neighbour structure that often reflects the figurative meaning rather than the surface words. The same residuals can also be averaged into steering directions that move held-out phrase embeddings toward idiomatic interpretations, indicating that the signal is not only diagnostic but geometrically actionable. The appendix replication on 836 LLM-generated pairs supports the same conclusion at larger scale: effects are sometimes smaller in absolute size, as expected from a noisier non-manually-curated list, but the larger sample makes the statistical evidence stronger.

The layer-wise and training-time analyses make this picture more specific. For Qwen3-0.6B and Pythia-1B, the alignment between the synergy vector and the phrase embedding emerges in the middle-to-late layers. In Pythia-1B, the effect is strongest late in training and is much clearer at 1B parameters than in the 410M and 160M parameter models, suggesting that scale and training both increase the model's ability to represent non-compositional meaning as structured geometry. This is the main representation-level claim of the paper, and the point of contact with mechanistic interpretability: the residual quantifies what the encoder contributes *beyond* additive composition, and that contribution behaves like a learnable representational feature rather than like idiosyncratic embedding noise. This presents an example where a naive version of the linear representation hypothesis fails while preserving interpretable decompositions of latent representations.

Limitations The claim remains representation-level rather than circuit-level. The main experiments use sentence-level embedding models or pooled hidden states, and therefore do not yet identify the attention heads, MLP features, or residual-stream mechanisms that construct the synergy vector. The dataset is also English-only and small: the primary n=107 short idiom/literal pairs are manually validated, but still cover only a narrow slice of English idioms. The larger n=836 3-word list in Appendix B was generated with LLM assistance and has not been manually curated, so it should be read as a robustness check rather than as a replacement benchmark. The 4-word robustness set has only n=27 and the 5–7 word subsets are too small for stable conclusions.

Finally, the 2×2 meaning-embedding control is underpowered for Qwen3-Embedding-0.6B at $n = 107$ but becomes significant in both models at $n = 836$; broader multilingual and independently sampled datasets are still needed before treating the meaning-control result as architecture-general.

The most important next step is therefore to apply the same algebra directly to token-level residual streams in larger decoder-only LLMs. Computing Möbius residuals directly in residual streams would connect the span calculus used here to causal interventions and circuits-style mechanistic interpretability. This could also reveal whether synergy directions are useful for downstream tasks such as idiom disambiguation or generation steering. The present results

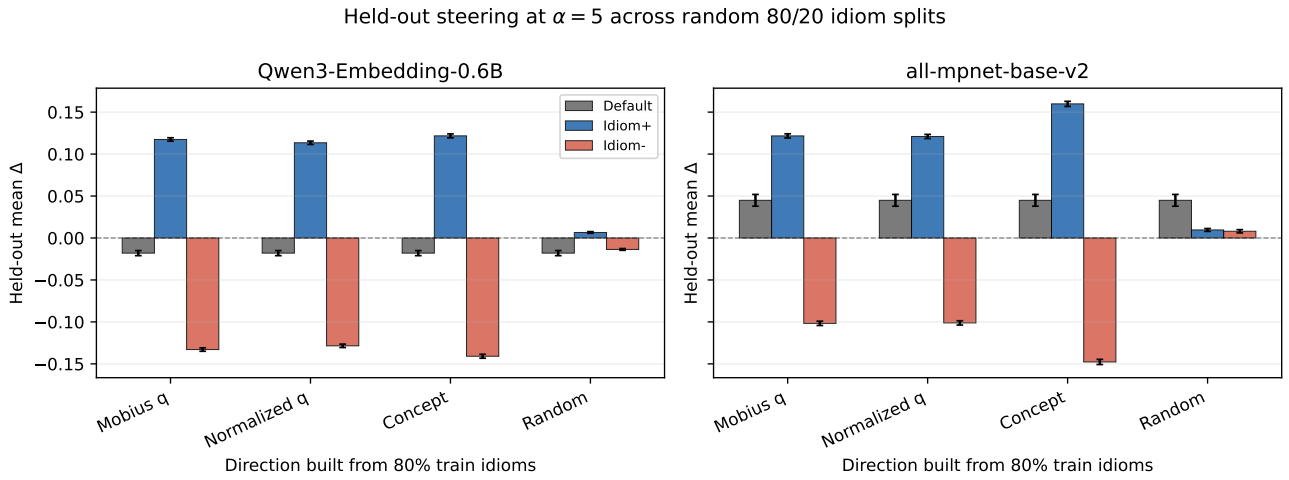

Figure 5. Held-out steering validation across 200 random 80/20 idiom splits on the 107-pair benchmark. Directions are constructed from the 86 training idioms in each split and evaluated on the 21 held-out idioms. Bars show split-level mean idiomaticity index Δ at $\alpha = 5$; error bars are 95% confidence intervals over splits.

show that Möbius inversion supplies a compact calculus for non-compositional meaning in embedding space; the remaining question is which internal mechanisms write these residuals into the model.

Acknowledgements

The author thanks Patrick Forré, Martha Lewis, Melanie Mitchell, and Fernando Rosas for helpful discussions and comments. This work was supported by the Dutch Institute for Emergent Phenomena (DIEP) cluster via the Foundations and Applications of Emergence (FAEME) programme, as well as through the PIBBSS $\times$ Iliad Research Residency and Principles of Intelligence.

References

Abel Jansma. Semantic synergy data repository. https://github.com/AJnsm/semantic_synergy_data/tree/main, 2026.

Baroni, M., Bernardi, R., and Zamparelli, R. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9:241–346, 2014. URL https://aclanthology.org/2014.lilt-9.5/.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Van Der Wal, O. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Forré, P. and Jansma, A. Möbius transforms and shapley values for vector-valued functions on weighted directed acyclic multigraphs. *arXiv preprint arXiv:2510.05786*, 2025.

