STRUCTURE RETENTION IN EMBEDDING SPACES CORRELATES WITH BENCHMARK PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we show that high-performing embedding models organize their embedding spaces in a consistent way. We evaluate 25 contemporary embedding models on five MTEB tasks spanning four diverse task categories (retrieval, bitext mining, pair classification, and summarization) in both English and multilingual settings, and reveal that nearest-neighbor overlap and magnitude differences in independent component analysis (ICA) between paired text instances strongly correlate (even up to 0.97) with performance on the given task. Ultimately, we show that embedding tasks display varying degrees of linearity and reliance on retention of local information. Our results further the understanding of embeddings and their relation to model performance beyond cosine similarity.

1 INTRODUCTION

Embedding language models, or embedding models, are language models optimized to produce numerical representations of text for a variety of tasks, such as clustering, retrieval, and classification (Xie et al., 2016; Huang et al., 2020; Wang et al., 2016). Because embeddings are a core component of many current NLP applications, evaluating them—along with studying their structure and optimization methods—is essential. To date, evaluation of embedding models has primarily focused on downstream performance: models are assessed by embedding large datasets, which is both time-consuming and requires large amounts of compute. The application-focused perspective has also left comparatively less attention for understanding *why* one embedding model outperforms another. This gap in understanding which properties are beneficial or detrimental limits our ability to design effective training objectives, select models that generalize beyond test sets, analyze failure cases, and develop new optimization approaches, such as condition-aware embeddings.

We hypothesize that a model's suitability for a task is reflected in how it organizes its embedding space. Ideally, task-relevant information is presented systematically, or figuratively on a *silver platter*: for example, translation should benefit from placing parallel sentences in different languages in analogous configurations, in which case moving between languages would correspond to a relatively simple mapping in the space. We apply the same principle to a broader set of embedding tasks. Rather than comparing embeddings *across* models, we analyze each model's embedding space in isolation by exploiting the paired structure of common embedding evaluation tasks (e.g., question-answer pairs). This allows us to study what drives task success, rather than only reporting an aggregate evaluation score. Specifically, we investigate the degree to which query and target sides of common embedding benchmark datasets align with each other using kNN-overlap and a linear method based on independent component analysis (ICA) as probes. We uncover strong relationships between embedding space geometry and benchmark performance, and our results provide evidence that these properties can help explain performance differences across tasks. Our contributions are:

Finding 1 We show that high k-Nearest Neighbors overlap between queries and targets strongly correlates (even up to 0.97) with performance on the given task. This suggests that strong embedding models organize paired texts in a mutually consistent way in their embedding spaces.

Finding 2 We introduce the magnitude difference in ICA components ("*peaks*") as a proxy for embedding model performance and show that it can be used to analyze the linearity of task information.

Finding 3 We tie these measures together by showing the differences in structure they can uncover and validate our results using the interpretable features of ICA.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Background and Motivation

Language model representational spaces, such as embeddings, outputs before next-token prediction, or internal states, have been extensively studied in prior work. One recurring theme is *linearity*: the idea that concepts can be represented as linear combinations of others. This observation traces back to early word embeddings (Mikolov et al., 2013) and has since been examined in transformer models (Nanda et al., 2023; Gurnee & Tegmark, 2024; Park et al., 2024), with evidence that high-level concepts are often encoded linearly. Consistent with this, embeddings have been successfully analyzed with linear methods, particularly Independent Component Analysis (ICA) (Hyvärinen, 1999) in, e.g., Yamagiwa et al. (2023); Li et al. (2024); Musil & Mareček (2024). Another emerging approach for analyzing representations is the superposition viewpoint (Elhage et al., 2022), which argues that concepts can be encoded more efficiently as nearly orthogonal rather than fully orthogonal representations; notable examples include Cunningham et al. (2023); Gurnee et al. (2023); Chen et al. (2025). Finally, Park et al. (2025) report that high-level concepts (e.g., language, gender, or animacy) exhibit linearity, whereas lower-level concepts are encoded in a more complex structure.

