| Idiom | Matched phrase | Syntactic pattern | Log frequency |
|---|---|---|---|
| Devil's advocate | Baker's town | JJ/dep/2 NN/pobj/0 | 2.398 |
| Act of darkness | Abandonment of institution | NN/dobj/0 IN/prep/1 NN/pobj/2 | 4.304 |
| School of hard knocks | Field of social studies | NN/pobj/0 IN/prep/1 JJ/amod/4 NNS/pobj/2 | 6.690 |

Table 1: Examples of idioms with their matched phrases, selected based on having the same syntactic pattern and most similar log frequency in the Syntactic Ngrams dataset. Examples depicted here have the same log frequency. Note that the frequency is based on the most common dependency and constituency pattern found in Syntactic NGrams. Humans were asked to rate each phrase for its compositionality.

remaining 10% were divided into a test set (5%) and dev set (5%).[5]

To fairly compare probes, we used minimum description length probing (Voita and Titov, 2020).This approximates the length of the online code needed to transmit both the model and data, which is related to the area under the learning curve. Specifically, we recorded average cosine similarity of the predicted vector and actual vector on the test set while varying the size of the training set from 0.005% to 100% of the original.[6] We compare the AUC of each probe under these conditions to select the most parsimonious approximation for each model.

## 4.2 Results

We find that **affine probes** are best able to capture the composition of phrase embeddings from their left and right subphrases. A depiction of probe performance at approximating representations across models and representation types is in Figure 2. However, we note that scores for most models are very high, due to the anisotropy phenomenon. This describes the tendency for most embeddings from pretrained language models to be clustered in a narrow cone, rather than distributed evenly in all directions (Li et al., 2020; Ethayarajh, 2019). We note that it is true for both word and phrase embeddings.

Since we are comparing the probes to each other relative to the same anisotropic vectors, this is not necessarily a problem. However, in order to compare each probe's performance compared to chance, we correct for anisotropy using a control task. This task is using the trained probe to predict a random phrase embedding from the set of treebank phrase embeddings for that model, and recording the distance between the compositional probe's prediction and the random embedding. This allows us to calculate an error ratio $\frac{\text{dist}_{\text{probe}}}{\text{dist}_{\text{control}}}$, where $\text{dist}_{probe}$ represents the original average distance from the true representation, and $\text{dist}_{\text{control}}$ is the average distance on the control task. This quantifies how much the probe improves over a random baseline that takes anisotropy into account, where a smaller value is better. These results can be found in Appendix E. The results without anisotropy correction can be found in Appendix G. In most cases, the affine probe still performs the best, so we continue to use it for consistency on all the model and representation types.

We also compare the AUC of training curves for each probe and find that the affine probe remains the best in most cases, except $\text{RoBERTa}_{\text{CLS}}$ and $\text{DeBERTa}_{\text{CLS}}$. Training curves are depicted in Appendix C. AUC values are listed in Appendix H.

Interestingly, there was a trend of the right child being weighted more heavily than the left child, and each model/representation type combination had its own characteristic ratio of the left child to the right child. For instance, in BERT, the weight on the left child was 12, whereas it was 20 for the right child.

For example, the approximation for the phrase "green eggs and ham" with BERT [CLS] embeddings would be: $r_{CLS}(\text{"green eggs and ham"}) = 12r_{CLS}(\text{"green eggs"}) + 20r_{CLS}(\text{"and ham"}) + \beta$.

[5] The learned probes were trained with early stopping on the dev set with a patience of 2 epochs, up to a maximum of 20 epochs. The Adam optimizer was used, with a batch size of 512 and learning rate of 0.512.

[6] We look at milestones of 0.005%, 0.01%, 0.1%, 0.5%, 1%, 10% and 100% specifically. This was because initial experimentation showed that probes tended to converge at or before 10% of the training data. Models were trained separately (with the same seed and initialization) for each percentage of the training data, and trained until convergence for each data percentage condition.

