# Measuring Idiomaticity in Text Embedding Models with $\varepsilon$-compositionality

**Sondre Wold**[*] and **Étienne Simon**[*] and **Erik Velldal** and **Lilja Øvrelid**
University of Oslo

## Abstract

The principle of compositionality, which concerns the construction of meaning from constituent parts, is a longstanding topic in various disciplines, most commonly associated with formal semantics. In NLP, recent studies have focused on the compositional properties of text embedding models, particularly regarding their sensitivity to idiomatic expression, as idioms have traditionally been seen as non-compositional. In this paper, we argue that it is unclear how previous work relates to formal definitions of the principle. To address this limitation, we take a theoretically motivated approach based on definitions in formal semantics. We present $\varepsilon$-compositionality, a continuous relaxation of compositionality derived from these definitions. We measure $\varepsilon$-compositionality on a dataset containing both idiomatic and non-idiomatic sentences, providing a theoretically motivated assessment of sensitivity to idiomaticity. Our findings indicate that most text embedding models differentiate between idiomatic and non-idiomatic phrases, although to varying degrees.

## 1 Introduction

This paper studies the principle of compositionality of meaning and its relation to sentence representations produced by text embedding models. There are numerous definitions of the principle, some of which will be discussed in this work, but the principle is most commonly associated with the idea that the meaning of a compound expression is constructed from the meanings of its parts (Janssen, 1986).

Compositionality has been studied particularly in relation to sentence meaning, most notably in the field of formal semantics (Montague et al., 1970; Janssen, 1986; Partee et al., 1995). Historically, *idioms* have been seen as a counterexample to compositional meaning construction, where the meaning of an expression is predominately determined by convention rather than composition (Baroni et al., 2014; Nefdt and Potts, 2024). As an informal example, the meaning of the phrase "couch potato",[1] is less compositional than the phrase "red wine", which is more directly composed of the two parts of the expression.

Recent work in NLP has focused on how existing sentence embedding models capture compositionality in the context of idiomatic language and how they align with human compositionality judgments (He et al., 2025; Fodor et al., 2025). There is also recent work on how LLMs handle compositionality in idiomatic nominal compounds in the text–visual domain (Kurtyigit et al., 2025). Common to these works is the treatment of compositionality as a general guiding principle about meaning construction in natural language. It is not clear, however, how the formal definitions of the principle are implemented in these works, and as such, there is a potential for an improved theoretical underpinning of the empirical measurements of compositionality.

To address this limitation, we take a theoretically motivated approach, starting with a precise formalization of the principle based on traditional definitions in formal semantics. We then introduce a continuous relaxation of this formalization, which we call $\varepsilon$-compositionality. In practical terms, $\varepsilon$-compositionality provides a way of distinguishing between compositional and non-compositional text embeddings that is grounded in theoretical accounts of compositionality. Intuitively, words are used in a $\varepsilon$-compositional context if their representations are pushed together when they are contextualized, whereas they are used in a non-$\varepsilon$-compositional context when contextualizing them moves their representation away. We measure $\varepsilon$-compositionality

[*] Equal contribution.

[1] "couch potato": an idler spending most of their time lying down.

*Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 2239–2252
March 24-29, 2026 ©2026 Association for Computational Linguistics

for a range of different text embedding models on the NCIMP dataset by He et al. (2025); Garcia et al. (2021), which consists of minimal sentence pairs with both compositional and idiomatic multi-word expressions (MWE). We use those compositional MWEs to control for confounding variables: instead of analyzing $\varepsilon$-compositionality values in isolation, we relate them to the $\varepsilon$ of compositional samples through a Wilcoxon signed-rank test. The shift in distribution between those two $\varepsilon$s is what characterize non-compositionality.

Expressions in NCIMP are labeled according to how compositional they are on a three-level scale (non-compositional, partially compositional, compositional). In our experiments, we evaluate a range of techniques for constructing text embeddings: different types of aggregation of embeddings from pre-trained encoder-based models, sentence embedding models based on contrastive post-training, and state-of-the-art embedding models based on instruction-tuning. We find that most of the evaluated models can differentiate between idiomatic and non-idiomatic language, but to varying degrees.

In summary, our contributions are, following the structure of the paper: *i)* We review and summarize current approaches to compositionality in the NLP and ML literature; *ii)* we introduce a theoretically motivated measurement of compositionality that we derive from existing definitions in formal semantics, which we call $\varepsilon$-compositionality; and *iii)* we evaluate $\varepsilon$-compositionality on a range of text embedding models, with results indicating that our measurement can distinguish between the sensitivity to compositionality displayed by said models.

## 2 Background

Compositionality is often regarded as a central challenge for current language models, driving both research (Press et al., 2023; Dziri et al., 2023) and commercially oriented benchmarking like the ARC-AGI challenge (Chollet et al., 2025). In machine learning, however, compositionality rarely aligns with its traditional meaning in formal semantics, philosophy of language, or mathematical logic. Instead, it often refers to the general decomposability of a problem, or the composability of suboperations in a machine learning system, with few theoretical constraints. This ambiguity, despite its frequent use in both machine learning and NLP (Sinha et al., 2024; McCurdy et al., 2024), makes it challenging to understand what the implications of the principle are for a given problem setting.

Our work aims to connect formal semantic accounts of compositionality of meaning in natural languages with the evaluation of compositionality in text embedding models in a principled manner. By doing this, we hope to make clear what the implication of compositionality is with respect to embedding models, and how it relates to its theoretical formalization.

### 2.1 Compositionality in Formal Semantics

The principle of compositionality is attractive primarily because it allows for a finite method to interpret the semantic content of an infinite number of expressions (van Bethem et al., 1991). That is, in its most basic form, the principle says something about how the semantics of a specific language parallels its syntax.

Compositionality has been of particular interest in the field of formal semantics, where the exact relationship between syntax and semantics has been formalized primarily using two different approaches. In one approach (Montague et al., 1970; Janssen, 1986), the interaction is defined using a *many-sorted algebra*, where the syntactic operations are only specified for specific *sorts* of linguistic constructions, such as nominal compounds, adverbial phrases etc. The other approach, taken for example by Hodges (2001); Pagin and Westerståhl (2010), is to define a *partial algebra*, so that the operation is undefined for sorts that do not match the specification of the operation arguments.