Jansma, A. Mereological approach to higher-order structure in complex systems: From macro to micro with möbius. *Physical Review Research*, 7(2):023016, 2025a.

Jansma, A. Decomposing interventional causality into synergistic, redundant, and unique components. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=yPnEvPq3kV.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, 2013.

Pagin, P. and Westerståhl, D. Compositionality I: Definitions and variants. *Philosophy Compass*, 5(3):250–264, 2010. doi: 10.1111/j.1747-9991.2009.00228.x.

Qwen Team. Qwen3 embedding, 2025. https://huggingface.co/Qwen/Qwen3-Embedding-0.6B.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019.

Rota, G.-C. On the foundations of combinatorial theory I: Theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

Wikipedia. English-language idioms, 2026. https://en.wikipedia.org/wiki/English-language_idioms.

Wold, S., Simon, É., Velldal, E., and Øvrelid, L. Measuring idiomaticity in text embedding models with epsilon-compositionality. In Demberg, V., Inui, K., and Marquez, L. (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2239–2252, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.99. URL https://aclanthology.org/2026.eacl-long.99/.

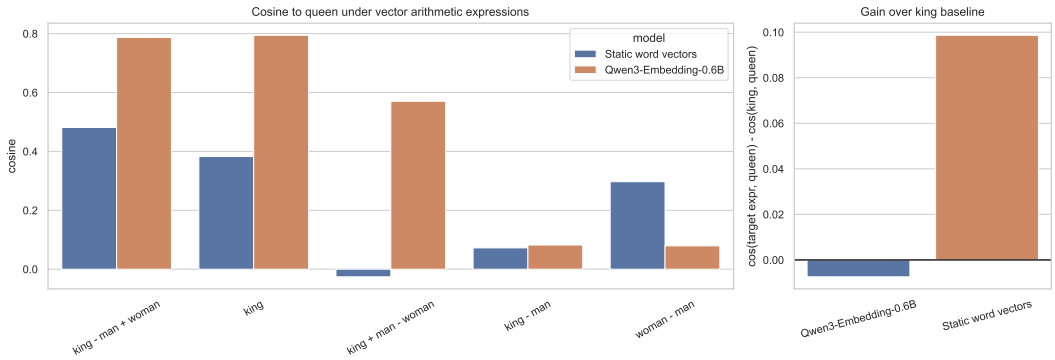

Figure 6. Token-level vector arithmetic (`king−man+woman`) in spaCy `en_core_web_md` v3.8.0 static word vectors (left bar groups) and Qwen3-Embedding-0.6B (right bar groups). Right panel shows cosine gain over the `king` baseline. Static vectors support this kind of arithmetic; contextual encoders do not, motivating a decomposition that respects contextual span structure.

Table 4. Extended-list idiom vs. literal contrast for $\langle q, T(\text{meaning})\rangle$. All p-values are Welch's t-test.

Model	Idiom mean	Literal mean	Cohen's d	p-value
Qwen3-Embedding-0.6B	0.043	−0.026	1.13	3.4×10^{-102}
all-mpnet-base-v2	0.039	−0.084	1.04	5.6×10^{-89}

A. Word-Vector Arithmetic Motivation

Figure 6 gives the motivating negative example behind the span-Möbius construction. Classic static word-vector spaces often support simple arithmetic reasoning: subtracting a gender-associated direction from `king` and adding it back in another direction can increase similarity to `queen`. The static baseline in the figure uses the 300-dimensional vector table from spaCy `en_core_web_md` v3.8.0 (`en_vectors`). In contextual phrase encoders, however, token and phrase embeddings are produced in a context-sensitive way, so the same linear recipe is no longer a reliable model of how meanings combine. This motivates replacing latent space arithmetic with a *calculus* over the partial order of spans: rather than asking whether isolated word vectors add, we ask what contribution the full phrase has after all sub-span contributions have been removed.

B. Extended LLM-Generated 836-Pair Replication

The primary results in the main text use the manually curated 107-pair benchmark. To check whether the findings survive a broader but less controlled idiom inventory, we also generated an auxiliary list of 836 3-word idiom/literal pairs using ChatGPT 5.5 Pro, prompted from the original 107 examples. The generated list includes idiom phrases, matched literal controls, figurative meaning glosses, and literal paraphrases, but it has not been manually curated. Both the original and extended CSV files are available in the GitHub data repository (Abel Jansma, 2026). The extended list therefore tests scale and robustness, not benchmark quality.

Across the extended-list analyses, the qualitative findings match the main text. In several cases the absolute effect size is smaller than on the hand-checked set, which is expected for a noisier LLM-generated inventory, but the much larger sample lowers the p-values for the corresponding positive findings.

The alternative-residual comparison gives the same basic pattern as in the main text (Figure 8). The exact Möbius residual remains the cleanest sign-centred diagnostic. Bigram-only is competitive in all-mpnet-base-v2, but both idiom and literal means are negative, so it lacks the zero-threshold interpretation of the Möbius residual. Raw phrase embeddings and unigram-mean residuals again point in the wrong direction.

Layer-wise phrase alignment also replicates on the extended list (Figure 9). At the final layer, Qwen3-Embedding-0.6B gives idiom 0.015, literal -0.013, $d = 0.47$, $p = 1.7 \times 10^{-21}$; all-mpnet-base-v2 gives idiom 0.045, literal -0.042, $d = 0.63$, $p = 1.2 \times 10^{-35}$. These effects are smaller than the 107-pair effects ($d = 0.67$ and 0.95), but substantially more significant.