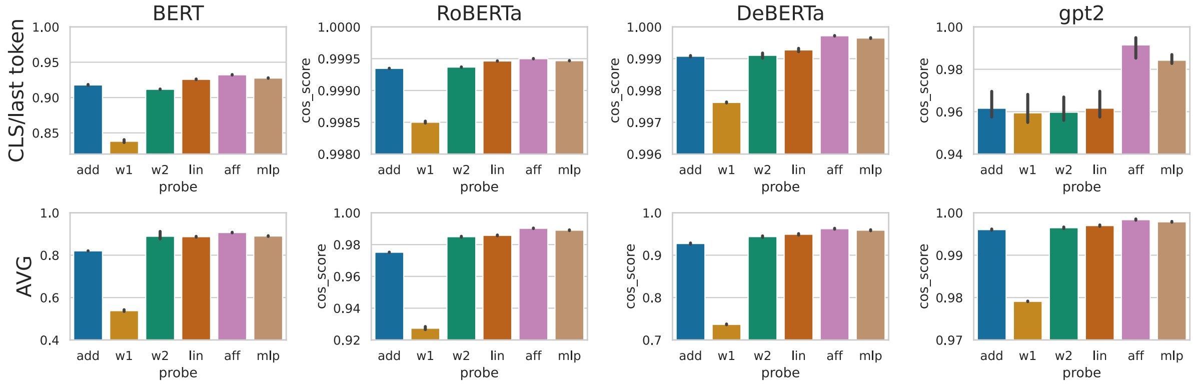

Figure 2: Mean compositionality score (cosine similarity) and standard deviation of each approximative probe across 10 folds. Error bar indicates 95% CI.

## 5 Examining Compositionality across Phrase Types

### 5.1 Methods

Intuitively, we expect the phrases whose representations are close to their predicted representation to be more compositional. We call similarity to the expected representation, $\text{sim}(r(x), \hat{f}(r(a), r(b)))$, the *compositionality score* of a phrase.

We record the mean reconstruction error for each tree type and report the results. In addition to comparing tree types to each other, we also examine the treatment of named entities in subsubsection 5.2.1. We examine the relationship between length of a phrase in words and its compositionality score in subsubsection 5.2.2.

### 5.2 Results

There is a significant difference between the mean compositionality score of phrase types. Particularly, the **AVG** representation assigns a lower compositionality score to NP → NNP NNP phrases, which is expected since this phrase type often corresponds to named entities. By contrast, the **CLS** representation assigns a low compositionality score to NP → DT NN, which is unexpected given that such phrases are generally seen as compositional. The reconstruction error for the most common phrase types is shown in Figure 5.

Because different phrase types may be treated differently by the model, we examine the relative compositionality of phrases within each phrase type. Examples of the most and least compositional phrases from several phrase types are shown in Table 2 for RoBERTa$_{\text{CLS}}$. Patterns vary for model and representation types, but long phrases are generally represented more compositionally.

#### 5.2.1 Named Entities

We used SpaCy to tag and examine named entities (Honnibal and Montani, 2017), as they are expected to be less compositional. We find that named entities indeed have a lower compositionality score in all cases except RoBERTa$_{\text{CLS}}$, indicating that they are correctly represented as less compositional. A representative example is shown in Figure 3. Full results can be found in Appendix J. We break down the compositionality scores of named entities by type and find surprising variation within categories of named entities. For numerical examples, this often depends on the unit used. For example, in RoBERTa$_{\text{AVG}}$ representations, numbers with "million" and "billion" are grouped together as compositional, whereas numbers with quantifiers ("about", "more than", "some") are grouped together as not compositional. The compositionality score distributions for types of named entities are presented in Figure 4.