Regardless of the approach, the principle of compositionality is often presented in one of two forms: the function version or the substitution version. Following Pagin and Westerståhl (2010), let $\boldsymbol{E} = (E, A, \Sigma)$ be a grammar consisting of linguistic expressions $E$ generated from syntactic operations (rules) in $\Sigma \subset (E^* \rightarrow E)$ from a set of linguistic atoms $A \subset E$, and let $\mu$ be a function that assigns a semantic interpretation to an expression $u \in E$. For example, $A$ can be the set of English words, $\text{cat}, \text{cute}, \text{fluffy} \in A$, and those words can be combined using grammar rules such as $\texttt{attribute} \in \Sigma$ to form all English expressions $\texttt{attribute}(\text{fluffy}, \text{cat}) \in E$ and $\texttt{attribute}(\text{cute}, \texttt{attribute}(\text{fluffy}, \text{cat})) \in E$.

The function version of compositionality can then be formulated as:

**Definition 1** (Function version)**.** For every rule $\alpha \in \Sigma$ there is a meaning operation $r_\alpha$ such that if $\alpha(u_1, \dots, u_n)$ has meaning, $\mu(\alpha(u_1, \dots, u_n)) = r_\alpha(\mu(u_1), \dots, \mu(u_n))$.

In other words, the meaning ($\mu(\cdot)$) of a compound expression ($\alpha(u_1, \dots, u_n)$) can be derived from the meaning of its components ($\mu(u_1), \dots, \mu(u_n)$) and the way they are composed ($r_\alpha$). Using the same notation, the substitution version can be formulated as:

**Definition 2** (Substitution version)**.** If $s[u_1, \dots, u_n]$ and $s\,[t_1, \dots, t_n]$ are meaningful terms, and $\mu(u_i) = \mu(t_i)$ for $1 \leq i \leq n$, then $\mu(s\,[u_1, \dots, u_n]) = \mu(s\,[t_1, \dots, t_n])$,

where the notation $s\,[u_1, \dots, u_n]$ indicates that the expression $s \in E$ contains subexpressions $u_1, \dots, u_n \in E$.

As noted by Pagin and Westerståhl (2010), the substitution version is more general than the function version. If we assume that all subexpressions of meaningful expressions are also meaningful, then the two versions are strictly equivalent.

From these formalizations, we can derive some immediate consequences. Firstly, that if $\mu$ is the identity function, then the principle becomes trivially true for all inputs. A related point is famously made in Zadrozny (1994). This point is especially relevant for how compositionality is treated in the machine learning literature, which we will discuss in the next section. Secondly, that the principle is based on the premise that two expressions can be equivalent under the interpretation function $\mu$. In linguistic terms, this entails synonymy, and in the context of embedding models, where the model is $\mu$, it is hard to see how one would have $\mu(a) = \mu(b)$ for two strings $a, b$ without also having $a = b$, questioning the practical utility of the principle in the context of NLP. We further discuss these limitations in Section 3.

### 2.2 Compositionality in ML

In the machine learning literature, compositionality is often defined as the ability to decompose a function $\mu$ over simple components $(c_1, \dots, c_n) \in P_1 \times \dots \times P_n$ that entirely define the input $x = (c_1, \dots, c_n)$. Crucially, this decomposition into simple components is problem-dependent. In Wiedemer et al. (2023) for example, a function $\mu$ is compositional if it decomposes into two functions:

- $\varphi_i \colon P_i \to M_i$ mapping a component to its meaning,
- $\mathcal{C} \colon M_1 \times \dots \times M_n \to M$ combining component meanings into a meaning for the whole sample.

The intuition behind this is that $\mathcal{C}$ should be somewhat simple, but formally this is not enforced: If we let $\varphi$ be the identity function, we get a trivial compositional decomposition, which opens up the possibility of a compositional interpretation of *any* function. To circumvent this, Wiedemer et al. (2023) adapts the decomposition in a way that caters to their specific problem (predicting simple geometrical shapes on a grid), but the vacuousness of the general decomposition strategy remains. Ram et al. (2024) applies a similar approach but to natural language, defining $\mathcal{C}$ as a DAG over $\Sigma$ instead of the text sequence, but this definition can still be made vacuous using the identity as a component function and letting $\mathcal{C}$ be unrestricted.

Lippl and Stachenfeld (2025) constrains $\mu$ to be definable in term of a kernel invariant to – problem-specific – component-identity. This leads to a model restricted to a class of problems they refer to as *conjunction-wise additive tasks*. As they point out, these do not include some tasks that appear otherwise compositional such as transitive equivalence[2] and symbolic multiplication. Their kernel constraint would be difficult to measure for modern language models and appear too restrictive to capture the complexity of language.

These examples illustrate that an important aspect of the principle of compositionality relates to what precisely constitutes a *relevant part*, i.e. which parts of the input we compose. As pointed out by Pelletier (1994), the formalization should specify what constitutes a relevant part for a given problem to not be vacuous. On the other hand, if the specification is too problem-specific, it might not be generalizable to other problem settings.

### 2.3 Compositionality and Text Embeddings

Text embeddings models are language models trained to capture the semantic properties of text. As such, they parallel the $\mu$ operator in the formal compositional framework introduced in Section 2.1.[3] A core challenge in semantics is explaining how finite vocabularies generate meaning, and in computational semantics, the question is whether

[2] Transitive equivalence: $a \equiv b \wedge b \equiv c \implies a \equiv c$.

[3] Although text embeddings are rarely considered proper meaning representations in formal semantics, the principle of compositionality as specifically stated is agnostic to the choice of semantics.

these models mirror the compositional structure of human language processing (Fodor et al., 2025).

Idioms have historically been understood as a challenge to compositional theories of natural language semantics, where the meaning is determined more by convention than composition.[4] Recently, He et al. (2025) studied how well current embedding models capture idiomaticity, and they find that current models often rely on compositional strategies, even when presented with idiomatic language. As current models are predominately based on the distributional hypothesis (Harris, 1954; Sahlgren, 2008), which is thought to be well-suited for representing contextual signal, this finding might seem counter-intuitive. However, the observation that Transformers in particular struggle to accurately represent idiomatic language aligns with previous work on compositionality in machine translation (Dankers et al., 2022).