The Pythia-1B checkpoint analysis shows the same late-emergence trajectory (Figure 10). The final-layer difference is near zero at initialization ($\Delta = -0.0028$, $d = -0.067$, $p = 0.17$), becomes weakly positive at 1k steps ($\Delta = 0.0086$, $d = 0.097$,

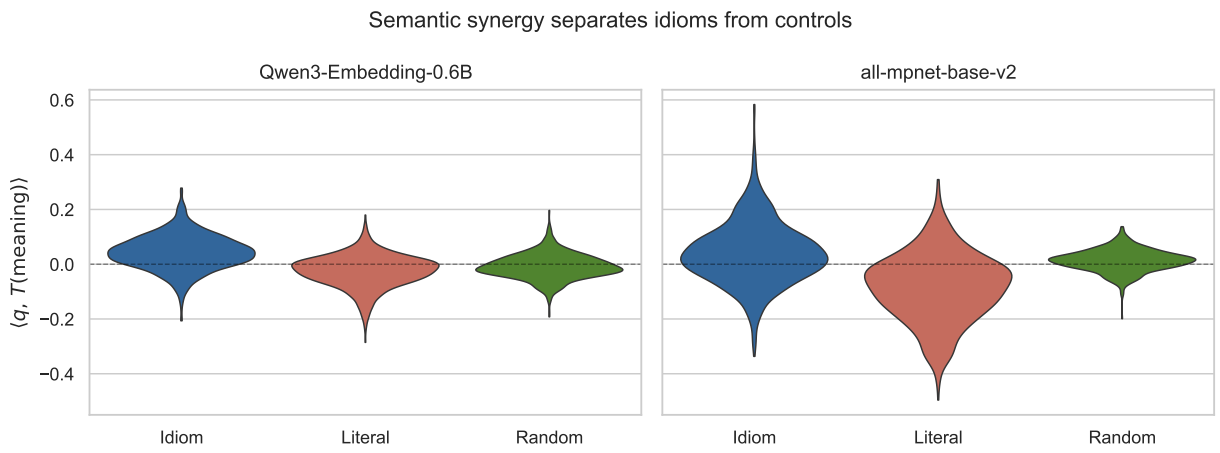

Figure 7. Extended-list idiomaticity detection for 836 LLM-generated idiom/literal pairs. The score is $\langle q, T(\text{meaning})\rangle$. Compared with the 107-pair benchmark, absolute effect sizes decrease but significance increases.

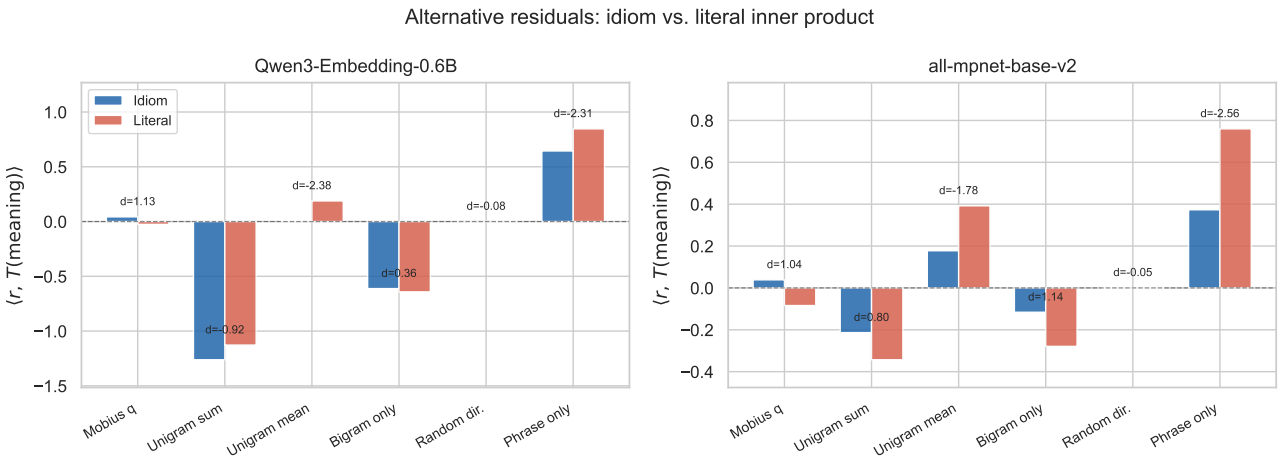

Figure 8. Extended-list comparison with alternative residuals for $\langle r, T(\text{meaning})\rangle$. The Möbius residual preserves the correct sign-centred idiom/literal contrast; several alternatives are significant but reversed or not zero-centred.

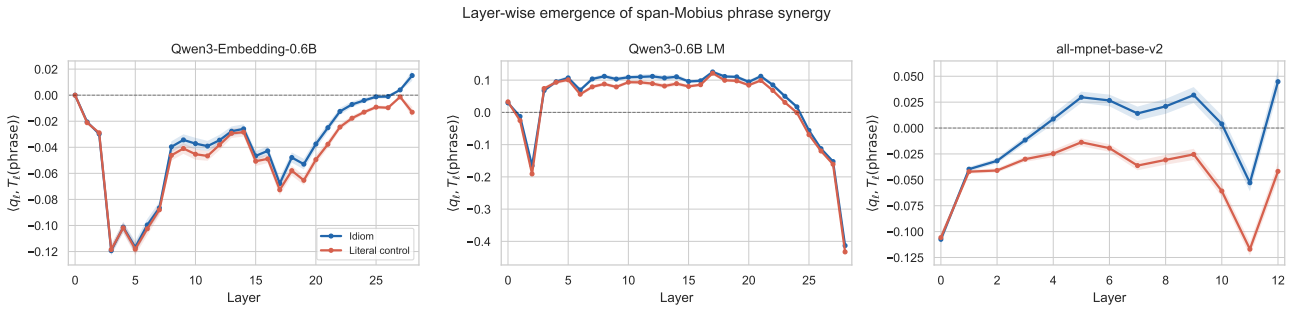

Figure 9. Extended-list layer-wise phrase-alignment score $\langle q_\ell, T_\ell(\text{phrase})\rangle$ for idioms and matched literal controls.