#### 5.2.2 Examining Compositionality and Phrase Length

There is no consistent relationship between phrase length and compositionality score across models and representation types. However, **CLS** and **AVG** representations show divergent trends. There is a strong positive correlation between phrase length and compositionality score in the **AVG** representations, while no consistent trend exists for the **CLS** representations. This indicates that longer phrases are better approximated as an affine transformation of their subphrase representations. This trend is summarized in Appendix D. All correlations are highly significant.

| Phrase type | Most compositional | Least compositional |
|---|---|---|
| PP → IN NP | ("of", "two perilous day spent among the planters of Attakapas, . . .)<br>("of", "the cloth bandoleers that marked the upper part of his body . . .) | ("of", "September")<br>("like", "the Standard & Poor 's 500") |
| S → NP-SBJ VP | ("him", "to suggest it's the difference between the 'breakup' value . . .)<br>("it", "was doing a brisk business in computer power-surge protectors . . .") | ("other things", "being more equal")<br>("less", "is more") |
| NP → NNP NNP | ("M.", "Bluthenzweig")<br>("Dr.", "Volgelstein") | ("Edward", "Thompson")<br>("Alexander", "Hamilton") |

Table 2: Phrases rated most and least compositional using RoBERTa$_{\text{CLS}}$ representations, from several syntactic phrase types. ". . ." indicates that a phrase continues but is too long to display. Long phrases and abbreviated names tend to have a higher compositionality score.

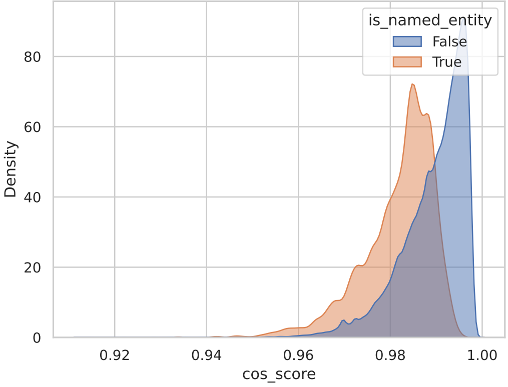

Figure 3: Density plot for compositionality scores of named entities and non-named-entities with RoBERTa$_{\text{AVG}}$ representations. Higher means more compositional.

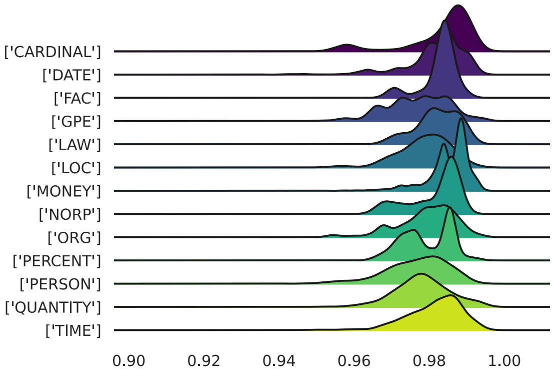

Figure 4: Density plots for compositionality scores of different named entity types with RoBERTa$_{\text{AVG}}$ representations. Higher means more compositional.

## 6 Comparing Compositionality Judgments of Humans and Models

### 6.1 Methods

#### 6.1.1 Human Annotation

Human annotators assigned labels to each phrase in the matched dataset from subsection 3.2: 1 for not compositional, 2 for somewhat compositional, and 3 for fully compositional. They could also decline to answer if they felt that the phrase didn't make sense on its own. Furthermore, they were asked how much each subphrase (left and right) contributed to the final meaning, from 1 for not at all, to 3 for a great deal. The Likert scale of 1-3 was chosen based on analysis of previous compositionality annotation tasks, which found that extreme values of compositionality were the most reliable (Ramisch et al., 2016a).

Initially, six English-speaking graduate students were recruited. The six initial annotators all annotated the first 101 examples and the subset of three annotators with the highest agreement who agreed to continue (Krippendorff $\alpha = 0.5750$) were recruited for the full study, annotating 1001 examples. For the full study, the agreement was higher ($\alpha = 0.6633$). We took the mean of compositionality judgments to be the final score for phrases. The instructions shown to annotators are in Appendix F. Examples judgments from an annotator can be found in Table 3.