A shared thread in these works is their reliance on the principle of compositionality to explain human language understanding. However, the exact relationship between the principle and experimental settings targeted at compositionality is often unspecified. In He et al. (2025), for example, the principle is invoked by a reference to works by Richard Montague and Gottlob Frege. The formal definitions from these authors, however, are not explicitly included in the empirical work, making it unclear exactly how compositionality is supposed to be understood in that context. Fodor et al. (2025) refers to the functional definition of compositionality from Pelletier (2017), which is similar to Definitions 1 and 2, but it is unclear how the formalization is applied in their measurement of sentence similarity.

We argue that there is room for improvement with respect to how the principle of compositionality is used empirically in NLP. In the next section, we will address this limitation by deriving a measurement of compositionality from the formal framework presented in Section 2.1.

## 3 Operationalizing Compositionality

This section re-formalizes the principle of compositionality from formal semantics and examines its relationship to existing definitions in the literature. We then address the key challenges posed by current definitions, both in formal semantics and machine learning, and propose a continuous relaxation of the principle grounded in our formalization.

### 3.1 Re-formalization

Following the formalization of Pagin and Westerståhl (2010), let $A$ be a set of words and $E$ a way to compose those words into expressions $E \supset A$ from grammar rules $\Sigma$ such as $\alpha \colon E^* \to E$. Let $\mu \colon E \to M$ be a meaning assignment function, where $M$ is the space of possible meanings.

Using this notation, the principle of compositionality can be re-formulated as the inability to distinguish between two elements in $E$ when composed:

$$\forall a, b, c \in E : \forall \alpha \in \Sigma : \mu(a) = \mu(b) \implies \mu(\alpha(a, c)) = \mu(\alpha(b, c)). \quad (1)$$

This is equivalent to the substitution version of the principle of compositionality from Definition 2, and if we assume that subterms of meaningful terms are also meaningful (the domain principle), it is also equivalent to Definition 1. Assuming a single grammar rule $\Sigma = \{\langle \cdot, \cdot \rangle\}$, it is also equivalent to the definition of compositionality from Andreas (2019).

As argued by Janssen (1986), any formalization of the principle must derive the conditions that determine whether the meaning function at hand aligns with compositional semantics. In our case, this means that our formalization should be able to separate a compositional $\mu$ from a non-compositional $\mu$. We argue that Equation 1 provides an intuitive procedure for such a separation: if a meaning function $\mu$ assigns the same meaning to two expressions, then it must also assign the same meaning to these two expressions if they are composed with the same context.

### 3.2 Relaxation

The problem with the strict definition of compositionality we present in 3.1 is that it presupposes equality in the meaning space, which for natural languages entails the existence of perfect synonymy. This is common for all formal accounts of compositionality, as pointed out by Pagin and Westerståhl (2010). Perfect synonymy is a contentious premise, especially when evaluating embedding models, and as such, it is unclear whether this can be used in any practical sense given the equality constraint.

A solution to this issue is to do a continuous relaxation of Equation 1 so that compositionality is a matter of *degree*:

[4] A nice overview of how idioms and compositionality have been understood historically can be found in Dankers et al. (2022), together with references that approach broader linguistic issues related to idioms, such as lexicalization.

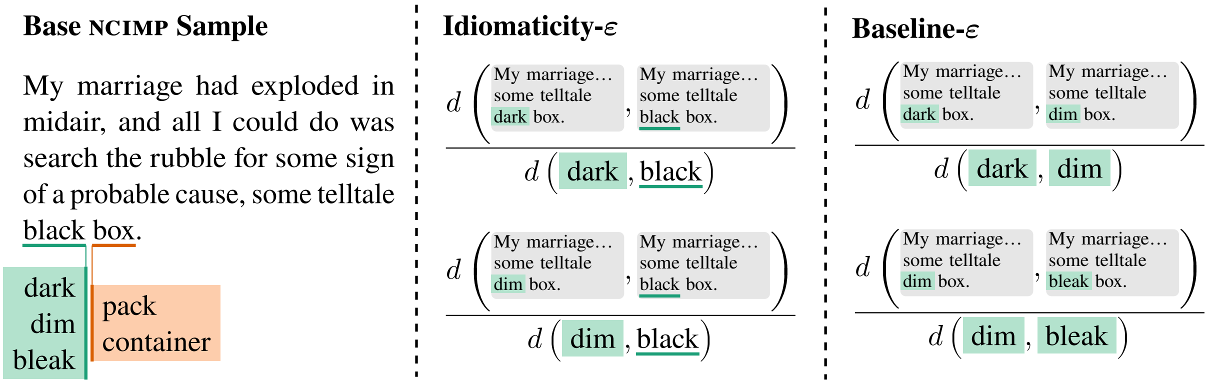
Base NCIMP Sample
My marriage had exploded in midair, and all I could do was search the rubble for some sign of a probable cause, some telltale black box.
dark
dim
bleak
pack
container
Idiomaticity-ε
d
My marriage… some telltale dark box.
My marriage… some telltale black box.
d
dark
black
d
My marriage… some telltale dim box.
My marriage… some telltale black box.
d
dim
black
Baseline-ε
d
My marriage… some telltale dark box.
My marriage… some telltale dim box.
d
dark
dim
d
My marriage… some telltale dim box.
My marriage… some telltale bleak box.
d
dim
bleak

Figure 1: A dataset sample from NCIMP (He et al., 2025) and how to compute $\varepsilon$ from it. For readability, we removed the $\mu(\cdot)$ operators and the $-1$ from Equation 3. Furthermore we only show our construction for synonyms of the modifier ("black"), but a similar procedure is applied to the head ("box").

**Definition 3** ($\varepsilon$-compositionality)**.** Given two words $a$ and $b$ and some context $c$, an interpretation function $\mu$ is $\varepsilon$-compositional if it assigns similarly distant meanings to $\alpha(a, c)$ and $\alpha(b, c)$, within some error $\varepsilon$:

$$d(\mu(\alpha(a,c)), \mu(\alpha(b,c))) \leq (1+\varepsilon) \times d(\mu(a), \mu(b)). \quad (2)$$

If $M$ is a vector space, $d$ could be any distance measure. The intuition behind this relaxation is that the distance in $M$ between two elements from $E$ should not increase if both components are composed with the same context via $\alpha$, within some error $\varepsilon$.

