Rethinking the Idiomaticity Decomposability Hypothesis: Evidence from Distributional Learning

Maggie Mi[1] **Golzar Atefi**[2] **Atsuki Yamaguchi**[1] **Felix Alexander Gers**[2]
Aline Villavicencio[1,3,4] **Nafise Sadat Moosavi**[1]
[1]University of Sheffield [2]Berliner Hochschule für Technik (BHT)
[3]University of Exeter [4]Federal University of Rio Grande do Norte, Brazil
{zmi1, ayamaguchi1, a.villavicencio, n.s.moosavi}@sheffield.ac.uk
{golzar.atefi, FelixAlexander.Gers}@bht-berlin.de

Abstract

Idioms can be analysed in terms of their decomposability, the extent to which constituent meanings contribute to the figurative whole. Decomposability is thought to predict syntactic flexibility. Usage-based accounts instead attribute idiom behaviour to distributional experience, such as speaker familiarity and predictability. We examine these views using contextualised language models as controlled distributional learners. We propose a model-internal measure of decomposability and relate it to human ratings, syntactic flexibility, and predictability while tracking idiom learning during pretraining. Model-derived decomposability correlates weakly with human judgments and shows a small but consistent negative relationship with syntactic flexibility. Pretraining analyses show that stabilisation of idiom representations in models is not explained by frequency alone. Instead, surprisal, decomposability, and frequency all contribute, with decomposability showing the strongest training-dependent effect.

https://github.com/mi-m1/idiom_decomp

1 Introduction

Idiomatic expressions, such as "*spill the beans*" (meaning to reveal a secret), "*pop the question*" (propose marriage), or "*kick the bucket*" (to die), exhibit a well-known tension between fixed form and non-compositional meaning. Although idiomatic meanings are always not fully predictable from the composition of the meanings of their individual parts (Sag et al., 2002; Fraser, 1970; Strassler, 1982), idioms vary in how transparently those parts relate to the overall interpretation (Nunberg, 1978; Nunberg et al., 1994). In some cases, individual words map transparently onto aspects of the figurative meaning (e.g., *spill* → reveal, *beans* → secrets). In others, the figurative meaning cannot be attributed to any constituent (e.g., *kick, bucket,* → *?*). This property, defined as *decomposability* (Gibbs et al., 1989b; Wasow et al., 1983), has been linked to various behavioural and grammatical phenomena, most notably patterns of syntactic flexibility (Nunberg, 1978).

This intuition is formalised in the Idiom Decomposability Hypothesis (IDH; (Gibbs and Nayak, 1989; Gibbs et al., 1989b)). The IDH holds that idioms differ in semantic decomposability and that speakers share stable intuitions about how constituent meanings contribute to the whole. These intuitions are predicted to govern syntactic behaviour: the more decomposable an idiom is, the more syntactic flexibility it allows.

Decomposability has played a central role in prior works not merely as a descriptive notion, but as part of a broader explanatory account. Hybrid processing theories argue that decomposability reflects internal semantic structure (Cacciari and Levorato, 1998). According to this view, when an idiom's figurative meaning is distributed over its constituent words, those constituents remain accessible to grammatical operations. Consequently, an idiom can undergo syntactic modification without losing its figurative interpretation (Nunberg et al., 1994; Gibbs and Nayak, 1989; Riehemann, 2001). This predicts a systematic relationship between semantic decomposability and syntactic flexibility, as illustrated in (1–4). While the decomposable idiom "*pop the question*" retains its figurative interpretation under passivisation (1-2), the non-decomposable idiom "*kick the bucket*" does not (3-4).

(1) He **popped the question**.

(2) The **question** was **popped** by him.

(3) He **kicked the bucket.**

(4) # The **bucket** was **kicked** by him.[1]

[1]# is used to denote loss of idiomatic meaning.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34208–34229

An alternative perspective emerges from usage-based and constructionist approaches to linguistic knowledge. These accounts argue that form-meaning pairings are learned through exposure. Such pairings become entrenched as a function of frequency, predictability, and contextual diversity, rather than being derived from constituent-to-meaning mappings (Bybee, 2010; Goldberg, 2005). Applied to idioms, this perspective predicts that syntactic behaviour reflects distributional experience and familiarity. It also implies that the effects of decomposability may be unstable across tasks and items, as observed in psycholinguistic studies (Libben and Titone, 2008; Tabossi et al., 2009b; Nordmann et al., 2014). These two perspectives therefore differ in what they take to be the source of idiom generalisations: internal semantic structure on the one hand, or distributional experience on the other.

Human ratings of decomposability have long served as the primary empirical basis for evaluating these competing accounts (Partee, 1995). While such judgments provide valuable insight into speakers' linguistic intuitions, they reflect the full range of cognitive resources humans bring to language use and acquisition (Hubbard et al., 2023). As a result, they offer limited leverage on a distinct but fundamental question: which aspects of idiom behaviour arise from exposure to usage alone? Contextualised language models provide a principled way to probe this question. Trained on large amounts of text, these models acquire representations through distributional exposure, without explicit access to semantic role structure or acceptability judgments (Chang and Bergen, 2022; OpenAI, 2022; Touvron et al., 2023; Ficarra et al., 2025). Studying idioms in this setting isolates what can be learned from distributional exposure alone (Mi et al., 2025b).

In this work, we use contextualised language models as controlled distributional learners to investigate the relationship between decomposability, syntactic flexibility, and usage-based factors. We introduce a representational diagnostic of decomposability and relate it to human judgments, corpus-based measures of syntactic flexibility, and predictability. This perspective allows us to revisit the classic IDH theory from a mechanism-sensitive standpoint. Together, these analyses test whether decomposability explains idiom learning under purely distributional exposure, or whether usage-based factors account for the same variance.

We address the following **research questions**:

1. To what extent do measures of decomposability derived from models align with decomposability judgments by humans? (Section 5.2)

2. How does representational decomposability relate to syntactic flexibility and usage-based factors such as predictability in a distributional learner? (Section 5.5)

3. How do idiomatic representations evolve during pretraining, and are these dynamics better explained by decomposability or by usage-based predictability? (Section 6)

Contributions. We make three contributions. First, we show that model-derived decomposability exhibits only a weak positive correspondence with human judgments. This suggests that the two operationalisations reflect overlapping but distinct idiom properties. Second, we find a weak but consistent negative correlation between representational decomposability and corpus-based measures of syntactic flexibility across models and layers, with the strongest effects for prepositional phrasal idioms, contrary to decomposability-based predictions. Third, analysing training checkpoints reveals that the emergence and stabilisation of idiomatic representations are best explained by surprisal and decomposability. Frequency alone offers limited explanatory contribution, underscoring the dominant role of distributional predictability over raw exposure.

2 Related Work

IDH and the Compositionality Debate. Early accounts treated phrasal idioms as noncompositional units whose meanings are not derived from their parts (Chomsky, 1980; Fraser, 1970; Heringer, 1976; Katz, 1973). While this view captures the semantic opacity of many idioms, it offers limited explanation on why idioms differ systematically in their syntactic flexibility. In response, later work proposed the Idiom Decomposability Hypothesis (IDH), according to which idioms vary in the extent to which their figurative meanings can be distributed over their component words, and this variation constrains grammatical behaviour. In influential psycholinguistic work, Gibbs and Nayak (1989) show that many idioms are judged to be partially decomposable, and that speakers' beliefs

about decomposability predict acceptability judgments of syntactic alternations. On this view, differences in idiom flexibility are taken to reflect internal semantic structure, shaped by semantic and pragmatic factors in addition to syntax (Nunberg et al., 1994). At the same time, subsequent work has shown that decomposability judgments are gradient, task-dependent, and subject to considerable inter-speaker variability, and that their relationship to syntactic flexibility is less robust than originally assumed (Libben and Titone, 2008; Tabossi et al., 2009b; Wierzba et al., 2023; Sheinfux et al., 2019). This empirical tension motivates our work, revisiting the decomposability-flexibility relationship from the perspective of distributional learning.

Decomposability and Idiom Processing. A related line of work has examined whether idiom decomposability influences online processing and recognition. Early accounts propose that decomposable idioms are processed compositionally, using the same lexical retrieval and syntactic parsing mechanisms as literal language, whereas non-decomposable idioms are retrieved as single lexical units (Gibbs et al., 1989a). However, subsequent studies have challenged this sharp distinction (Sprenger et al., 2006; Tabossi et al., 2009a). In particular, Tabossi et al. (2009a) find no interaction between decomposability and processing speed in semantic judgment tasks, with both idiom types recognised equally quickly. More broadly, these findings suggest that decomposability effects are task-dependent and may not generalise across different aspects of idiom behaviour, such as recognition speed versus syntactic flexibility. This variability further motivates a mechanism-sensitive examination of which idiom patterns emerge under different learning and processing settings.

Challenges with Quantifying Decomposability. Human ratings have been central to investigations of IDH. However, several norming and processing studies suggest that such judgments are neither uniform nor stable across speakers and tasks. A norming study by Titone and Connine (1994) found that fewer than half of idioms elicited consistent judgments, with substantial inter-speaker disagreement. This pattern persisted in later eye-tracking work (Titone and Connine, 1999), which revealed only weak, participant-specific effects of decomposability. As observed by Libben and Titone (2008), idioms are "*multidetermined*" and Tabossi et al. (2009b) suggests that, these findings undermine the assumption that idioms can be cleanly classified as decomposable or not based on shared intuitions. These factors motivate our work, which complements this line of research by examining decomposability from a perspective that does not rely solely on human semantic judgments.