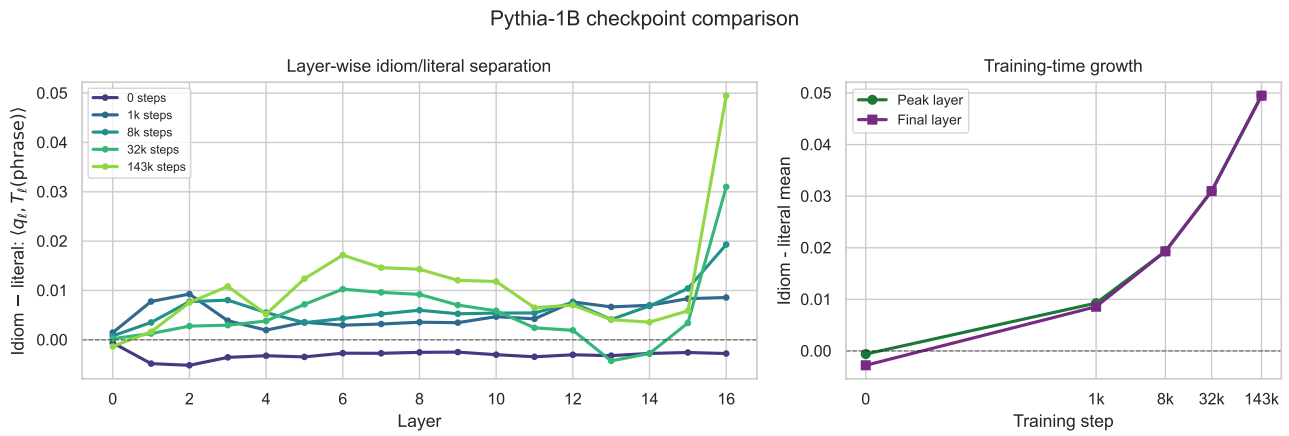

Figure 10. Extended-list Pythia-1B training-time emergence for $\langle q_\ell, T_\ell(\text{phrase})\rangle$.

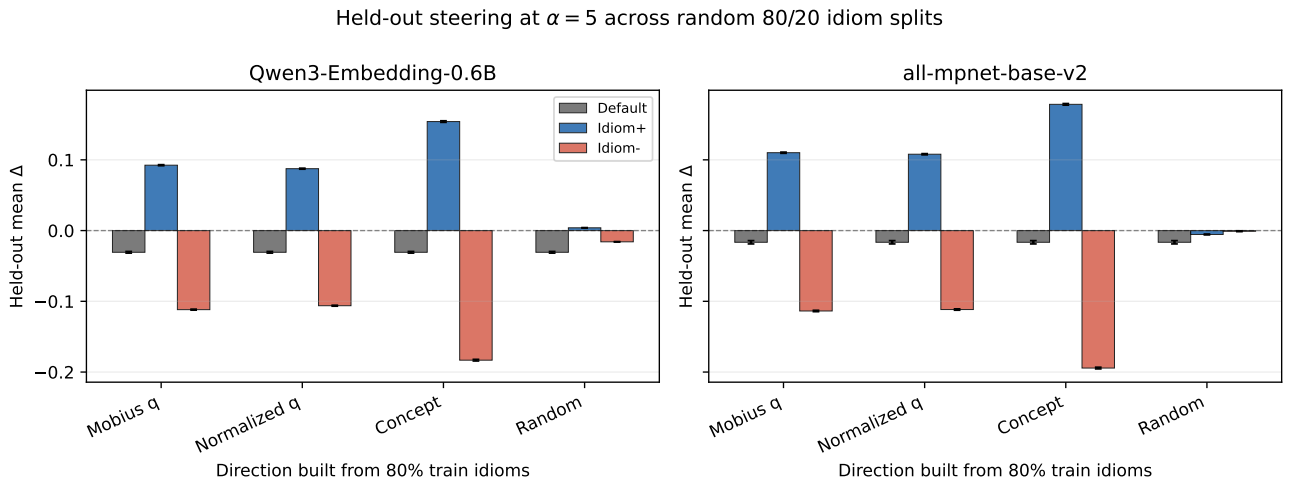

Figure 11. Extended-list held-out steering validation across 200 random 80/20 idiom splits.

$p = 0.048$), and is strongest at the final 143k checkpoint ($\Delta = 0.049$, $d = 0.31$, $p = 5.5 \times 10^{-10}$).

Held-out steering also replicates (Figure 11). Across 200 random 80/20 splits of the 836-pair list, the train-only Möbius direction moves Qwen3-Embedding-0.6B from default $\Delta = -0.031$ to $+0.092/-0.112$ under Idiom+ / Idiom− steering ($p = 5.4 \times 10^{-262}$). For all-mpnet-base-v2 the corresponding values are -0.017 to $+0.110/-0.114$ ($p = 1.6 \times 10^{-257}$). The supervised concept vector is stronger on the extended list, but the unsupervised Möbius direction still generalises to held-out idioms.

Finally, the same-meaning idiom alignment check becomes clearer when unrelated comparisons are drawn from the full 836-idiom background (Figure 12). For Qwen3-Embedding-0.6B, raw q vectors have higher same-cluster than unrelated-pair cosine (0.327 vs. 0.163; Welch $p = 0.002$), and the centered $q - \bar{q}$ contrast is similar (0.152 vs. -0.001; $p = 0.003$). For all-mpnet-base-v2, raw q separates same-meaning from unrelated pairs (0.304 vs. 0.142; $p = 0.033$), and centered q gives a comparable contrast (0.164 vs. 0.001; $p = 0.048$). Full phrase embeddings and token-mean controls also contain some synonymy signal, so this analysis should be treated as exploratory evidence that q preserves meaning-specific structure, not as a definitive synonymy benchmark.