Definition 3 does not depend on there being a notion of perfect synonymy in $M$, nor does it require a problem-specific decomposition, while still maintaining a close resemblance with the traditional definitions from formal semantics. Indeed, Equation 1 is implied by Equation 2: if $a$ and $b$ are perfectly synonymous, both sides of the inequality in Equation 2 are 0, and $\alpha(a, c)$ must be perfectly synonymous with $\alpha(b, c)$.

## 4 Experiments

By gradually relaxing the principle of compositionality, we can quantify the degree to which different encoding functions ($\mu$) exhibit compositionality, as captured by the $\varepsilon$ value. In the next section, we outline our experimental framework, demonstrating how $\varepsilon$-compositionality can differentiate between embedding models. Our experiment specifically evaluates how well these models handle idiomatic multi-word expressions (MWEs), using $\varepsilon$ as a metric of their deviation from strict compositionality.

All code and data related to our experiments are made publicly available online.[5] We provide details on reproducing our results in Appendix F.

### 4.1 Data

In our experiments, we use the NCIMP dataset introduced by He et al. (2025), which is an extension to work by Garcia et al. (2021). The dataset consists of MWEs, each contextualized into three natural sentences. These are categorized according to human judgment into one of three classes: compositional (e.g. "winter solstice" which has a transparent meaning), partially compositional (e.g. "face value" which is a value but does not immediately relate to faces), and non-compositional (e.g. "hot potato" which is semantically opaque[6] ). All the MWEs are nominal compounds composed of two words, and each word is annotated with synonyms (e.g. "worth" and "price" are synonyms of "value"). An example from the dataset can be seen on the left of Figure 1, showing a sample for the idiomatic nominal compound "black box".

This dataset enables us to easily operationalize our definition of $\varepsilon$-compositionality. Specifically, given $\mu$ as text embedding model, we take $d$ to be the cosine distance in the embedding space produced by $\mu$. Note that despite not being a true distance, the cosine distance provides the best fit for $d$ as discussed in Appendix B. Data-wise, we take $b$ to be the original word ("black" in Figure 1), $a$ to be a synonym ("dark"), and $c$ to be the context sentence with the remaining compound word. From the composition of NCIMP as modifier–head com-

[5] https://github.com/ltgoslo/epsilon-compositionality

[6] "hot potato": an unwanted problem which is awkward to deal with.

pounds, we know that a syntactic operator $\alpha$ holds between those, its exact nature being irrelevant to the computation, we can safely ignore it. Thus, we can check whether Definition 3 holds and for which value of $\varepsilon$. We compute $\varepsilon$ as the lowest value that would satisfy the inequality in Equation 2:

$$\varepsilon(a, b, c) = \frac{d(\mu(ac), \mu(bc))}{d(\mu(a), \mu(b))} - 1, \qquad (3)$$

where $ac$ (respectively $bc$) denotes the infixing of $a$ (resp. $b$) in $c$. For each model $\mu$, we compute $\varepsilon$ over 275 nominal compounds embedded in 825 total sentences with 5.2 alternatives each on average.

### 4.2 Making Sense of $\varepsilon$

Our formalization posits that for a compositional encoding of semantics, two expressions that are close in meaning space should not be more distant in that same space when composed with the same context. In other words, given the embeddings of a set of phrases, inserting those phrases in the same larger context should result in the embeddings getting closer to each other. This should translate into a negative $\varepsilon$.

However, while the codomain of $\mu$ is voluntarily under-specified in formal semantics, identifying it with the embedding space of a language model comes with some challenges. In particular, the contextualized distance in the numerator of Equation 3 is not on the same scale as the phrase distance in the denominator. Indeed, the embedding of a full sentence $\mu(ac)$ is more dependent on the context $c$ than it is on a single word $a$. This is related to the fact that all embeddings are normalized the same way regardless of the complexity of the sentence. We refer to this phenomenon as *context-dominance*: $c$ is a possible confounder that makes it difficult to measure the influence of compositionality precisely.

To circumvent this problem, we always compare values of *idiomaticity-*$\varepsilon$ with values of *baseline-*$\varepsilon$ built from the same context as a form of regularization. He et al. (2025) approaches this by regularizing the similarities using randomly drawn samples from the overall dataset. Instead, we build a set of strong negatives from the set of synonyms in the dataset. In practice, we compute the baseline-$\varepsilon$ by comparing synonyms between themselves instead of comparing them to the compound expression. In particular, we align the computations of our idiomaticity-$\varepsilon$ with the computations of the baseline-$\varepsilon$ to have one phrase in common. This is shown in Figure 1, in the top row, the right $\varepsilon$ ("dark"–"dim") is a strong negative for the left $\varepsilon$ ("dark"–"black").

This gives us pairs of dependent $\varepsilon$-samples: one idiomatic and one baseline. If the models do not have a special treatment of MWEs, the samples should be independent and have the same $\varepsilon$ values, on the contrary if models break compositionality by assigning distant embeddings to MWEs, this should result in an idiomaticity-$\varepsilon$ larger than the baseline-$\varepsilon$. To test these hypotheses, we employ a (paired) Wilcoxon signed-rank test with a one-sided alternative: $\mathrm{H}_1$ : idiomaticity-$\varepsilon >$ baseline-$\varepsilon$. We report both the $p$-value and the effect size as measured by the rank-biserial correlation. While the $p$-value gives the probability of observing our set of $\varepsilon$ if we reject $\mathrm{H}_1$, the rank-biserial correlation measures how often one group ranks higher than the other.

### 4.3 Models

Our model selection is meant to represent different approaches to distributional encoding of semantics, while also being representative of models actually being used in applications. All models are loaded using either the `transformers` library or the `sentence-transformers` library (Reimers and Gurevych, 2019) for the sentence models. More details about the models, and their licenses, can be found in Appendices C and D.

#### 4.3.1 Plain Encoders

We evaluate three pre-trained encoder-based language models as baselines. To produce sentence embeddings from these models, we experiment with three different approaches to constructing a single representation for each text input. The most naive approach simply consists of taking the representation of the special CLS token. The second approach consists of averaging the representations of the subtokens using the four last layers, as done in He et al. (2025). Lastly, we experiment with summing the special CLS and SEP tokens.