3 Formalisations

3.1 Theoretical Stipulations

We formalise decomposability under a set of theoretical stipulations grounded in prior work:

- **(IDH) Decomposability-syntactic flexibility link:** Decomposability is assumed to correlate (positively) with syntactic flexibility (Nunberg et al., 1994).
- **(S1) Semantic alignment:** Decomposability can be understood as the degree of semantic alignment between constituent representations and the idiom's overall figurative meaning (Nunberg et al., 1994).
- **(S2) Distributed meaning:** Decomposability reflects the extent to which an idiom's figurative meaning is distributed across its constituent parts (Nunberg et al., 1994).
- **(S3) Gradient:** Decomposability and syntactic flexibility are gradient properties that vary along a continuum[2](Sheinfux et al., 2019).

By operationalising decomposability in a manner consistent with **(S1–S3)**, our formulation allows us to test the Idiom Decomposability Hypothesis (**IDH**) independently, by evaluating whether the resulting decomposability measure predicts syntactic flexibility in usage.

3.2 Decomposability

Idioms are defined as contextually dependent semantic units characterised by a tension between their holistic, figurative interpretation and the literal meanings of their constituent words (Gibbs Jr and Colston, 2012; Nunberg et al., 1994). Figurative meanings arises when an expression's interpretation cannot be compositionally derived from its parts and instead emerges in context. Modelling this form of non-compositionality therefore

[2]Earlier accounts have proposed categorical distinctions, including a binary classification (Nunberg et al., 1994) a three-way distinction (Gibbs and Nayak, 1989), whereas more recent work argues for a continuum-based view (Sheinfux et al., 2019).

requires representations integrating information across the entire expression while also providing context-sensitive token-level representations. Contextualised language models meet these requirements by jointly encoding sentence-level context and token-level representations learned from distributional exposure (Vaswani et al., 2017; Devlin et al., 2019). In particular, bidirectional transformer-based models allow each token representation to be informed by its surrounding context, making them well-suited for analysing how figurative meaning is distributed across idioms.

From this perspective, hidden-state geometry offers a principled basis for operationalising decomposability. If an idiom is decomposable, we expect that (i) perturbing the idiom's constituent tokens will induce systematic changes in the hidden-state representation of the full expression, and (ii) the influence of individual constituents will be reflected in, and partially traceable from, the resulting representations. Conversely, if an idiom is non-decomposable, modifying a component should have weaker or less systematic effects on the model's internal representation of the sentence.

Let f denote a pretrained bidirectional transformer encoder (BERT) that maps an input sentence to contextualised token representations. For an input sentence $s = (w_1, \ldots, w_n)$, the encoder produces hidden states

$$\mathbf{h}_j(s) = f(s)_j \in \mathbb{R}^d, \qquad j = 1, \ldots, n.$$

where j denotes the token position. Because the encoder is bidirectional, each token representation incorporates information from both left and right contexts. This is important for modelling idioms as multi-token expressions whose figurative meaning depends on context.

We obtain a sentence-level representation via a pooling operation $P(\cdot)$ over token representations:

$$\mathbf{e}(s) = P(\mathbf{h}_1(s), \ldots, \mathbf{h}_n(s)) .$$

Let s_g denote the sentence obtained by replacing the idiom in context with a paraphrastic figurative gloss expressing its intended meaning. The corresponding pooled representation is:

$$\mathbf{e}(s_g) = P\big(\mathbf{h}_1(s_g), \ldots, \mathbf{h}_{n_g}(s_g)\big) .$$

In line with **S1**, we operationalise semantic alignment between an idiom in context and its figurative meaning. Let $\mathrm{sim}(\cdot, \cdot)$ denote a similarity measure between sentence representations (e.g., cosine similarity, Centered Kernel Alignment (CKA) (Kornblith et al., 2019), or the Wasserstein distance (Kantorovitch, 1958; Vaserstein, 1969)). We define the figurative similarity score for the full idiom as $S_{\text{fig}} = \mathrm{sim}(\mathbf{e}(s), \mathbf{e}(s_g))$.

To estimate token-level contributions, let $I \subseteq \{1, \ldots, n\}$ denote the set of token indices corresponding to the idiom span. For each idiom token $j \in I$, we construct a leave-one-out variant by masking that token:

$$s^{(-j)} = (w_1, \ldots, w_{j-1}, \texttt{[MASK]}, w_{j+1}, \ldots, w_n).$$

We compute the pooled representation $\mathbf{e}(s^{(-j)})$ and its similarity to the gloss-replaced sentence:

$$S^{(j)}_{\text{mask}} = \mathrm{sim}\Big(\mathbf{e}\Big(s^{(-j)}\Big), \mathbf{e}(s_g)\Big) .$$

If a constituent contributes to the idiomatic meaning, removing it should disrupt the alignment between the idiom-in-context sentence and its figurative paraphrase. We therefore define the contribution of token j to the idiomatic meaning as the absolute change in similarity when it is masked, consistent with the semantic alignment assumption in **S2**[3]. Taking the absolute value ensures contributions are non-negative; using signed values would allow large positive and negative effects to cancel, thus obscuring each token's true influence and rendering distributional measures unstable or uninterpretable.

$$\Delta_j = \Big| S_{\text{fig}} - S^{(j)}_{\text{mask}} \Big| .$$

Aggregation. Given the set of token-level contributions scores for an idiom, $\mathcal{D} = \{\Delta_j \mid j \in I\}$, we obtain an expression-level decomposability score by aggregating these values. Aggregation is required because decomposability is defined at the level of the idiomatic expression, while contributions are estimated at the level of individual constituents. We consider four aggregation functions that capture different aspects of contribution distribution across constituents, including mean, maximum, Gini dispersion, and entropy[4]. This produces a continuous measure of decomposability, consistent with **S3**.

[3] See Appendix A.1 for worked examples

[4] Formal definitions of these aggregation functions are provided in Appendix A.2.

3.3 Syntactic Flexibility

We operationalise syntactic flexibility as the diversity of constructional environments in which an idiom occurs. Specifically, we group each attested occurrence of an idiom by constructional type and quantify flexibility as the evenness of its distribution across these types. Following Tabossi et al. (2009b) and Gibbs and Nayak (1989), we distinguish four constructional types in addition to the base form: adverb insertion, adjective insertion, passivization, and action nominalization. Idioms that are highly flexible occur across multiple constructions, whereas rigid idioms are concentrated in a single form.

Let C denote the set of mutually exclusive construction types covering the attested syntactic realisations of an idiom. For an idiom i, we estimate the probability of construction $c \in C$ from corpus counts as $p_{i,c} = \frac{n_{i,c}}{N_i}$, where $n_{i,c}$ is the number of occurrences of idiom i in construction c, and $N_i = \sum_{c \in C} n_{i,c}$ is the total number of occurrences of idiom i. We define the syntactic flexibility of idiom i as the Shannon entropy of its constructional distribution, which captures both the diversity of constructions an idiom occurs in and how evenly its occurrences are distributed across them.

$$H(i) = -\sum_{c \in C} p_{i,c} \log_2 p_{i,c}.$$

Entropy increases as an idiom appears in a larger number of constructions and as its occurrences are distributed more evenly across them. The maximum entropy occurs when all construction types are equally probable: $H_{max} = \log_2(|C|)$. We provide a worked example in Appendix B.

3.4 Frequency

The frequency of an idiom is calculated as the sum of its counts across all constructional frames described above.

3.5 Predictability

Following prior idiom norming work, we operationalise predictability as the probability of the idiom-final word given its preceding context (i.e., cloze completion; e.g., Vulchanova et al. (2019); Cacciari and Tabossi (1988)). This captures the degree of contextual constraint: once sufficient context is processed, the final word often completes the idiomatic configuration and triggers the figurative interpretation (Configuration Hypothesis; (Cacciari and Glucksberg, 1991; Titone and Connine, 1994)). As such, predictability reflects distributional expectations learned from usage rather than semantic transparency alone. To remain comparable with human norms, we use the model analogue, $\log P(\text{final word} \mid \text{context})$, computed via masked prediction in bidirectional models. When the final word is split into multiple subword tokens, we calculate predictability as the average of their log probabilities.

4 Experimental Set-up

4.1 Models

We employ bidirectional transformer models, as idiomatic interpretation often depends on information from both preceding and following context. Psycholinguistic evidence from eye-tracking studies indicates that human readers consult surrounding context during idiom interpretation (Titone and Connine, 1999); we treat this evidence as a motivation for using architectures that explicitly encode both left and right context. Accordingly, we use bidirectional architectures, namely, BERT (Devlin et al., 2019) `base, large` variants and ModernBERT (Warner et al., 2025), a reimplementation of the BERT architecture with updated training practices, which conditions on context of both sides of an idiom. Details for the models are provided in Appendix C.1.

4.2 Datasets

We use datasets composed of sentences in which idiomatic expressions are employed figuratively (e.g., "How have you *weathered the storm*?"), paired with glossed paraphrases that convey the same meaning without the idiom (e.g., "How have you *succeeded in getting through the difficult situation*?"). Our experiments draw on the IMPLI dataset (Stowe et al., 2022), which contains figurative sentences together with paraphrases in which each idiom is replaced by the gloss. A summary of the datasets is provided in Appendix C.2.