The 2×2 meaning-embedding control also strengthens on the extended list (Figure 13). The Qwen3-Embedding-0.6B interaction becomes significant ($F = 4.26$, $p = 0.039$), and all-mpnet-base-v2 remains strongly significant ($F = 184.6$, $p = 5.8 \times 10^{-41}$). This is the clearest example where the original 107-pair result is directionally consistent but underpowered in one model, while the larger list makes the same effect significant in both.

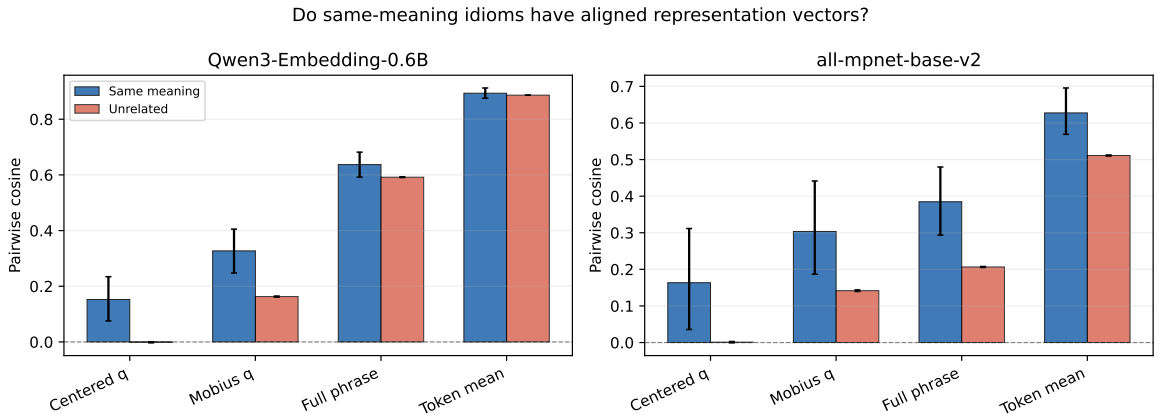

Figure 12. Extended-list same-meaning idiom alignment. Same-meaning pairs are 14 curated synonym pairs; unrelated comparisons are drawn from the 836-idiom LLM-generated background.

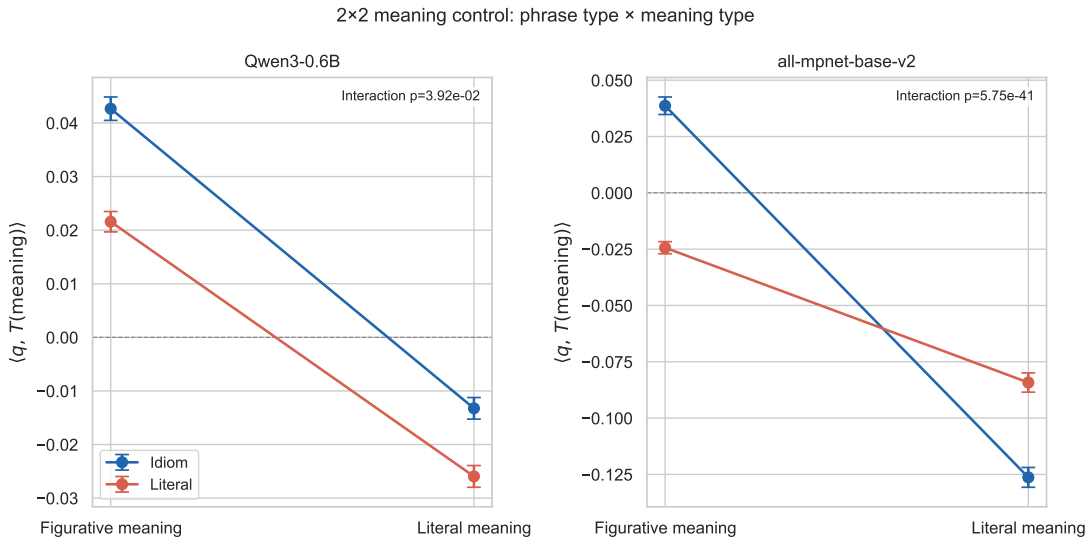

Figure 13. Extended-list 2×2 meaning-embedding control crossing phrase type with meaning type.

Table 5. Idiom vs. literal contrast on the 107-pair manually curated benchmark ($\langle r, T(\text{meaning})\rangle$, Welch's t-test) for the Möbius residual and five alternatives. d = Cohen's d; sign + means idiom > literal. Bold indicates the Möbius residual.

Model	Residual	Idiom mean	Literal mean	Cohen's d	p-value
Qwen3-Embedding	**Möbius** q	$\mathbf{+0.078}$	$\mathbf{-0.033}$	$\mathbf{+1.81}$	$\mathbf{1.6 \times 10^{-29}}$
	Bigram-only	-0.584	-0.633	$+0.58$	3.0×10^{-5}
	Unigram sum	-1.256	-1.141	-0.75	1.2×10^{-7}
	Unigram mean	$+0.001$	$+0.118$	-1.52	9.1×10^{-23}
	Random direction	-0.000	$+0.005$	-0.28	4.5×10^{-2}
	Phrase only	$+0.630$	$+0.745$	-1.29	7.1×10^{-18}
all-mpnet-base-v2	**Möbius** q	$\mathbf{+0.078}$	$\mathbf{-0.122}$	$\mathbf{+1.80}$	$\mathbf{2.5 \times 10^{-29}}$
	Bigram-only	-0.079	-0.280	$+1.58$	2.8×10^{-24}
	Unigram sum	-0.198	-0.364	$+1.00$	4.4×10^{-12}
	Unigram mean	$+0.182$	$+0.273$	-0.74	1.9×10^{-7}
	Random direction	-0.001	-0.005	$+0.10$	0.46
	Phrase only	$+0.373$	$+0.589$	-1.44	4.6×10^{-21}

Table 6. Name-level span synergy (Qwen3-Embedding-0.6B), mean ± SD over $n = 32$ names per group, each averaged across three carrier sentences. All pairwise Welch's p-values < 0.05; see text.