**BERT** As our simplest embedding model, we use the original BERT model from (Devlin et al., 2019), resulting in text embeddings of size 768.

**ModernBERT** We also use the updated encoder model by Warner et al. (2024), which includes more modern optimization techniques. The produced text embeddings are of size 768.

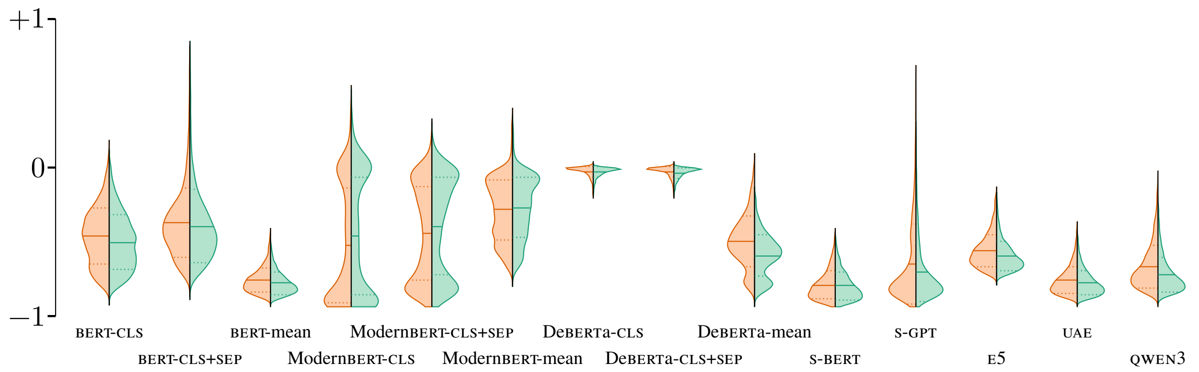

Figure 2: The distribution of idiomaticity-$\varepsilon$ (left, orange) and baseline-$\varepsilon$ (right, green) for all models (with a distance correction of $+0.05$ for visualization purposes). The means and unbiased standard deviations are shown in solid and dotted lines, respectively. This figure illustrates the need for the Wilcoxon signed-rank test as the two distributions depend too much on the context $c$ and are not directly comparable without taking that confounder into account.

**DeBERTa-v3** To include a model with a different style of attention mechanism (disentangled attention), we use the newest version of the model introduced in He et al. (2021). The produced text embeddings are of size 768.

### 4.3.2 Text Embedding Models

We also experiment with models fine-tuned specifically for producing text embeddings. We experiment with both encoder-based and autoregressive backbones, monolingual and multilingual, and instruction-tuned variants.

**S-BERT** As a baseline for this category, we use the original Sentence-BERT model from Reimers and Gurevych (2019), an embedding model that is wrapped around BERT and fine-tuned on contrastive sentence pairs. The embeddings are produced using the mean of the token embeddings, with a size of 768.

**S-GPT** Similarly to S-BERT, we also use a model based on the same approach but with an autoregressive language model as the backbone (Muennighoff, 2022). The embeddings are produced using a position-weighted mean pooling on the tokens, with embedding size 768.

**E5** We also use the multilingual E5 model from Wang et al. (2024), which extends the original English model from Wang et al. (2022). The embeddings are produced using mean pooling on the final output layer. We use the instruction-tuned version, following the official instructions for producing sentence embeddings of size 1 024.

**UAE** We use the angle-optimized model from Li and Li (2024). The model is trained both with contrastive training and with a special objective that minimizes the angle between the embeddings of text pairs. The model uses BERT as a backbone and produces embeddings of size 1 024 by taking the representation of the CLS token.

**QWEN3** Finally, we use a state-of-the-art embedding model based on the QWEN3 family of large language models (Zhang et al., 2025). This model is multilingual and instruction-tuned, and the embeddings are produced using mean pooling on the final output layer. We use the 600 million parameter version with an embedding size of 1 024. The rationale for not using a larger version is that it needs to be somewhat comparable to the simpler baselines, which are not available in larger versions.

## 5 Results

The distributions of $\varepsilon$ using the various models we selected for $\mu$ are shown in Figure 2. The violin plots show both the idiomaticity-$\varepsilon$ (in orange) and baseline-$\varepsilon$ (in green). When comparing them side-by-side, without taking into account the context $c$ confounder, it might appear that there are no significant differences between the two distributions. However, there are two cases we need to distinguish: for the same $a$ and $c$, either the idiomaticity-$\varepsilon$ and the baseline-$\varepsilon$ are randomly ordered, resulting in similar looking-distributions, or the idiomaticity-$\varepsilon$ is always slightly larger than the baseline-$\varepsilon$, resulting in distributions that might look similar because of large variance due to $c$, but are actually revealing of an non-compositional interpretation.

To resolve this ambiguity, we report the Wilcoxon signed-rank test statistics in Table 1. A $p$-value of 0 indicates that the observations are unlikely

| Model | | $p$-values | | | Rank-biserial (%) | | |
|---|---|---|---|---|---|---|---|
| | | C | PC | NC | C | PC | NC |
| BERT | CLS | 0.323 | 0.000 | 0.005 | 50.7 | 56.8 | 54.0 |
| | CLS + SEP | 0.789 | 0.022 | 0.137 | 48.8 | 53.1 | 51.7 |
| | mean | 0.731 | 0.000 | 0.042 | 49.1 | 58.9 | 52.7 |
| ModernBERT | CLS | 1.000 | 1.000 | 1.000 | 43.1 | 43.9 | 41.1 |
| | CLS + SEP | 0.999 | 1.000 | 1.000 | 45.4 | 40.7 | 39.9 |
| | mean | 0.988 | 0.964 | 1.000 | 46.7 | 47.2 | 44.4 |
| DeBERTa-v3 | CLS | 0.000 | 0.000 | 0.000 | 55.6 | 69.7 | 75.2 |
| | CLS + SEP | 0.000 | 0.000 | 0.000 | 56.8 | 71.2 | 76.2 |
| | mean | 0.000 | 0.000 | 0.000 | 80.8 | 91.6 | 94.6 |
| S-BERT | | 1.000 | 1.000 | 0.850 | 34.6 | 43.2 | 48.4 |
| S-GPT | | 0.100 | 0.000 | 0.000 | 51.9 | 56.3 | 56.2 |
| E5 | | 1.000 | 0.000 | 0.000 | 42.0 | 60.8 | 69.8 |
| UAE | | 1.000 | 0.000 | 0.000 | 44.4 | 60.8 | 61.3 |
| QWEN3 | | 0.000 | 0.000 | 0.000 | 62.2 | 72.6 | 78.8 |