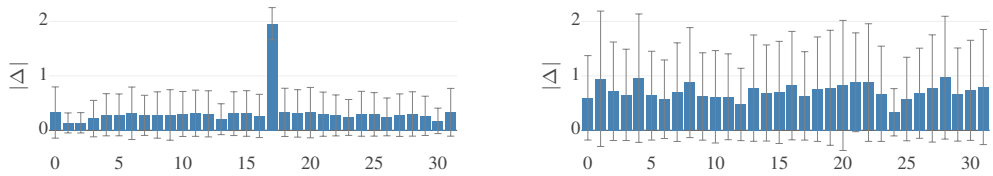
Several studies show that different embedding models share substantial geometric structure and that their spaces can be aligned with relatively simple mappings. Ren et al. (2023) argue that embedding spaces of differently initialized models are isomorphic up to some tolerance. Yoon & Arik (2025) propose the Embedding Converter, a framework for mapping one embedding space to another. Lin & Smith (2019) compare models via a nearest neighbor overlap score to quantify shared local structure. In addition, Wang et al. (2022a) connect embedding space structure to evaluation scores by analyzing exclusively semantic text similarity (STS) tasks and cosine similarity.

Our motivation is as follows: if modern embedding models exhibit structures described above, then a simple linear method such as ICA should be able to isolate a dominant concept-level signal (e.g., language) in their embedding spaces. To illustrate the intuition, consider a dataset of N translation pairs $D = \big[(t_1, t'_1), (t_2, t'_2), \ldots, (t_N, t'_N)\big]$ from English texts t_i (queries) to French t'_i (targets). Let $f(t) \in \mathbb{R}^d$ denote the embedding of t. We fit ICA on the embeddings of both sides of the pairs,

$$\text{ICA} \xleftarrow{\text{fit}} \big\{f(t_1), \ldots, f(t_N), f(t'_1), \ldots, f(t'_N)\big\}.$$

If language is, as assumed, the primary systematic difference between queries and targets and is represented approximately linearly, ICA should allocate a dedicated component to this signal. Consequently, when we transform both embeddings and take the element-wise difference for each pair, $\text{ICA}\big(f(t_i)\big) - \text{ICA}\big(f(t'_i)\big)$, most of the magnitude of this difference should concentrate in a small number of ICA dimensions. When we plot the mean absolute difference across query and target pairs for each ICA dimension, this concentration should appear as a pronounced "*peak*". Figure 1 shows this effect for two models, multilingual-e5-large-instruct and embedding-gemma-300m. A peak is clearly observable for multilingual-e5-large-instruct, indicating that ICA isolates a language-specific direction in its output space, and the model coincidentally receives a high score on an English-French task, whereas embedding-gemma-300m does not produce such a peak and receives a lower score. We show that the **prominence of the peak is descriptive of the performance of the model**.

However, not all concepts are represented linearly in representational spaces (e.g., Park et al. (2025)). Therefore, we also consider a more general notion of structural similarity between queries and tar-



(a) Intfloat/multilingual-e5-large-instruct (b) Google/embedding-gemma-300m

Figure 1: Average absolute difference ($|\Delta|$) over English-French translation pairs on ICA (dim=32) transformed embeddings per output dimension, with standard deviations as error bars. Multilingual-e5-large instruct, which receives high scores on English-French tasks, shows a characteristic "peak", while embedding-gemma-300 does not, and correspondingly receives lower scores.

gets. Rather than explicitly learning a mapping from $f(t_i)$ to $f(t'_i)$, we probe it directly using the query–target structure of embedding evaluation datasets. Concretely, we compute the *neighborhood retention rate*: the fraction of k-nearest neighbors shared between $f(t_i)$ and $f(t'_i)$, and find strong evidence that **models with higher neighborhood retention also achieve higher evaluation scores**.

In this paper, we widen the scope from the exemplified translation task and target five diverse tasks from four different task categories, showing that the above-described relations between embedding space properties and evaluation scores hold for 25 contemporary embedding models: across these settings, we establish that the two described structural measures correlate strongly with benchmark scores. We further analyze how these measures relate to each other, use ICA to characterize task-specific linear structure, and leverage the interpretability of ICA components to validate the findings.

3 METHODS

3.1 DATASETS

The datasets used in this study are ARCChallenge (Clark et al., 2018), WebFAQ (Dinzinger et al., 2025), Tatoeba (Tatoeba community, 2021), RTE3 (Giampiccolo et al., 2007), and SummEval (Fabbri et al., 2021). A summary of these datasets is presented in Table 1, and the languages included in multilingual tasks are shown in the result tables (e.g., Table 2) as ISO language codes. The datasets cover four task types in the MTEB benchmark and were selected because they contain well-defined paired text instances. For example, ARCChallenge contains question–answer pairs, SummEval document–summary pairs; we refer to these partitions as queries and targets, respectively. We use the test/development sets and give further dataset construction details in Appendix A.1, such as the selection of contradicting passages for RTE3, and discuss data contamination in Appendix A.2.3.