Group	n	$\|q\|_2$	$\langle q, T(\text{name})\rangle$
Non-compositional (Michael Jackson, Michael Jordan, ...)	32	1.211 ± 0.061	-0.493 ± 0.089
Unfamiliar (Gary Elsworth, Teresa Norwood, ...)	32	1.065 ± 0.044	-0.641 ± 0.070
Compositional (Albert Einstein, Marie Curie, ...)	32	1.039 ± 0.056	-0.700 ± 0.132

C. Alternative Residual Details

Table 5 expands the compressed residual comparison in the main text. For each phrase, we compare the residual vector r with the embedding of the supplied meaning gloss, and ask whether this score is larger for idioms than for matched literal controls. The alternatives test whether the result is just an artifact of phrase embeddings, unigram aggregation, bigram overlap, or arbitrary high-dimensional directions. The expected signature is a large positive idiom–literal effect: idiomatic residuals should align more strongly with figurative meanings than literal residuals align with literal paraphrases.

The Möbius residual is the strongest correct-direction effect in Qwen3-Embedding-0.6B and the cleanest sign-centred effect in both models. Several controls are statistically significant even at $n = 107$, but they either point in the wrong direction despite low p-values (e.g. phrase-only and unigram-mean scores), are null or unstable (random direction), or lack the centred sign interpretation of the Möbius residual (bigram-only).

D. Proper-Name Extension

Proper names offer a supporting domain check: *non-compositional* names like *Michael Jackson* or *James Brown* refer to famous entities whose meaning cannot be recovered from the component first- and last-name embeddings, since both parts are common names. *Compositional* names like *Albert Einstein* or *Ada Lovelace* have sufficiently distinctive components that the parts themselves carry substantial information about the referent. As an additional group we include *unfamiliar* names (e.g., *Laura Whitfield*) that have no intended strong public association.

We embed 32 names per group (96 total; list available in Abel Jansma 2026) each in three carrier sentences and compute the span synergy $q(\text{first last}) = T(\text{first last}) - T(\text{first}) - T(\text{last})$. Statistical significance is assessed by Welch's t-test and bootstrap 95% confidence intervals (10 000 resamples) on the difference of means.

The three groups are ordered as predicted: non-compositional names show the largest synergy norms and least-negative inner products; compositional names show the smallest (Figure 14). All three pairwise comparisons are statistically significant:

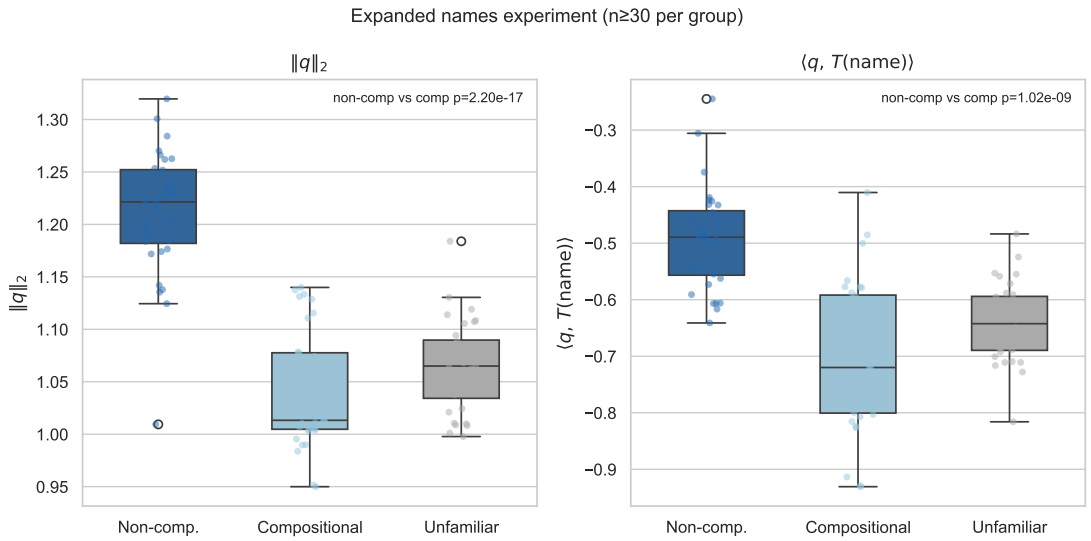

Figure 14. Span synergy norm (left) and inner product with the full-name embedding (right) for three name groups ($n = 32$ each). Each point is one name (averaged over three carrier sentences). All pairwise differences are significant (Welch's t-test and bootstrap CI).

Table 7. 2×2 meaning-embedding control on the 107-pair benchmark: cell means of $\langle q, T(\text{meaning})\rangle$. Interaction p-values from two-way ANOVA (Type II).

Model		Figurative meaning	Literal meaning	Interaction p
Qwen3-Embedding	Idiom	$+0.078$	-0.007	0.11
	Literal	$+0.034$	-0.033	
all-mpnet-base-v2	Idiom	$+0.078$	-0.134	4.9×10^{-9}
	Literal	-0.028	-0.122	

non-comp vs. compositional ($p_{\|q\|} = 2.2 \times 10^{-17}$, bootstrap CI $[0.144, 0.199]$; $p_{\langle q,T\rangle} = 1.0 \times 10^{-9}$, CI $[0.155, 0.262]$); non-comp vs. unfamiliar ($p_{\|q\|} = 8.7 \times 10^{-16}$, CI $[0.120, 0.171]$; $p_{\langle q,T\rangle} = 5.7 \times 10^{-10}$, CI $[0.111, 0.188]$); compositional vs. unfamiliar ($p_{\|q\|} = 0.048$, CI $[-0.049, -0.001]$; $p_{\langle q,T\rangle} = 0.030$, CI $[-0.108, -0.008]$).