4.3 Frequency Extraction

We extract idiom frequencies from the enTenTen corpus via Sketch Engine (Jakubíek et al., 2013)[5]. enTenTen is a large, web-scale corpus of English that provides broad coverage of contemporary usage and supports lemmatised queries, making it well suited for estimating idiom frequencies across

[5] `www.sketchengine.eu`

diverse syntactic contexts. For each idiom, we lemmatise its base form and generate Corpus Query Language (CQL) patterns corresponding to each syntactic frame. We then query the corpus and record the resulting frequency counts.

4.4 Evaluation of Hypotheses

We evaluate competing accounts of idiom decomposability through a set of complementary analyses examining the relationship between representational decomposability, syntactic flexibility, and usage-based factors. In line with the IDH, we assess whether decomposability is positively associated with syntactic flexibility by computing Spearman's rank correlations between decomposability scores and entropy-based flexibility measures. To test usage-based accounts, we additionally examine how representational decomposability relates to frequency and predictability, and whether these factors account for variance attributed to decomposability in human judgments. All correlations are computed for representations extracted from each layer of each model, allowing us to assess the consistency of effects across representational levels. Finally, we analyse how idiomatic representations evolve over training time, testing whether decomposability or usage-based predictability better explains stabilisation dynamics during pretraining. We compare different similarity functions and embedding aggregation strategies to ensure that observed patterns are not driven by a particular representational choice.

5 Decomposability Across Humans, Models, and Syntax

This section assesses the relationship between idiom decomposability and syntactic flexibility across human judgments and model-derived representations. We first test the classic decomposability-flexibility claim using human ratings, then examine how decomposability is encoded in contextualised language models, and finally assess the role of usage-based factors. Given the large number of model configurations, layers, and decomposability metrics, we focus on robust qualitative patterns, full results are provided in Appendix E.1.

5.1 Human Decomposability Ratings and Syntactic Flexibility

We first test a central assumption of the Idiom Decomposability Hypothesis: that idioms judged as more decomposable are more syntactically flexible (Nunberg et al., 1994). Contrary to this assumption, we find no significant relationship between human decomposability ratings and corpus-based measures of syntactic flexibility across the 90 idioms shared between the Bulkes and Tanner dataset and IMPLI. This result aligns with prior work questioning the stability and predictive power of decomposability judgments and suggests that, while such ratings reflect perceived semantic transparency, they do not reliably predict syntactic behaviour. These findings motivate a broader comparison: if decomposability judgments are poor predictors of syntactic behaviour in human data, how do they relate to decomposability as encoded in distributional models?

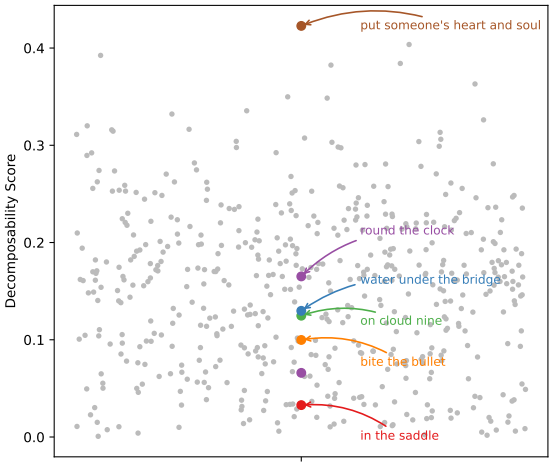


Figure 1: Decomposability scores obtained from the setup that most resembles human decomposability ratings. Most phrases cluster in the low-to-mid range, while "*put someone's heart and soul*" stands out as highly decomposable and "*in the saddle*" as among the least.

5.2 Human Ratings and Model-Derived Decomposability

We use contextualised models to assess whether a distributional learner recovers patterns similar to those reflected in human decomposability judgments. Across models, layers, and representational configurations, correlations between human ratings and model-derived decomposability are consistently weak but positive. The strongest correspondence is observed for BERT-large (Uncased) using final-layer representations, Wasserstein distance, and sum-based aggregation ($r(90) = .24$, $p = .005$). Figure 1 illustrates the decomposability scores calculated in this way.

Although statistically reliable, with robustness checks reported in Appendix D, this effect size indicates only a partial overlap. These results suggest that contextualised language models encode some aspects of the intuitions underlying human decomposability ratings, while also diverging from them in systematic ways.

5.3 Representational Decomposability and Syntactic Flexibility

We next examine whether model-derived decomposability covaries with syntactic flexibility, as predicted by semantic accounts of idioms. Across models and layers, Spearman correlations are consistently small and frequently negative (maximum $r(527) = -.16$, $p = .0002$), indicating little systematic alignment between representational decomposability and syntactic variability. Larger models, such as BERT-large and ModernBERT-large, yield a greater number of statistically significant correlations, though effect sizes remain modest. These effects are most consistently observed in earlier layers, particularly layer 2, which prior work has associated with local syntactic and phrasal information rather than abstract semantic composition (Jawahar et al., 2019; Tenney et al., 2019).

5.4 Idiom-Type Differences in the Decomposability-Flexibility Relation

The relationship between decomposability and syntactic flexibility is not uniform across idiom types. Using the most human-aligned decomposability configuration, we analyse correlations separately by coarse syntactic category (Table 1). Prepositional phrase idioms (PP+NP), such as "*off the hook*" and "*in a nutshell*", exhibit consistently stronger and statistically significant negative correlations than other idiom types. For these expressions, higher representational decomposability is associated with reduced syntactic flexibility. This pattern plausibly reflects the morpho-syntactic rigidity of prepositional constructions: lacking a verbal head, they support fewer syntactic alternations, such as passivisation or inflection, independently of how figurative meaning is distributed across constituents. Notably, for verb phrase idioms-the class most directly targeted by the **IDH**-we do not observe a significant relationship between decomposability and syntactic flexibility. This absence of effect in the theoretically most relevant domain weakens support for a decomposability-based account.

Structures	n	ρ	p
VP	284	-0.02	0.68
PP	127	-0.24	0.01*
NP	85	-0.15	0.18
ADJP	15	0.23	0.40
ADVP	10	0.12	0.75
OTHER(NUM)	3	-	-
S	2	-	-
OTHER(INTJ)	1	-	-

Table 1: Spearman's rank-order correlations between decomposability and syntactic flexibility by idiom type. Correlations are computed using decomposability measures derived from the most human-aligned setting (i.e., BERT-large uncased, last layer, Wasserstein and sum-based aggregreation); n indicates the number of idioms per coarse syntactic category. * denotes significant results.

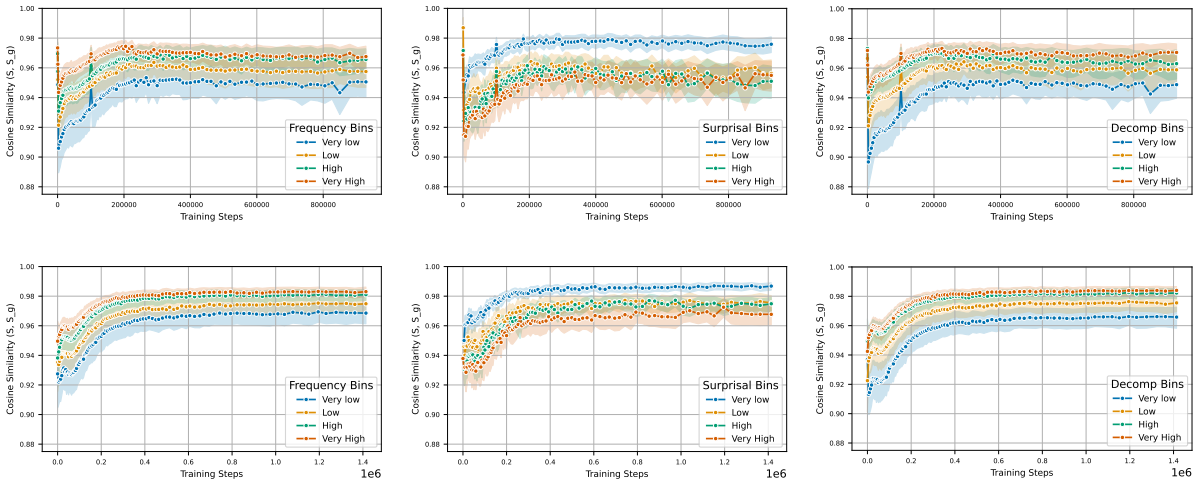
Human Ratings			
Variable	Coef	z	p
Predictability	-0.52	-0.33	0.73
Frequency	-0.20	-2.26	0.02*
Predictability x frequency	0.03	0.24	0.80
Model-derived Measures (BERT Large (Cased)			
Variable	Coef	z	p
Predictability	0.002	1.43	0.15
Frequency	-0.29	-4.07	0.000*
Predictability x frequency	0.002	0.98	0.32

Table 2: Regression results for measures from humans and BERT Large (Cased). Frequency is log-transformed. * denotes significant results.

5.5 Usage-Based Influences on Decomposability

To clarify what decomposability ratings reflect, we examine their relationship with two usage-based variables: predictability and frequency. Frequency corresponds to corpus-derived usage statistics, while familiarity ratings from Bulkes and Tanner (2017) reflect perceived exposure; for more information on human-derived measures and their model-derived counterparts, see Table 8.