3.2 EMBEDDING MODELS

The 25 embedding models used in this study are listed in Appendix A.2.2. We select models that vary along several dimensions to ensure the soundness of our experiments: (1) scores on the MTEB multilingual leaderboard (spanning a wide range); (2) model developer (covering a diverse set, with a few models chosen per family); (3) model accessibility (primarily open-weight models, with a few prominent closed-weight models); and (4) language scope (including both English-only and multilingual models). This selection helps ensure that observed differences reflect actual model capability variation rather than only confounding factors. Additionally, most contemporary embedding models are intended to be used with a prompt, a short instruction added to the query to help steer the embedding towards the desired output. We therefore run our main experiments both with and without prompts.The prompts, the prompt template, and additional model information are presented in Appendix A.2. We use the `sentence-transformers` (Reimers & Gurevych, 2019) library.

3.3 MTEB EVALUATION

We use the evaluation scores from the official MTEB repository (accessed 12th Feb. 2026). However, this repository only contains language-aggregated score for some tasks, in which case we evaluate the tasks ourselves with the `mteb`-library (Muennighoff et al., 2023; Enevoldsen et al., 2025) with default settings. Unfortunately, the scores for WebFAQ are not available for closed-source models, nor is it feasible to evaluate this task on them due to the large number of examples (>100k per language). Similarly, evaluating WebFAQ on the largest models using `mteb`-library would require computational resources beyond our reach, so for this specific task, we report the results for a truncated set of models, with the omitted models shown in Table 6.

3.4 NEIGHBORHOOD RETENTION

We evaluate the neighborhood retention between the query and target (t, t') as

$$\text{Ret}_k = \sum_{(t,t') \in D} \frac{\text{Jacc}\left(k\text{NN}(f(t)), k\text{NN}(f(t'))\right)}{k}$$

Table 1: Datasets used in our study. N denotes the number of paired instances.

Dataset name	MTEB category	Pair difference	Language(s)	N
ARCChallenge	Retrieval	Question–Answer	English	937
WebFAQ	Retrieval	Question–Answer	13 languages	5000
Tatoeba	Bitext Mining	English–Non-English	10 languages	5000
RTE3	Pair Classification	Premise–Contradiction	4 languages	90
SummEval	Summarisation	Document–Summary	English	80

where $k\text{NN}(f(t))$ is the set of k nearest neighbors for the embedding of text t and $\text{Jacc}(\cdot, \cdot)$ is the Jaccard-overlap. This quantity is inspired by the local training loss of Yoon & Arik (2025), which they use to preserve the local structure in their Embedding-Converter, and the N2O metric used by Lin & Smith (2019). We roughly optimize k with a grid search covering $k = \{5, 10, 20, 50\} \cup \{0.01N, 0.02N, 0.05N, 0.1N, 0.2N\}$ where N is the number of paired instance in the dataset D, accounting for diversity in dataset sizes. We report the results for $k = 10$, which was found to be in the top 3 values for all datasets. It is to be noted that while the value of k has an impact on the results, **all tested values of k support our conclusions** (see Appendix B.2).

3.5 Independent Component Analysis and Peak Prominence

Independent component analysis (Hyvärinen, 1999) is a linear blind source separation algorithm. A more detailed explanation of ICA is presented in Appendix A.3. We use ICA in this study for three reasons: we aim to evaluate linearity, it offers straightforward interpretability, and fitting the required number of ICA models for our experiments is substantially less costly than training the same number of other interpretable models, such as Sparse Autoencoders (SAEs). We use the FastICA algorithm from the `sklearn`-library (Pedregosa et al., 2011) with default parameters, except for the maximum number of iterations, which we set to 10,000 to ensure convergence.

In our experiments, we fit the ICA models, subtract paired instances, and analyze the prominence of the characteristic peak with the Gini coefficient: a statistic measuring the inequality of a distribution, 0 indicating a uniform distribution, and 1 a totally unequal distribution. Using the Gini coefficient was motivated by the level of agreement between Gini values and visual inspection of peaks, and its stability against other suitable metrics (e.g., peak-to-mean ratio), as evaluated in Appendix B.3. While ICA is developed for settings where the number of sources is known upfront, it has been successfully applied to embeddings with a smaller number of assumed sources d_{ICA} (Huertas-García et al., 2023). In preliminary experiments, we observed that large values of d_{ICA} produced no peaks or failed to converge, while smaller values led to a clear signal, which we attribute to the homogeneity and small size of our datasets. We therefore set $d_{\text{ICA}} \ll d_{\text{emb}}$ and compare $\{32, 64, 128\}$ and PCA-estimated number of dimensions accounting 50% and 90% of explained variance. The clearest peaks occur at 32 and 64 dimensions, so we report all results for ICA^{32} (and ICA^{64} in Appendix B.1).