E. Held-Out Steering Alpha Sweep

The main text reports held-out steering at $\alpha = 5.0$. To check that this choice is not a narrow tuning artefact, Figure 15 shows the full held-out alpha sweep on the 107-pair benchmark. For each of the 200 random 80/20 splits, directions are fit only on the 86 training idioms and evaluated only on the 21 held-out idioms. The figure plots split-level mean Δ and 95% confidence intervals over splits.

The effect is smooth across steering strengths. For both Qwen3-Embedding-0.6B and all-mpnet-base-v2, the Möbius direction and the normalized-q direction move held-out phrase embeddings in the expected directions: Idiom+ increases Δ, and Idiom− decreases it. The curves rise rapidly at small α and then saturate, so the reported $\alpha = 5$ value lies in a stable regime rather than at a sharp optimum. The supervised concept vector is generally strongest, while the random direction is much smaller and less consistent.

F. Meaning-Embedding Control

A potential confound is that the figurative meaning descriptions for idioms can differ from the literal meaning descriptions in length and surface overlap. To disentangle the contribution of q from the similarity structure of the meaning embedding, we construct a 2×2 design: **phrase type** (idiom / literal) $\times$ **meaning type** (figurative gloss / literal paraphrase). Each phrase is crossed with both meaning types from its matched pair, giving four cells.

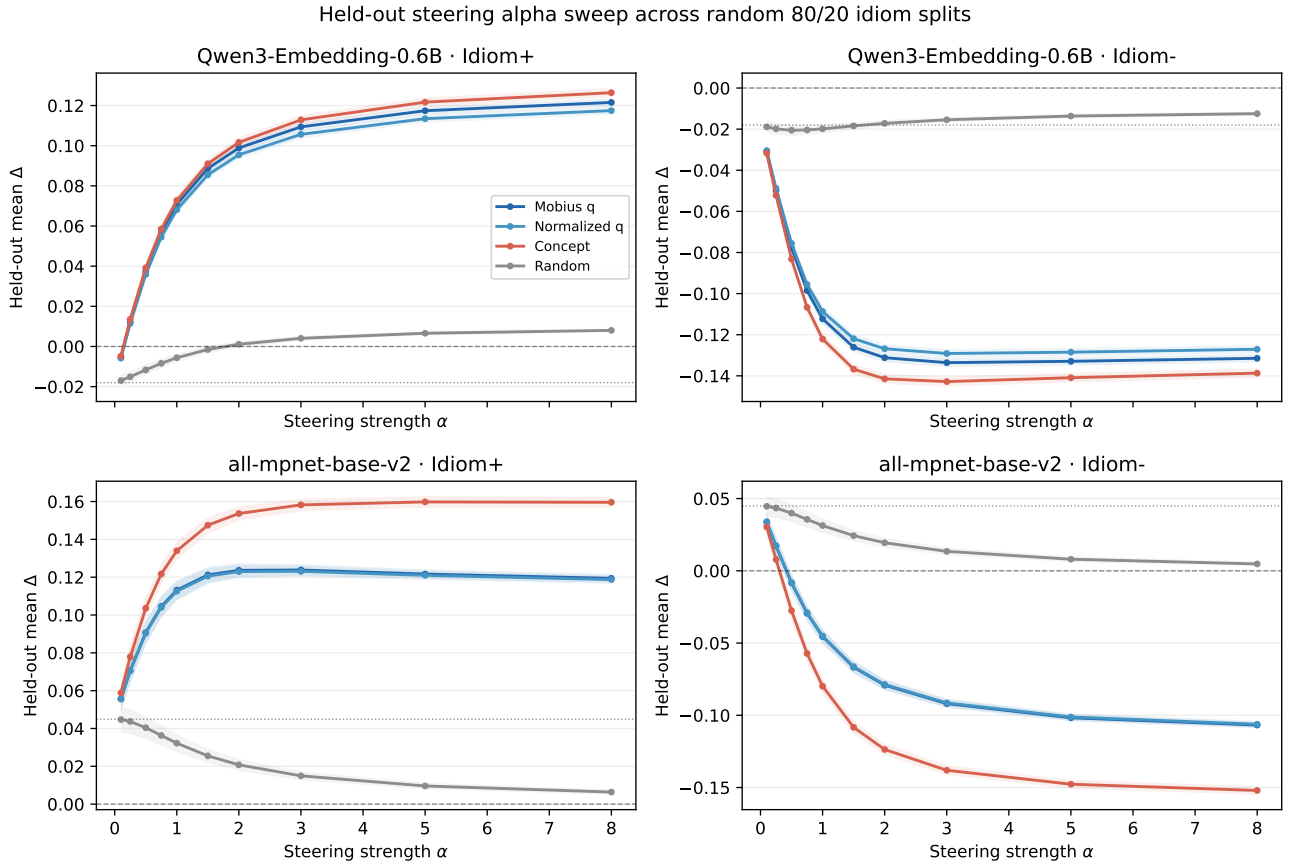

Figure 15. Held-out steering alpha sweep on the 107-pair benchmark. Each panel shows split-level held-out mean Δ across 200 random 80/20 splits. Solid curves are steering directions fit only on the train idioms in each split; shaded bands are 95% confidence intervals over splits. Dotted grey lines indicate the unsteered default mean for the corresponding model.