We fit separate regression models for human- and model-derived decomposability measures (results in Table 2). For human ratings, neither familiarity nor predictability yields a significant effect. However, familiarity and corpus frequency may index different constructs: familiarity reflects subjective experience, whereas frequency captures objective distributional regularities. When familiarity ratings are replaced with corpus-derived frequency



(a) Frequency (b) Surprisal (c) Decomposability

Figure 2: Representation similarity over pretraining for OLMo-2 7B (top row) and OLMo-3 7B (bottom row), measured across 100 checkpoints. Each checkpoint is plotted individually, with colour indicating idiom characteristics (frequency, surprisal, and decomposability). All results are from layer 13.

counts, we observe a significant negative relationship between log-transformed frequency and decomposability ratings (coef $= -0.20$, $z = -2.26$, $p = 0.02$), indicating that more frequent idioms are judged as less decomposable. A parallel pattern emerges for model-derived decomposability. For BERT-large (cased), frequency shows a significant negative effect (coef $= -0.29$, $z = -4.07$, $p < .001$). No significant frequency effects are observed for other models. The negative effect of frequency on decomposability aligns with previous findings that higher frequency lowers the processing cost of multi-word expressions (Arnon and Snider, 2010; Sosa and MacFarlane, 2002). It suggests that frequently-used idioms are treated as single, holistic units that are stored and retrieved from memory, which is a characteristic of how non-decomposable idioms are processed (Gibbs et al., 1989b).

6 Acquisition Over-time

We examine how representations of idioms develop and become stable during pretraining in OLMo-2 (7B) (Walsh et al., 2025) and OLMo-3 (7B) (Team Olmo et al., 2025), using internal representations extracted from 100 checkpoints for each model. This allows us to study idiom learning as a dynamic process rather than as a static end-state. Our analysis focuses on how idiom-intrinsic properties interact with usage-based predictability during representation formation.

Specifically, we track learning trajectories for idioms varying in decomposability and contextual surprisal, where surprisal captures predictability beyond raw frequency (see Figure 2). This design allows us to disentangle the roles of structural interpretability (decomposability) and distributional expectations (surprisal) (Mi et al., 2025a) in shaping idiomatic representations over time, and to assess whether frequency alone (Mi et al., 2025b) suffices to explain their emergence under exposure-only learning.

Frequency in Pretraining Data. We estimate idiom frequency in the models' pretraining data using Infini-gram (Liu et al., 2024), which allows us to approximate the distribution of each idiom in the pretraining data as accurately as possible. See Appendix F for additional details.

Surprisal. Derived from Shannon's information theory (Shannon, 1948), surprisal quantifies how unexpected a word is within its context (Hale, 2001; Levy, 2008). It is defined as the negative log-probability of the token given its preceding context: $\text{Surprisal}(t_i) = -\log_2 P(t_i \mid t_{<i})$.

Decomposability. We use the model-derived decomposability score that produced the highest correlation with human ratings for this analysis.

Performance Measure. To track idiom learning, we measure representational similarity between a sentence containing an idiomatic expression (S) and a corresponding glossed sentence (S_g) that explicitly encodes the idiom's intended figurative

	Coef	z	p
Steps x Frequency	-0.0008	-24.69	<0.001
Steps x Surprisal	-0.0007	-22.301	<0.001
Steps x Decomposability	-0.0010	-36.367	<0.001

Table 3: Interaction effects between training progress and idiomaticity characteristics. Coefficients from a linear regression of model scores on training steps and input properties. Robust (HC3) standard errors; all variables z-scored. Frequency is log-transformed.

meaning. We compute this similarity using cosine similarity across layers and across 100 training checkpoints[6].

6.1 Results

Training Dynamics of Idiom Properties. We examine how the influence of idiom properties changes over training using a linear regression with interactions between training progress (steps) and three standardized predictors: log frequency, surprisal, and decomposability. All variables were z-scored, and HC3 robust standard errors were used. Results are summarised in Table 3; full results are presented in Appendix E.1.1.

All three interaction terms are negative and highly significant ($p < .001$), which indicates that the effects of frequency, surprisal, and decomposability systematically diminish as training progresses. Pearson correlation analysis among the regression predictors further indicates that multicollinearity is unlikely to affect the estimates (see Appendix D.3).

Relative Contributions of Idiom Properties. Among the three predictors, decomposability shows the largest interaction magnitude, which indicates the strongest training-dependent shift. This suggests that decomposability plays a more prominent role in shaping representations early in training, relative to frequency and surprisal.

Early Versus Late Training Effects. The observed interactions do not imply that idiom properties are inherently detrimental to learning. Rather, their total effect depends on training stage: they exert stronger influence early in training, with their impact attenuating as representations stabilise. This pattern is consistent with decreasing reliance on expression-level cues as the model accumulates distributional evidence over training.

[6]We focus on `Stage 1` pretraining, prior to instruction tuning (`Stage 2`), and DPO (`Stage 3`) in OLMo's training.

7 Conclusion

Decomposability has long been central to debates about whether idiom behaviour reflects internal semantic structure or emerges from usage and exposure. In this work, we revisited this question using contextualised language models as controlled distributional learners, allowing us to isolate what can be learned from distributional information alone.

Our results offer a clear challenge to IDH. We found that decomposability exhibits limited and unstable links to syntactic flexibility-most notably failing to hold even within verb phrase idioms, the core empirical domain where the IDH should, in theory, be most robust. This absence of effect suggests that decomposability is not a reliable mechanism-neutral explanation for how idioms behave syntactically.

Instead of syntactic links, we find a robust negative relationship between frequency and decomposability, supporting the view that high-frequency idioms are represented holistically. However, our analysis of pretraining dynamics reveals a more nuanced picture: frequency alone fails to explain how idiomatic representations emerge. Rather, surprisal and decomposability drive the stabilization of these meanings over time.

Ultimately, our work demonstrates that investigating the internal representations of language models provides a powerful method for adjudicating theoretical debates. By isolating the signal available from exposure alone, we can distinguish between properties inherent to semantic structure and those that emerge naturally from distributional learning-thereby helping to disentangle the confounded factors that often underlie human linguistic judgments.

Limitations

We recognise that our proposed decomposability metric represents just one possible way of operationalising this phenomenon. Decomposability is a complex property of idioms and has not yet been empirically studied. As such, we aim to offer an initial exploration in this direction; however, we do not claim that our approach is the only valid one.

A limitation of our pretraining analysis is the reliance on decomposability scores derived from a specific architecture (BERT-large) to predict the dynamics of another (OLMo). While using model-internal measures as fixed predictors is conceptually analogous to using human ratings or

corpus-based frequency, model-derived decomposability is not architecture-neutral. Given that different models exhibit varying correlations with human judgments and syntactic flexibility, these scores may carry architecture-specific biases. Ideally, decomposability would be computed directly from the model under study; however, the context-sensitive nature of idiomaticity requires a bidirectional model, which is not applicable to causal models like OLMo. Future work should explore architecture-agnostic measures to ensure broader generalisability across different model families.

Finally, our analysis is based on a set of English idioms and may not extend to other languages. Investigating how decomposability manifests in cross-lingual studies is an interesting avenue for future work.

Ethical Considerations

AI-assisted writing and coding tools were used in compliance with the ACL Policy on the Use of AI Writing Assistance.

Acknowledgments

We thank the reviewers for their comments and feedback. We acknowledge IT Services at The University of Sheffield and the University of Oxford Advanced Research Computing (ARC) for the provision of services for High Performance Computing. Finally, we thank Agata Savary and members of the MWE community for their discussions.

MM is supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. For the purpose of open access, we have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. GA is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – FIP-12 – [Project-ID 528483508], and the European Union [grant number 101079894]. AY is supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/W524360/1] and the Japan Student Services Organization (JASSO) Student Exchange Support Program (Graduate Scholarship for Degree Seeking Students). AV's research is partly supported by UKRI (grants MR/U506734/1 and EP/T02450X/1), CNPq (406926/2025-5) and EQUATE.

References

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82.

Nyssa Z. Bulkes and Darren Tanner. 2017. "going to town": Large-scale norming and statistical analysis of 870 american english idioms. *Behavior Research Methods*, 49(2):772–783.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.

Cristina Cacciari and Sam Glucksberg. 1991. Chapter 9 understanding idiomatic expressions: The contribution of word meanings. In Greg B. Simpson, editor, *Understanding Word and Sentence*, volume 77 of *Advances in Psychology*, pages 217–240. North-Holland.

Cristina Cacciari and Maria Chiara Levorato. 1998. The effect of semantic analyzability of idioms in metalinguistic tasks. *Metaphor and Symbol*, 13(3):159–177.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language*, 27(6):668–683.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences*, 3(1):1–15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Filippo Ficarra, Ryan Cotterell, and Alex Warstadt. 2025. A distributional perspective on word learning in neural language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11184–11207, Albuquerque, New Mexico. Association for Computational Linguistics.

Bruce Fraser. 1970. Idioms within a transformational grammar. *Foundations of language*, pages 22–42.

Raymond W Gibbs and Nandini P Nayak. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21(1):100–138.

Raymond W Gibbs, Nandini P Nayak, John L Bolton, and Melissa E Keppel. 1989a. Speakers' assumptions about the lexical flexibility of idioms. *Memory & cognition*, 17(1):58–68.

Raymond W. Gibbs, Nandini P. Nayak, and Cooper Cutting. 1989b. How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, 28(5):576–593.

Raymond W Gibbs Jr and Herbert L Colston. 2012. *Interpreting figurative meaning*. Cambridge University Press.

Adele Goldberg. 2005. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

James T. Heringer. 1976. *Idioms and Lexicalization in English*, pages 205 – 216. Brill, Leiden, The Netherlands.