4 Experiments

We divide our experiments into two main experiments examining the relationship between embedding space structure and benchmark scores, and supporting experiments that provide interpretability and validate the main results. We detail the computational cost of our experiments in Appendix A.4.

In our first main experiment, we compute the neighborhood retention score Ret_k for all model-dataset combinations and correlate it with MTEB performance. This tests whether organizing the embedding space locally similarly on the query and target sides is beneficial for solving the task. In the second main experiment, we fit an ICA model on the embeddings, transform both sides of each pair, and compute the element-wise mean absolute difference between paired transformed embeddings. This produces a distribution over ICA dimensions with a varying "peak" (see Figure 1). We quantify peak prominence using the Gini coefficient and correlate it with MTEB performance. This experiment tests whether the query–target difference concentrates in a small number of independent directions, consistent with task-relevant information being encoded in an approximately linear

form, and whether this property is beneficial for model performance. For both main experiments, we report Spearman correlation coefficients, since in each case we compare two rankings.

To support the main findings, we further examine how neighborhood retention relates to ICA by interpreting the ICA components in terms of local (pair-specific) and global (dataset-level) signals shared between queries and targets. Additionally, in the Appendix, we validate our approach by (1) examining the stability of the fitted ICA models across runs with different random initializations (Appendix B.3); and (2) extracting the most salient words for each ICA dimension to verify that the observed peak corresponds to interpretable linguistic content rather than a spurious effect (Appendix B.4). All supporting experiments are described in more detail in their respective sections.

To validate that the results we obtain from the above experiments do not simply reflect the underlying cosine similarity, we also investigate 5 different levels of measuring cosine similarity and correlate them to model performance similarly to our first experiment described above. We explore raw, margin, max-margin, ratio, and z-normalized cosine similarity to reflect different levels of structural awareness in measuring cosine similarity, and control, in each direction, with our Ret_k method. We detail this experiment in Appendix B.6.

5 RESULTS

Neighborhood Retention vs. Performance The results for neighborhood retention against MTEB-score are presented in Table 2. Figure 2 presents an example of the relationship between retention and score. Overall, neighborhood retention correlates strongly with MTEB performance on three of the five datasets and moderately on one, with significance of $p < 0.001$ in almost all cases where a strong relationship is found. This suggests that embedding models that preserve local neighborhood structure between paired inputs for a given task tend to perform well on that task, whereas models with less consistent structure tend to perform worse. For Tatoeba, the relationship is close to perfect (≥ 0.9) with only a few exceptions. ARCChallenge and WebFAQ, both question-answering datasets, show similarly strong correlations (outside WebFAQ:eng, see Appendix Figure 9). SummEval shows a modest but significant correlation, whereas RTE3 shows little to no correlation. The effect of prompting varies by dataset: it is most beneficial for ARCChallenge, likely because our ARCChallenge prompt matches the one used in MTEB and therefore best reflects the intended evaluation setup. In contrast, prompting slightly reduces the correlation for other datasets.

Peak Prominence vs. Performance The results for Gini-coefficient of ICA^{32} (peak prominence) versus MTEB-score are presented in Table 3. Tatoeba shows correlations comparable to those ob-

Table 2: Spearman correlation (↑) for neighborhood retention with MTEB scores, with bolded values for strong relationship (≥ 0.7). Star ($*$) indicates significance at level $p < 0.05$, dagger (†) at $p < 0.01$, and double dagger (‡) at $p < 0.001$.

Dataset	Unprompted	Prompted	Dataset	Unprompted	Prompted
ARCChallenge	**0.890**‡	**0.945**‡	SummEval	0.643‡	0.588†
RTE3:eng	-0.230	-0.058	WebFAQ:bul	**0.926**‡	**0.866**‡
RTE3:fra	0.149	0.268	WebFAQ:deu	**0.889**‡	**0.856**‡
RTE3:ita	0.389	0.488*	WebFAQ:ell	**0.946**‡	**0.930**‡
RTE3:deu	0.025	0.162	WebFAQ:eng	0.620†	0.544*
Tatoeba:deu-eng	**0.888**‡	**0.904**‡	WebFAQ:fas	**0.934**‡	**0.893**‡
Tatoeba:bul-eng	**0.962**‡	**0.928**‡	WebFAQ:fin	**0.953**‡	**0.913**‡
Tatoeba:ukr-eng	**0.939**‡	**0.918**‡	WebFAQ:fra	**0.878**‡	**0.843**‡
Tatoeba:ita-eng	**0.917**‡	**0.884**‡	WebFAQ:hin	**0.909**‡	**0.891**‡
Tatoeba:fra-eng	**0.925**‡	**0.895**‡	WebFAQ:ita	**0.907**‡	**0.816**‡
Tatoeba:vie-eng	**0.945**‡	**0.916**‡	WebFAQ:jpn	**0.864**‡	**0.827**‡
Tatoeba:ell-eng	**0.968**‡	**0.922**‡	WebFAQ:ukr	**0.942**‡	**0.909**‡
Tatoeba:jpn-eng	**0.968**‡	**0.939**‡	WebFAQ:vie	**0.936**‡	**0.903**‡
Tatoeba:hin-eng	**0.931**‡	**0.876**‡	WebFAQ:zho	**0.891**‡	**0.841**‡
Tatoeba:fin-eng	**0.968**‡	**0.959**‡			