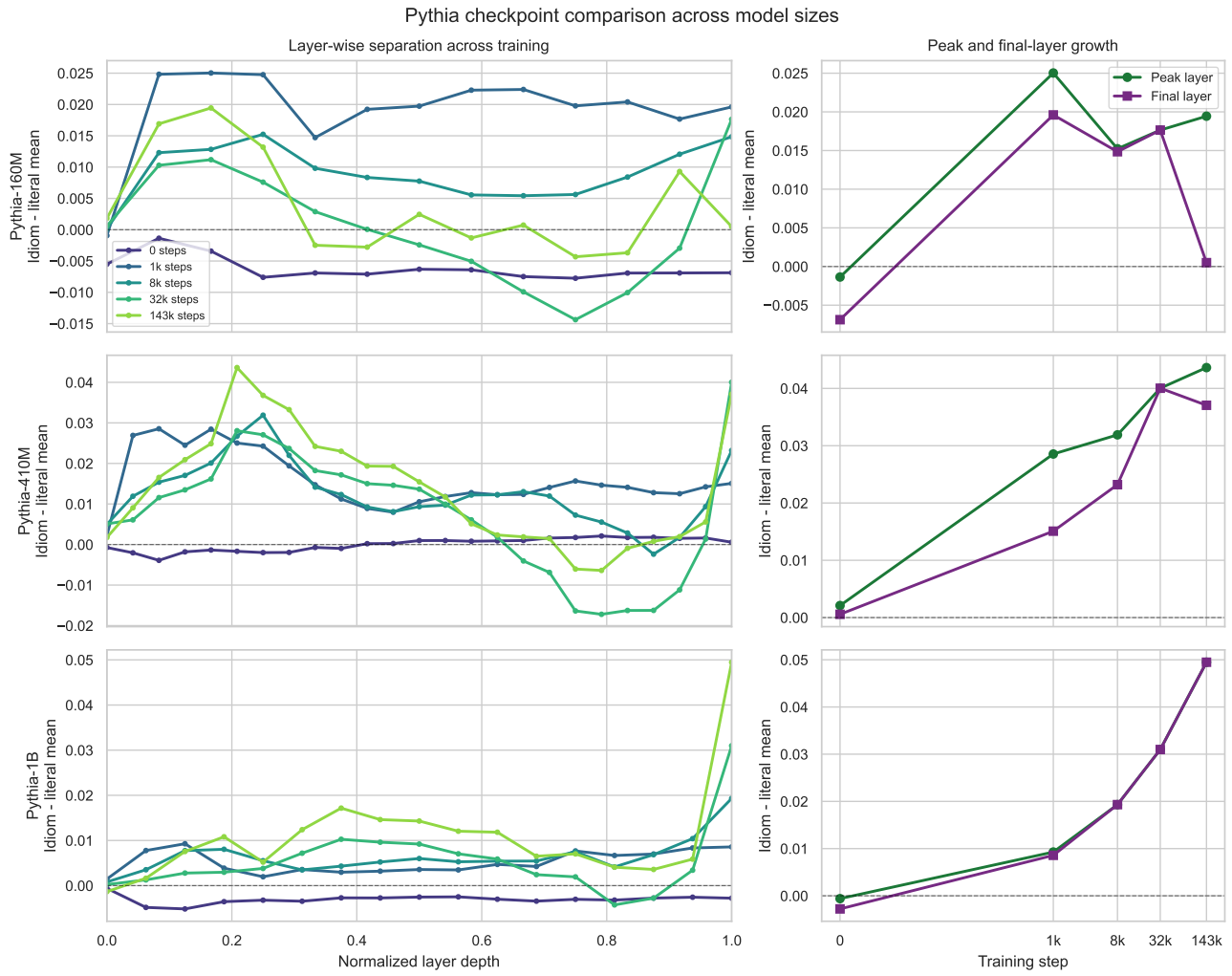

Figure 16. Pythia checkpoint comparison across 160M, 410M, and 1B models. The analysis uses the 836-pair LLM-expanded list. The 1B final-layer mean difference is strongest; the same pattern is visible in 410M but is not meaningfully preserved in the final layer of 160M.

The key contrast is whether idiom phrases show a larger figurative-vs.-literal meaning difference than literal phrases do (Table 7). For all-mpnet-base-v2 this interaction is large and highly significant ($F = 35.7$, $p = 4.9 \times 10^{-9}$): idiom q aligns strongly with figurative meaning and negatively with literal meaning, while literal q shows a smaller asymmetry. For Qwen3-Embedding-0.6B the interaction is directionally consistent but underpowered on the 107-pair benchmark ($F = 2.61$, $p = 0.11$). On the 836-pair extended list, the same control becomes significant in both models (Qwen3-Embedding-0.6B $p = 0.039$, all-mpnet-base-v2 $p = 5.8 \times 10^{-41}$; Appendix B). We interpret these results as support that the idiom/literal separation is not simply inherited from the meaning-gloss embeddings.

G. Extended-List Pythia Scale Comparison

To check whether the Pythia-1B training-time pattern is specific to that scale, we repeated the checkpoint analysis for EleutherAI/pythia-160m and EleutherAI/pythia-410m using the same five checkpoints, pooling rule, and 836 idiom/literal pairs.

The scale comparison suggests that model size increases the ability to pick up non-compositional phrase meaning. Pythia-160M does not meaningfully preserve the contrast in its final representation: its final-layer difference is tiny at the final checkpoint ($\Delta = 0.0005$, $d = 0.21$, $p = 1.5 \times 10^{-5}$), although a localized layer-2/12 peak is stronger ($\Delta = 0.019$, $d = 0.40$, $p = 4.5 \times 10^{-16}$). Pythia-410M shows the 1B pattern in weaker form: the final-layer contrast is near zero at

initialization, becomes significant during training, and remains positive at the final checkpoint (143k: $\Delta = 0.037$, $d = 0.33$, $p = 3.7 \times 10^{-11}$), with a slightly larger layer-5/24 peak ($\Delta = 0.044$, $d = 0.47$, $p = 4.3 \times 10^{-21}$). In Pythia-1B the corresponding final-checkpoint mean difference is larger ($\Delta = 0.049$, $d = 0.31$, $p = 5.5 \times 10^{-10}$). Thus the late final-layer mean effect is clearest in 1B, visible in 410M, and not convincingly present as a final-layer signal in 160M.