Table 1: Statistics for the Wilcoxon signed-rank test with the alternative hypothesis $H_1$ : idiomaticity-$\varepsilon$ > baseline-$\varepsilon$. The results are reported for our three data classes: **C**ompositional, **P**artially **C**ompositional and **N**on-**C**ompositional.

if $H_1$ : idiomaticity-$\varepsilon$ > baseline-$\varepsilon$ was false, in other words, $H_1$ is certainly true. By construction, the rank-biserial is increasing when the $p$-value decreases. We report it to have interpretable results even when the $p$-value saturates. Accordingly, a high rank-biserial indicates that idiomaticity-$\varepsilon$ is often larger than the baseline-$\varepsilon$.

As a sanity check, we can confirm that $p$-values generally do not increase from compositional to non-compositional compounds, which is to be expected. Based on these results, we can categorize the models we studied into three categories.

**Human-aligned models** like E5, UAE and to a lesser extent BERT. For these models, compositional compounds (C) are treated compositionally: a $p$-value of 1 indicates that their $\varepsilon$ is not significantly larger than the baseline. And on the other hand, non-compositional (NC) and partially compositional (PC) compounds are treated non-compositionally: their $\varepsilon$ is significantly larger than the baseline, indicating that the contextualized embeddings are getting farther away from synonym-substituted contexts.

**Compositionally-inclined models** like ModernBERT and S-BERT. These models treat every MWE in a compositional manner. This might be indicative of poor performance on non-compositional compounds.

**Anti-compositional models** like DeBERTa, S-GPT and QWEN. These models treat every MWE in an non-compositional manner. This might be a symptom of overfitting on frequent MWEs, building an ad-hoc representation for them even when they could be derived through compositional means. However, when examining the effect size as reported by the rank-biserial, it is clear that these models still distinguish between compositional and non-compositional MWEs, with larger $\varepsilon$ differences in the NC class.

In general, the rank-biserial is correctly aligned C < PC < NC especially for more modern models. This matches our intuitions: Models are more compositional on compositional samples and less so on the non-compositional (meaning idiomatic) samples. The identification of PC, however, is not always so clear-cut. For additional insight on the matter, refer to Table 4 in the appendix, which contains experiments where only the head or the modifier of the noun compound is substituted for a synonym in an attempt to better identify the PC class.

Finally, we can observe that the aggregation methods for BERT, ModernBERT and DeBERTa-v3 do not affect our results substantially. This is despite the fact that the resulting embeddings have widely different distributions (as shown in Table 3 in appendix). Variations in embedding magnitude is a potential confounder, and indeed those variations are visible in Figure 2 when comparing the mean aggregation with CLS and CLS + SEP. This confounder seem to be correctly handled by our statistical test as demonstrated by Table 1: the aggregation methods change the scale of the embeddings but do not significantly change the $p$-value reported by the Wilcoxon signed-rank test.

## 6 Previous Work

$\varepsilon$-compositionality is related to other measurements targeted at MWEs, such as the *Similarity Score* in He et al. (2025). Existing work, such as He et al. (2025) and Fodor et al. (2025) focus on comparing model predictions to human ratings of compositionality, while we focus more on theoretical aspects. In our work, we derive a measurement of compositionality from formal definitions that allows us to tell whether a specific embedding model can differentiate between idiomatic and non-idiomatic examples, as opposed to measuring how these models align with human annotations on the sample level. While their work focuses on highlighting the current differences between human and artificial representations of compositional meaning, our work focuses on connecting formal accounts of compositionality with the evaluation of such models, which we hope can motivate the development of models that have a more aligned representational capacity.

Naturally, our work is also a continuation of an established line of work in NLP centered on compositionality in relation to MWEs (Bannard et al., 2003; Reddy et al., 2011; Salehi et al., 2015; Constant et al., 2017). However, we also see our work as related to adjacent research in compositionality and machine translation (Dankers et al., 2022; Li et al., 2021).

## 7 Conclusion

In summary, this paper examines the principle of compositionality in relation to text embedding models. We introduce $\varepsilon$-compositionality as a measurement for how text embedding models differentiate between idiomatic and non-idiomatic language, grounding it in definitions from formal semantics. By applying this framework to the NCIMP dataset, we find that most of the tested text embedding models are capable of distinguishing between different levels of compositionality, though the degree of effectiveness varies. Our work highlights the need for clearer integration between theoretical definitions of compositionality and empirical measurements, and we hope that our findings motivate future work on how compositionality is to be understood not only in the context of text embedding models but in relation to neural language models in general.

## Limitations

**Use of models** It can be argued that the fixed-size encoding methods used for our plain encoders may not be well-suited to the short sequences of the MWEs in our experiments. On the other hand, the goal of our work is not to provide any measurement of sentence representations in general, but rather to showcase how embedding models, as instances of interpretation functions, fit into a compositional framework. Consequently, we argue that our usage of these models is justified, as it says something about how sensitive these models are to semantic shifts, in our case idiomaticity.

**Dataset quality** While selecting a good illustrating example, we found indications that the NCIMP dataset has some issues related to data quality. For example, it contains ungrammatical substitutions (e.g. “an acid test” becoming “an sour test”), and synonymy is understood in a very broad sense (e.g. “apple” and “pear”).

**Implication** $\varepsilon$-compositionality can be used to measure if text embedding models are able to differentiate between compositional and non-compositional language use. However, we do not discuss whether this is an attractive property for a model to have. In practice handling non-compositional expressions in a compositional way may also enable the model to understand unfrozen MWEs (e.g. “it’s raining tigers and wolves” is an unfrozen (emphatic) variant of “it’s raining cats and dogs”).

**Confounders** Using a statistical test to compare to baseline-$\varepsilon$ allows us to control for most confounders. However, it is difficult to identify all potential ones, and some of the confounders we identified can remain problematic. For example, the idiomatic expressions tend to be more frequent than the baseline ones, so if frequency is a covariate, it is not compensated by the Wilcoxon signed-rank test.