Ryan Hubbard, Nyssa Bulkes, and Vicky Tzuyin Lai. 2023. Predictability and decomposability separately contribute to compositional processing of idiomatic language. *Psychophysiology*, 60(8):e14269.

Milo Jakubíek, Adam Kilgarriff, Vojtěch Kovář, P. Rychlý, and Vít Suchomel. 2013. The tenten corpus family.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

L. Kantorovitch. 1958. On the translocation of masses. *Management Science*, 5:1–4.

Jerrold J. Katz. 1973. Compositionality, idiomaticity, and lexical substitution. In Stephen R. Anderson and Paul Kiparsky, editors, *A Festschrift for Morris Halle*, pages 357–376. Holt, Rinehart and Winston, New York.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Maya R Libben and Debra A Titone. 2008. The multidetermined nature of idiom processing. *Memory & cognition*, 36(6):1103–1121.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025a. From input perception to predictive insight: Modeling model blind spots before they become errors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34328–34341, Suzhou, China. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025b. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Emily Nordmann, Alexandra A. Cleland, and Rebecca Bull. 2014. Familiarity breeds dissent: Reliability analyses for british-english idioms on measures of familiarity, meaning, literality, and decomposability. *Acta Psychologica*, 149:87–95. Including Special section articles of Temporal Processing Within and Across Senses - Part-2.

Geoffrey Nunberg. 1978. *The Pragmatics of Reference*. Indiana University Linguistics Club, Bloomington, IN.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`. Accessed: 2026-01-05.

Barbara H. Partee. 1995. *Lexical semantics and compositionality.*, pages 311–360. An invitation to cognitive science. The MIT Press, Cambridge, MA, US.

Susanne Riehemann. 2001. A constructional approach to idioms and word formation.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, page 1–15, Berlin, Heidelberg. Springer-Verlag.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Livnat Herzig Sheinfux, Tali Greshler, Nurit Melnik, and Shuly Winter. 2019. Verbal multiword expressions: Idiomaticity and flexibility. *Representation and parsing of multiword expressions: Current trends*, 3:35.

Anna Vogel Sosa and James MacFarlane. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and language*, 83(2):227–236.

Simone A Sprenger, Willem JM Levelt, and Gerard Kempen. 2006. Lexical access during the production of idiomatic phrases. *Journal of memory and language*, 54(2):161–184.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. Impli: Investigating nli models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388.

Jurg Strassler. 1982. Idioms in english. *A Pragmatic Analysis. Tubingen: Narr*.

Patrizia Tabossi, Rachele Fanari, and Karoline Wolf. 2009a. Why are idioms recognized fast? *Memory & Cognition*, 37(4):529–540.

Patrizia Tabossi, Rachele Fanari, and Kristine Wolf. 2009b. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2):313–327.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Debra A. Titone and Cynthia M. Connine. 1994. Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbolic Activity*, 9(4):247–270.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12):1655–1674. Literal and Figurative Language.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

L. N. Vaserstein. 1969. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mila Vulchanova, Evelyn Milburn, Valentin Vulchanov, and Giosuè Baggio. 2019. Boon or burden? the role of compositional meaning in figurative language processing and acquisition. *Journal of Logic, Language and Information*, 28(2):359–387.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Thomas Wasow, Ivan Sag, and Geoffrey Nunberg. 1983. Idioms: An interim report. In *Proceedings of the XIIIth international congress of linguistics*, volume 29. Tokyo.

Marta Wierzba, Jessica MM Brown, and Gisbert Fanselow. 2023. Sources of variability in the syntactic flexibility of idioms. *Glossa: a journal of general linguistics*, 8(1).

A Decomposability Score Derivation

A.1 Examples of Derivation

In the following subsections, we demonstrate a worked example showing token-level Δ values for two canonical cases (e.g., rule of thumb (non/less decomposable idiom) vs. food for thought (more decomposable idiom).

It is important to note that, throughout our analysis, we use the ranked position of each idiom as an indicator of its degree of decomposability, rather than relying on the raw values themselves.

A.1.1 "rule of thumb"

Sentence: The rule of thumb for working out maximum heart rate (MHR) is 225 minus your age in years.

Sentence glossed: The general principle for working out maximum heart rate (MHR) is 225 minus your age in years.

Masked Variants:

- `[CLS] the [MASK] of thumb for working out maximum heart rate (MHR) is 225 minus your age in years. [SEP]`
- `[CLS] the rule [MASK] thumb for working out maximum heart rate (MHR) is 225 minus your age in years. [SEP]`
- `[CLS] the rule of [MASK] for working out maximum heart rate (MHR) is 225 minus your age in years. [SEP]`

Token-level Importance (Δ_j) - sorted.

$$\text{rule}_\Delta = 0.0774$$
$$\text{thumb}_\Delta = 0.0666$$
$$\text{of}_\Delta = 0.0663$$

Aggregation. sum

Idiomatic Decomposability Score. 0.2103 *(more transparent to figurative meaning)*

A.1.2 "food for thought"

Sentence. It won't do any harm, but I'd rather not give him food for thought, because I consider him an idiot and I don't think he's capable of interpreting it correctly.

Sentence glossed. It won't do any harm, but I'd rather not give him anything that should be thought about, because I consider him an idiot and I don't think he's capable of interpreting it correctly.

Masked Variants.

- `... rather not give him [MASK] for thought, because I consider him an idiot ...`
- `... rather not give him food [MASK] thought, because I consider him an idiot ...`
- `... rather not give him food for [MASK], because I consider him an idiot ...`

Token-level Importance (Δ_j) - sorted.

$$\text{thought}_\Delta = 0.0137$$
$$\text{for}_\Delta = 0.0105$$
$$\text{food}_\Delta = 0.0082$$

Aggregation. sum

Idiomatic Decomposability Score. 0.0324 *(less transparent to figurative meaning)*

A.2 Aggregation

The formulations of our aggregation metrics, for calculating expression-level decomposability score.

Mean

$$Decomp_{\text{mean}} = \frac{1}{n}\sum_{j=1}^{n}\Delta_j. \tag{1}$$

Maximum

$$Decomp_{\text{max}} = \max_{1\le j\le n}\Delta_j. \tag{2}$$

To define dispersion-based measures, we first normalise the contributions:

$$p_j = \frac{\Delta_j}{\sum_{i=1}^{n}\Delta_i}, \quad \sum_{j=1}^{n} p_j = 1. \tag{3}$$

Gini Dispersion

$$Decomp_{\text{Gini}} = 1 - \sum_{j=1}^{n} p_j^2. \tag{4}$$

Entropy

$$Decomp_{\text{Ent}} = -\sum_{j=1}^{n} p_j \log p_j. \tag{5}$$

B Syntactic flexibility computation

We provide an example for our computation of syntactic flexibility for the idiom *break somebody's heart*. We first extract the frequencies of the idiom in each constructional type using the corresponding queries shown in Table 4.

Construction	Query (lemmas)	Frequency	Probability
Base form	break [N/P]'s heart	84758	0.75
Adjective Insertion	break [N/P]'s [ADJ] heart	1719	0.01
Adverb Insertion	[ADV] break [N/P]'s heart	13092	0.12
Nominalization	breaking of [N/P]'s heart	15	0.00
Passive	[N/P]'s heart be break	13526	0.12

Table 4: Frequencies of the five constructional patterns of the idiom *break somebody's heart* and their corresponding search queries. Except for the action nominalization *breaking*, the verb was lemmatized in the searches to cover inflection and tense variation, and the slot somebody was implemented as a noun/pronoun search.

We next obtain the probabilities by dividing each individual frequency by the total frequency. The syntactic flexibility is then measured using Shannon entropy:

$$H(\text{break somebody's heart}) = -(0.75 \log 0.75 + 0.01 \log 0.01 + 0.12 \log 0.12 + 0.12 \log 0.12) \approx 0.77$$

C Experimental Details

C.1 Models

We use a total of 8 language models in this work. Table 5 provides information regarding models used this paper.

Model	Size	Tokens	Layers
bert-base-cased	110M	30K	12
bert-base-uncased	110M	30K	12
bert-large-cased	336M	30K	24
bert-large-uncased	336M	30K	24
ModernBert-base	149M	2T	22
ModernBert-large	395M	2T	28
Olmo-2-1124-7B	7B	4T	32
Olmo-3-1025-7B	7B	5.93T	32

Table 5: Comparison of language models used in this study.

C.2 Datasets

We provide an breakdown of the datasets used this in work in Table 7. The idioms in IMPLI fall primarily into the functional categories listed in Table 6.

Category	*n*	Category	*n*
VP	284	ADJP	15
PP	127	ADVP	10
NP	85	OTHER (NUM)	3
		S	2
		OTHER (INTJ)	1

Table 6: Distribution of idiomatic expressions across coarse-grained syntactic categories in IMPLI. We use the Bulkes and Tanner norms dataset (Bulkes and Tanner, 2017) for its human decomposability ratings. Specifically, we focus on the subset of idioms that overlap with IMPLI to enable direct comparison. The summary of the datasets are gathered in Table 7.

C.3 Model-Human Proxies

We present an illustration of the model and human equivalent features in Table 8.

D Robustness Analyses

D.1 Bootstrap Confidence Intervals for Model–Human Correlations

To assess the stability of the observed correlation, we conducted a bootstrap resampling analysis on the best-performing configuration (BERT-large, final-layer representations, Wasserstein distance, sum-based aggregation). Given the modest sample size (n = 90), this procedure provides a more robust estimate of variability. The resulting 95% confidence interval for the correlation coefficient was [0.07, 0.40]. As this interval excludes zero, the effect is unlikely to be attributable to sampling noise, although the width of the interval indicates substantial uncertainty in the magnitude of the relationship.