**Non-idiomatic non-compositional expressions** While our intention is to study compositionality, the NCIMP dataset focuses on idiomatic expressions. Some non-idiomatic non-compositional expressions such as “come on” are therefore excluded from our experiments. On the other hand, it could be argued that idioms provide the maximum degree of non-compositionality and are therefore a good target for analysis.

## References

Jacob Andreas. 2019. Measuring compositionality in representation learning. In *International Conference on Learning Representations*.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan. Association for Computational Linguistics.

Marco Baroni, Raffaella Bernardi, Roberto Zamparelli, and 1 others. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in language technology*, 9:241–346.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626.

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.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, and 1 others. 2023. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2025. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics*, 51:505–555.

Wilfrid Hodges. 2001. Formal features of compositionality. *Journal of Logic, Language and Information*, 10(1):7–28.

Theo Janssen. 1986. *Foundations and applications of Montague grammar*. Ph.D. thesis, University of Amsterdam.

Sinan Kurtyigit, Diego Frassinelli, Carina Silberer, and Sabine Schulte Im Walde. 2025. A couch potato is not a potato on a couch: Prompting strategies, image generation, and compositionality prediction for noun compounds. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10766–10776, Vienna, Austria. Association for Computational Linguistics.

Xianming Li and Jing Li. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780, Online. Association for Computational Linguistics.

Samuel Lippl and Kim Stachenfeld. 2025. When does compositional structure yield compositional generalization? a kernel theory. In *The Thirteenth International Conference on Learning Representations*.

Kate McCurdy, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao. 2024. Toward compositional behavior in neural models: A survey of current views. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9323–9339, Miami, Florida, USA. Association for Computational Linguistics.

Richard Montague and 1 others. 1970. Universal grammar. *1974*, pages 222–46.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *Preprint*, arXiv:2202.08904.

Ryan M. Nefdt and Christopher Potts. 2024. *Compositionality*, chapter 1. MIT Press. Https://oecs.mit.edu/pub/e222wyjy.

Peter Pagin and Dag Westerståhl. 2010. Compositionality i: Definitions and variants. *Philosophy Compass*, 5(3):250–264.

Barbara Partee and 1 others. 1995. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

Francis Jeffry Pelletier. 1994. The principle of semantic compositionality. *Topoi*, 13(1):11–24.

Francis Jeffry Pelletier. 2017. Compositionality and concepts—a perspective from formal semantics and philosophy of language. In *Compositionality and concepts in linguistics and psychology*, pages 31–94. Springer International Publishing Cham.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Parikshit Ram, Tim Klinger, and Alexander G. Gray. 2024. What makes models compositional? a theoretical view: With supplement. *Preprint*, arXiv:2405.02350.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of linguistics*, 20:33–53.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. 2024. A survey on compositional learning of AI models: Theoretical and experimental practices. *Transactions on Machine Learning Research*. Survey Certification.

JFAK van Bethem, JAG Groenendijk, DHJ de Jong, MJB Stockhof, and HJ Verkuyl. 1991. *Logic, Language, and Meaning, Volume 2: Intensional Logic and Intensional Grammar*. University of Chicago Press, Chicago, IL.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. 2023. Compositional generalization from first principles. *Advances in Neural Information Processing Systems*, 36:6941–6960.

Wlodek Zadrozny. 1994. From compositional to systematic semantics. *Linguistics and philosophy*, 17(4):329–342.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

## A Dataset

We use the dataset files from the original repository provided in He et al. (2025). The original data files include samples in both English and Portuguese; we only use the English version of the dataset since the Portuguese version does not contain enough synonyms to calculate the baseline-$\varepsilon$. We make use of some of the human annotations (the three compositionality classes) in our work. The details of this data collection can be found in the original dataset publication.

Our preprocessing of the data is fully automated and documented in the software package bundled with this publication.

**Note on dataset selection** A naive approach to comparing idiomatic and non-idiomatic cases might involve using separate datasets for each. However, this introduces a confounder to the evaluation, as $c$ would differ between datasets, complicating direct comparison. To address this, we ensure comparability by holding $c$ constant, using the same sentence in both idiomatic and non-idiomatic contexts. This eliminates the need to normalize embeddings across datasets or adjust for variations in the length or semantic density of $c$. While this reduces the number of suitable datasets, we argue that it strengthens methodological validity. In this paper, our goal is to introduce $\varepsilon$-compositionality in a principled way.

## B Distance Definition

We use a cosine-based measure for $d$ to follow current established practices with transformer embeddings. As we want to extract a singular scalar value to capture the degree of compositionality, $\varepsilon$ can be naturally interpreted as the slope of the line characterizing how much the distance between two expressions $a, b \in E$ contracts when they are put in the same context $c \in E$. If we expressed Equation 3 in term of cosine similarities with $-d = \cos$, this would force orthogonal expressions ($\cos(\mu(a), \mu(b)) = 0$, i.e. dissimilar expressions) to stay orthogonal once contextualized. This does not follow from the unrelaxed principle of Equation 1, which concerns itself with similar (or identical) expressions, not dissimilar ones. Instead, following the unrelaxed principle, we would expect that if two expressions have identical embeddings (implying $\cos(\mu(a), \mu(b)) = 1$), then so does their contextualized embeddings. This is further justified by the fact that for all practical purposes, transformer models are injective:

$$\begin{aligned}\mu(a) = \mu(b) &\implies a = b \\ &\implies \alpha(a, c) = \alpha(b, c) \\ &\implies \mu(\alpha(a, c)) = \mu(\alpha(b, c)).\end{aligned}$$

As a result, instead of having the point at $\cos(\cdot) = 0$ be fixed, we would expect the point at $\cos(\cdot) = 1$ to be. This can be achieved by defining $d(a, b) = 1 - \cos(a, b)$. This function is commonly known as the cosine distance in the literature. However, it should be noted that this is a misnomer as the cosine distance is not a true distance: it verifies neither the Cauchy–Schwarz inequality nor the coincidence axiom. As the principle of compositionality relies solely on a notion of equality, only the later of which could pose problem. However, modern embedding spaces are normalized and in such a high dimension that the following can be safely assumed for all practical purposes:

$$\cos(\mu(a), \mu(b)) = 1 \implies a = b$$

This in turn, implies the coincidence axiom which is sufficient to capture the formal semantic definition of compositionality.