D.2 Partial Correlation between Predictors

Table 9 presents the partial correlation matrices for the model and human datasets. Most partial correlations are small in magnitude, indicating that the predictors are largely distinct. However, a few significant correlations, especially those involving frequency, suggest some modest small relationships among the variables.

D.3 Pearson Correlation between Acquisition Predictors

The Pearson correlations among log frequency, surprisal, and decomposability are small in magnitude, suggesting minimal multicollinearity among the regression predictors. These relationships are presented in Table 10.

D.4 Variance Inflation Factors

The variance inflation factors (VIFs) were examined to assess potential multicollinearity among the predictors included in the regression models. As shown in Table 11, all VIF values are close to 1 for both the model data and the human data, indicating that multicollinearity is negligible. This suggests that the predictors contribute relatively independent information and that the regression estimates are unlikely to be distorted by strong linear relationships among the explanatory variables.

E Additional Results

E.1 Layer-wise Correlation

We present the Spearman ranked correlation results of layer-wise decomposability derivations and syntactic flexibility in the six bidirectional models we

Dataset	# Samples	# Unique idioms	Example
IMPLI	527	382	Sentence (S): How have you weathered the storm? Paraphrase (S_g): How have you succeeded in getting through the difficult situation?
Bulkes & Tanner (subset)	90	90	Be a dark horse $\rightarrow$ 0.09 Go back to basics $\rightarrow$0.97

Table 7: Summary of datasets used in this study. IMPLI consists of idiomatic sentences paired with paraphrase. Bulkes & Tanner dataset provides idioms annotated with human ratings across five dimensions (Familiarity, Predictability, Global Decomposability, Meaningfulness, and Literal Plausibility).

Human measure	**Instruction**	**Response metric**	**Model Equivalent**
Familiarity	Rating how often Participants hear or use an idiom.	Likert rating.	Frequency (section 3.4)
Predictability	Sentence-completion (cloze): providing the first word that comes to mind.	Proportion of responses matching the expected item.	Predictability (section 3.5)
Global decomposability	Classification of idioms whose component parts contribute to their overall meaning.	Binary categorization.	Decomposability (section 3.2)

Table 8: Human-elicited measures (adapted from Bulkes and Tanner (2017)) and their model-derived counterparts.

	Model Data		Human Data	
	Predictability	Frequency	Predictability	Frequency
Predictability	1.000	-	1.000	-
Frequency	0.262*	1.000	-0.148	1.000
Decomposability	0.050	-0.236*	-0.085	-0.356*

Table 9: Partial correlation matrices for human and model data. * denotes significance.

	Log_Frequency	Surprisal	Decomp
Log_Frequency	1.000	-0.066	-0.005
Surprisal	-0.066	1.000	-0.188
Decomp	-0.005	-0.188	1.000

Table 10: Pearson correlation matrix among regression predictors.

Predictor	Model Data	Human Data
Predictability	1.074	1.023
Log_Frequency	1.0134	1.164
Decomposability	1.059	1.146

Table 11: Variance Inflation Factors (VIFs).

have tested. Empty bars show non-statistically significant results. For each model, the results are presented in its corresponding figure:

- BERT-base (Uncased): Figure 3
- BERT-base (Cased): Figure 4
- BERT-large (Uncased): Figure 5
- BERT-large (Cased): Figure 6
- ModernBERT-base: Figure 7
- ModernBERT-large: Figure 8

E.1.1 Linear Regression Results

Table 12 reports the full set of coefficient estimates from an ordinary least squares (OLS) regression predicting *score*. The model includes fixed effects for network layer (treated as a categorical variable) and model identity, as well as several standardised continuous predictors and their interactions.

All continuous predictors (`steps`, `log_frequency`, `surprisal`, and `decomp`) were z-standardised prior to estimation. Interaction terms capture moderation of each linguistic predictor by `steps`. The model was estimated using heteroskedasticity-robust (HC3) standard errors.

The table reports coefficient estimates, robust standard errors, z-statistics, two-sided p-values, and 95% confidence intervals.

F Frequency Extraction with Infini-Gram

Whilst Infini-gram can approximate the frequency of expressions in model's training data, Infini-Gram does not support lemma-based queries -only exact strings. Thus, we devise a method to "unlemmatise" the expressions to try and capture as many variants as possible. Consequently, we convert idioms from their base form to surface forms in three steps:

1. Add all verb inflections
2. Add both singular and plural noun forms
3. Replace "somebody" and "something" with appropriate possessive pronouns or possessive adjectives, depending on their syntactic position.

We use the `word_forms` package to obtain singular and plural noun forms, as well as all morphological variants of verbs.

Example. Idiom: *break somebody's heart*

1. break: break, breaking, broke, broken
2. somebody's: my, your, his, her, its, our, their
3. heart: heart, hearts

For this idiom, we therefore produce (4 x 7 x 2) 56 distinct queries.

	coef	std err	z	P>|z|	[0.025]	[0.975]
Intercept	0.9142	0	3102.648	0	0.914	0.915
C(layer)[T.1]	0.0425	0	123.268	0	0.042	0.043
C(layer)[T.2]	0.0423	0	124.69	0	0.042	0.043
C(layer)[T.3]	0.0354	0	102.021	0	0.035	0.036
C(layer)[T.4]	0.0354	0	102.53	0	0.035	0.036
C(layer)[T.5]	0.0346	0	100.533	0	0.034	0.035
C(layer)[T.6]	0.0385	0	114.276	0	0.038	0.039
C(layer)[T.7]	0.0394	0	117.605	0	0.039	0.04
C(layer)[T.8]	0.0414	0	124.666	0	0.041	0.042
C(layer)[T.9]	0.0432	0	131.538	0	0.043	0.044
C(layer)[T.10]	0.0419	0	126.643	0	0.041	0.043
C(layer)[T.11]	0.0435	0	132.437	0	0.043	0.044
C(layer)[T.12]	0.0437	0	132.897	0	0.043	0.044
C(layer)[T.13]	0.0438	0	133.522	0	0.043	0.044
C(layer)[T.14]	0.0433	0	131.941	0	0.043	0.044
C(layer)[T.15]	0.0432	0	131.702	0	0.043	0.044
C(layer)[T.16]	0.0416	0	126.204	0	0.041	0.042
C(layer)[T.17]	0.0399	0	120.001	0	0.039	0.041
C(layer)[T.18]	0.0386	0	115.658	0	0.038	0.039
C(layer)[T.19]	0.038	0	113.435	0	0.037	0.039
C(layer)[T.20]	0.0363	0	107.521	0	0.036	0.037
C(layer)[T.21]	0.0358	0	105.709	0	0.035	0.036
C(layer)[T.22]	0.0342	0	99.947	0	0.033	0.035
C(layer)[T.23]	0.0335	0	97.523	0	0.033	0.034
C(layer)[T.24]	0.0336	0	97.999	0	0.033	0.034
C(layer)[T.25]	0.0334	0	97.374	0	0.033	0.034
C(layer)[T.26]	0.0337	0	98.104	0	0.033	0.034
C(layer)[T.27]	0.0344	0	100.512	0	0.034	0.035
C(layer)[T.28]	0.0366	0	107.535	0	0.036	0.037
C(layer)[T.29]	0.0414	0	123.712	0	0.041	0.042
C(layer)[T.30]	0.0537	0	169.44	0	0.053	0.054
C(layer)[T.31]	0.0648	0	212.713	0	0.064	0.065
C(layer)[T.32]	0.0653	0	215.456	0	0.065	0.066
C(model)[T.Olmo-3-1025-7B]	0.0041	6.04e-5	67.731	0	0.004	0.004
steps_z	0.0037	2.77e-5	134.436	0	0.004	0.004
log_frequency_z	0.0085	3.70e-5	229.318	0	0.008	0.009
surprisal_z	-0.0065	3.21e-5	-202.956	0	-0.007	-0.006
decomp_z	0.0099	2.93e-5	336.962	0	0.01	0.01
steps_z:log_frequency_z	-0.0008	3.39e-5	-24.692	0	-0.001	-0.001
steps_z:surprisal_z	-0.0007	3.13e-5	-22.301	0	-0.001	-0.001
steps_z:decomp_z	-0.001	2.79e-5	-36.367	0	-0.001	-0.001

Table 12: Full OLS regression results predicting *score*. The model includes layer and model fixed effects, z-standardised predictors (steps, log frequency, surprisal, and decomposability), and interactions between steps and each linguistic predictor. Robust (HC3) standard errors are reported.

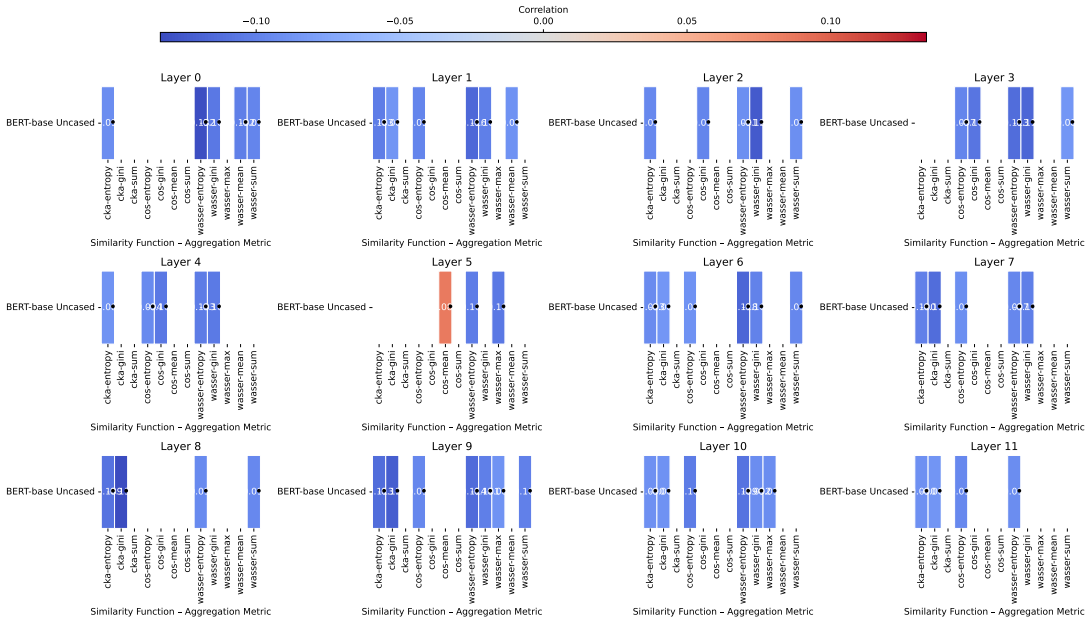


Figure 3: Correlation results for BERT-base Uncased

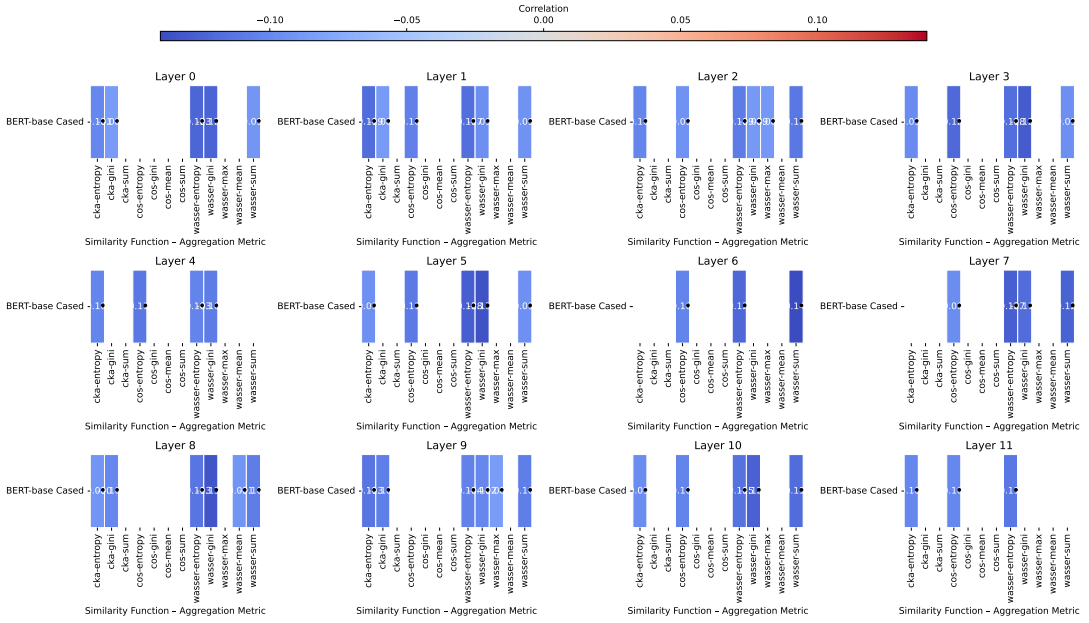


Figure 4: Correlation results for BERT-base Cased

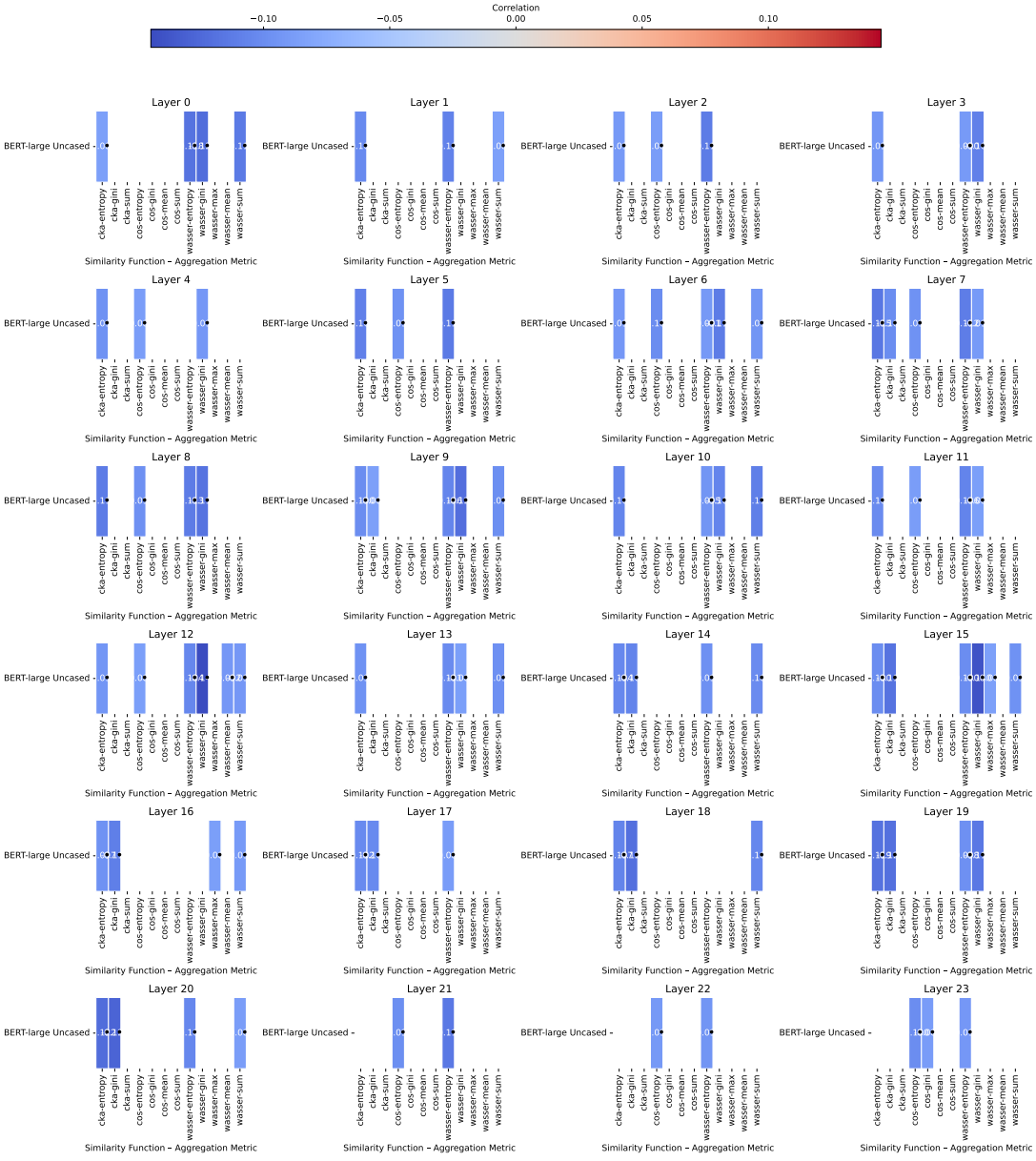


Figure 5: Correlation results for BERT-large Uncased