## C Models

Details about model keys, embedding sizes, and parameter sizes can be found in Table 2.

**Prompts for instruction-tuned models** For the E5 and QWEN3 models, we followed the official instructions for producing text embeddings, using the following prompt as the system task: *Given a sentence, encode its semantic properties.*

## D Licenses

We release our code under the APGL-3.0 license: https://github.com/ltgoslo/epsilon-compositionality.

The NCIMP dataset by He et al. (2025) does not include a license statement bundled with the software. Consequently, we assume that it follows the licensing of the publisher, in this case, MIT Press Direct, which is stated to be the Creative Commons Attribution 3.0 or 4.0 License (CC BY).

All models were used in accordance with their intended use, as specified in their respective licenses. The licenses for each model can be found in Table 2.

## E Use of AI assistants

An LLM-based service was used for minor paragraph-level editing and as a code assistant for a single file (`models.py`) for the experimental setup. The parts of the code that are in some way based on generation are clearly marked in the source code. We, as authors of this document, take full responsibility for the content of the paper and its results.

## F Reproduction

All the materials needed to reproduce our results can be found on `https://github.`

| Model | Model key | Emb.size | Params. | License |
|---|---|---|---|---|
| BERT | `bert-base-uncased` | 768 | ∼ 110 mill. | Apache-2 |
| ModernBERT | `answerdotai/ModernBERT-base` | 768 | ∼ 149 mill. | Apache-2 |
| DEBERTA-V3 | `microsoft/deberta-v3-base` | 768 | ∼ 86 mill. | MIT |
| S-BERT | `all-MiniLM-L6-v2` | 768 | ∼ 22 mill. | Apache-2 |
| S-GPT | `Muennighoff/SGPT-125M-weightedmean-nli-bitfit` | 768 | ∼ 125 mill. | None |
| E5 | `intfloat/multilingual-e5-large-instruct` | 1024 | ∼ 355 mill. | MIT |
| UAE$_{v1-large}$ | `WhereIsAI/UAE-Large-V1` | 1024 | ∼ 326 mill. | MIT |
| QWEN3 | `Qwen/Qwen3-Embedding-0.6B` | 1024 | ∼ 600 mill. | Apache-2 |

Table 2: Details about the models used in our experiments.

com/ltgoslo/epsilon-compositionality. Our software package contains an epsilon_compositionality Python module that can be run to reproduce all the numbers reported in this paper.

The experiments were conducted on a consumer-level laptop using the built-in CPU. Reproducing the results on a consumer-level CPU takes approximately 2 hours. Including failed experiments and development, we estimate the total compute time for the research to be around 50 CPU hours. As the dataset is relatively small and we are only considering inference over short sequences, there is no need for a GPU to reproduce our results.

## G Supplemental results

Table 3 shows the distribution of distances in our samples. We provide these statistics to give some sense of the scale that each of the respective embedding models operates on. As the distribution varies greatly between models, reporting a value of $\varepsilon$ without comparing to baseline would make little sense.

| Model | | Phrase $d(a, b)$ | | Context $d(ac, bc)$ | |
|---|---|---|---|---|---|
| | | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| BERT | CLS | 0.085 | 0.062 | 0.010 | 0.011 |
| | CLS + SEP | 0.073 | 0.042 | 0.021 | 0.029 |
| | mean | 0.259 | 0.108 | 0.015 | 0.014 |
| ModernBERT | CLS | 0.230 | 0.267 | 0.005 | 0.006 |
| | CLS + SEP | 0.096 | 0.112 | 0.003 | 0.004 |
| | mean | 0.029 | 0.028 | 0.002 | 0.006 |
| DeBERTa-v3 | CLS | 0.002 | 0.003 | 0.000 | 0.000 |
| | CLS + SEP | 0.002 | 0.003 | 0.000 | 0.000 |
| | mean | 0.100 | 0.076 | 0.008 | 0.009 |
| S-BERT | | 0.604 | 0.170 | 0.081 | 0.069 |
| S-GPT | | 0.351 | 0.173 | 0.050 | 0.044 |
| E5 | | 0.104 | 0.037 | 0.012 | 0.009 |
| UAE | | 0.362 | 0.102 | 0.042 | 0.036 |
| QWEN3 | | 0.211 | 0.093 | 0.021 | 0.021 |

Table 3: Distance distribution in the dataset.

| Model | | C | | PC | | NC | |
|---|---|---|---|---|---|---|---|
| | | modifier | head | modifier | head | modifier | head |
| BERT | CLS | 50.2 | 51.1 | 58.5 | 55.2 | 52.3 | 55.4 |
| | CLS + SEP | 49.1 | 48.6 | 54.3 | 52.0 | 50.9 | 52.4 |
| | mean | 52.5 | 45.6 | 62.4 | 55.7 | 53.0 | 52.2 |
| ModernBERT | CLS | 43.3 | 43.0 | 43.9 | 44.0 | 44.3 | 38.3 |
| | CLS + SEP | 46.7 | 44.2 | 42.1 | 39.4 | 45.4 | 35.0 |
| | mean | 46.5 | 46.8 | 46.3 | 48.2 | 47.2 | 41.8 |
| DeBERTa-v3 | CLS | 54.5 | 56.6 | 67.4 | 71.7 | 76.8 | 73.7 |
| | CLS + SEP | 54.5 | 58.9 | 70.6 | 71.6 | 77.7 | 74.7 |
| | mean | 79.9 | 81.8 | 93.0 | 90.4 | 95.1 | 94.2 |
| S-BERT | | 34.4 | 35.0 | 43.8 | 42.5 | 54.5 | 41.7 |
| S-GPT | | 54.4 | 49.3 | 58.8 | 53.9 | 57.3 | 55.1 |
| E5 | | 47.6 | 36.4 | 67.9 | 54.2 | 74.1 | 65.0 |
| UAE | | 44.8 | 43.8 | 63.4 | 58.3 | 65.8 | 56.1 |
| QWEN3 | | 67.7 | 56.6 | 77.7 | 67.2 | 80.9 | 76.5 |

Table 4: Rank-biserial distinguishing modifier and head substitutions.