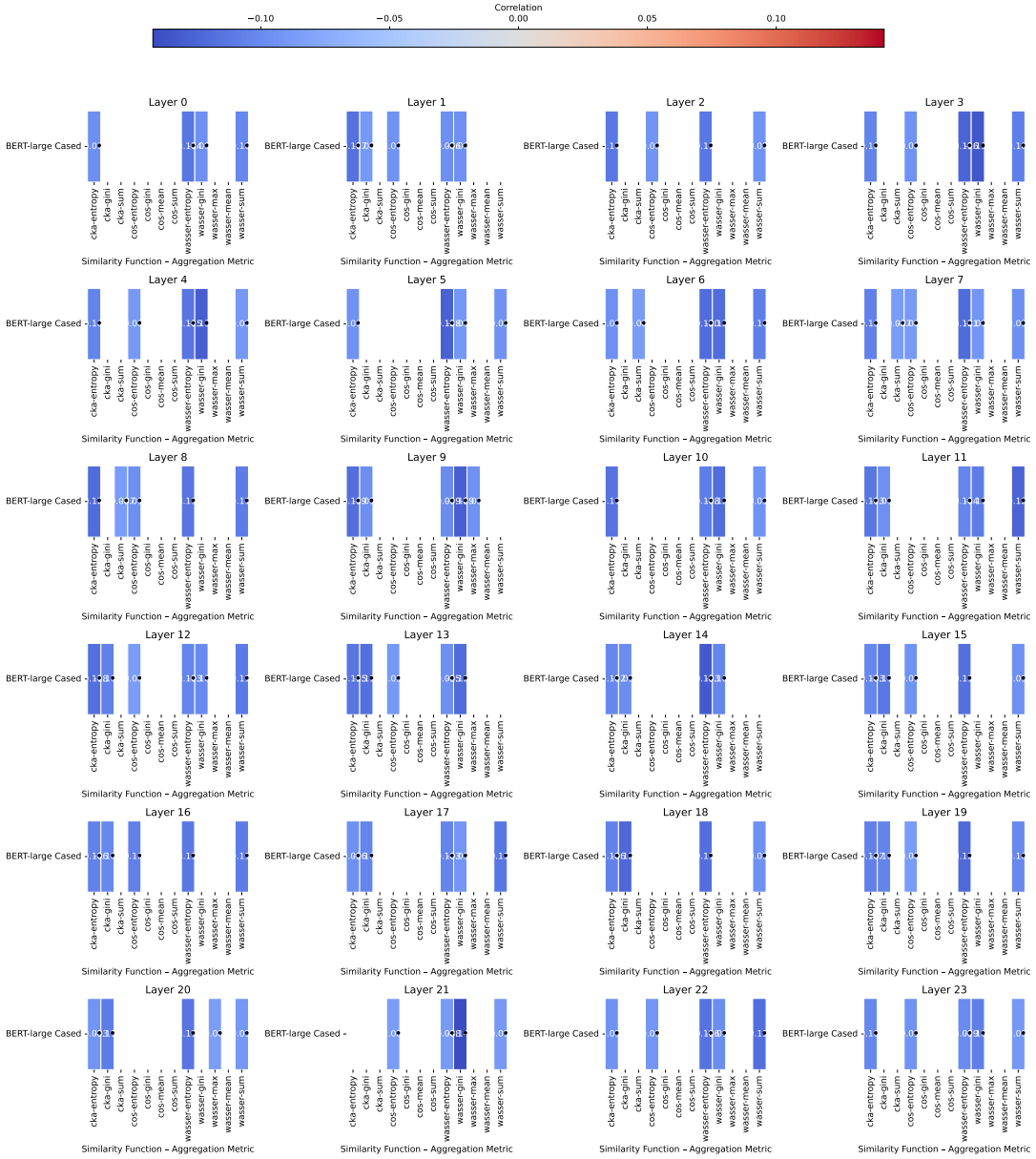


Figure 6: Correlation results for BERT-large Cased

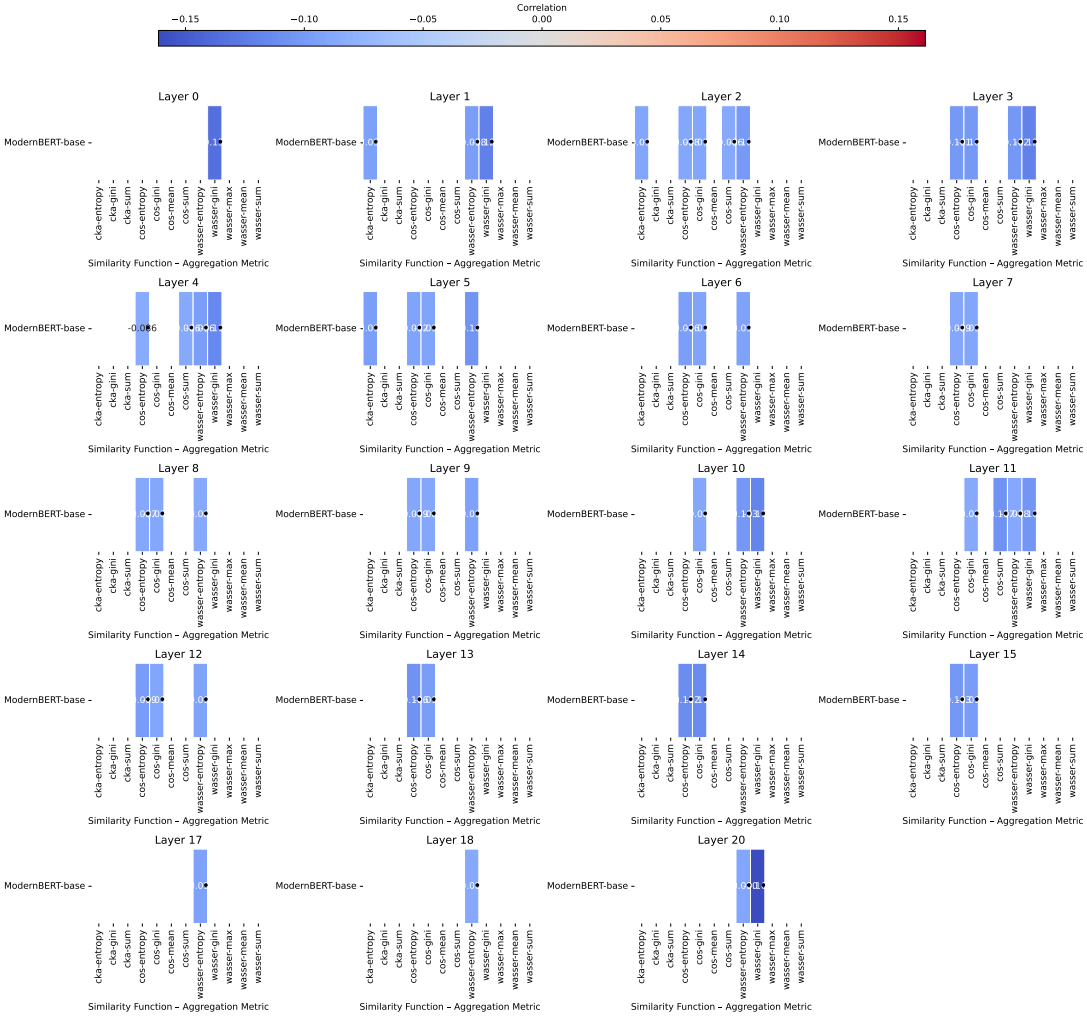


Figure 7: Correlation results for ModernBERT Base

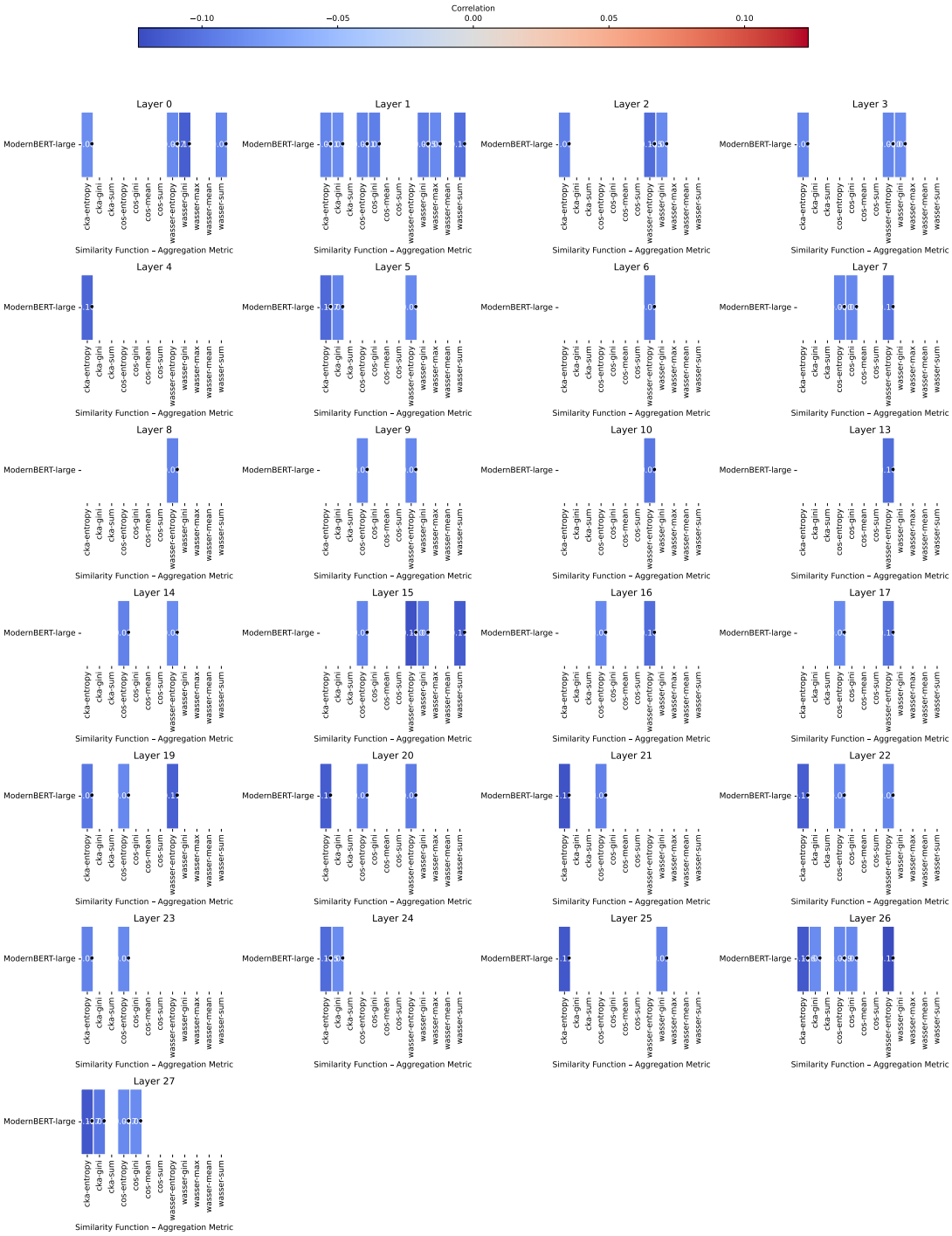


Figure 8: Correlation results for ModernBERT Large