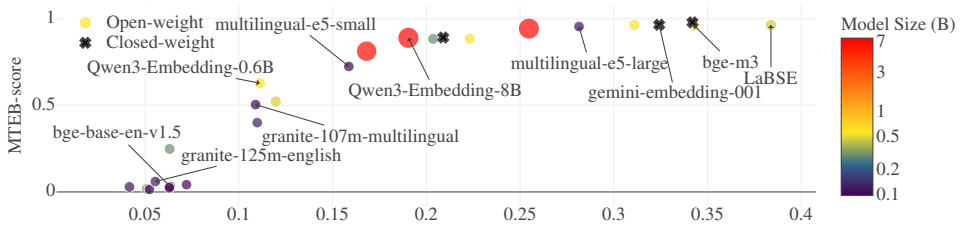


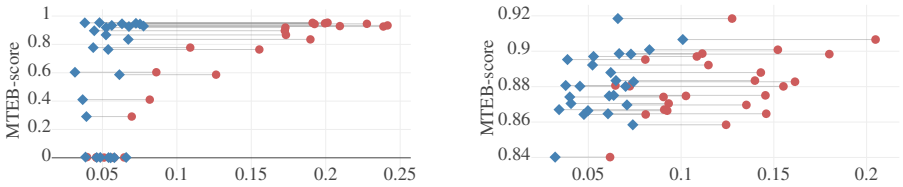
Figure 2: Visualization of the relationship between Ret_k (x-axis) and the MTEB-score for Tatoeba:fin-eng task, which shows one of the strongest correlations in Tatoeba (**0.968**‡). Annotated model names show that multilingual models outperform their English-only counterparts, and that model size is not the only contributing factor. Other examples in Appendix B.5.

Table 3: Spearman correlation (↑) for Gini-coefficient calculated from ICA^{32} with MTEB scores, with bolded values for strong relationship (≥ 0.7). Star ($*$) indicates significance at level $p < 0.05$, dagger (†) at level $p < 0.01$, and double dagger (‡) at level $p < 0.001$.

Dataset	Unprompted	Prompted	Dataset	Unprompted	Prompted
ARCChallenge	**0.810**‡	**0.840**‡	SummEval	0.178	0.428*
RTE3:eng	0.414*	0.318	WebFAQ:bul	0.286	0.278
RTE3:fra	**0.706**‡	0.542†	WebFAQ:deu	**0.756**‡	0.680†
RTE3:ita	0.652‡	0.541†	WebFAQ:ell	-0.106	0.284
RTE3:deu	0.525†	0.555†	WebFAQ:eng	0.255	0.542*
Tatoeba:deu-eng	**0.818**‡	**0.933**‡	WebFAQ:fas	-0.158	-0.133
Tatoeba:bul-eng	**0.861**‡	**0.928**‡	WebFAQ:fin	0.680†	**0.736**‡
Tatoeba:ukr-eng	**0.906**‡	**0.918**‡	WebFAQ:fra	**0.721**‡	0.622†
Tatoeba:ita-eng	**0.942**‡	**0.909**‡	WebFAQ:hin	0.152	0.183
Tatoeba:fra-eng	**0.796**‡	**0.859**‡	WebFAQ:ita	0.626†	0.564*
Tatoeba:vie-eng	**0.829**‡	**0.881**‡	WebFAQ:jpn	0.643†	0.614†
Tatoeba:ell-eng	**0.904**‡	**0.868**‡	WebFAQ:ukr	0.360	0.354
Tatoeba:jpn-eng	**0.853**‡	**0.866**‡	WebFAQ:vie	0.558*	0.657†
Tatoeba:hin-eng	**0.881**‡	**0.893**‡	WebFAQ:zho	0.257	0.422
Tatoeba:fin-eng	**0.945**‡	**0.928**‡			

served in the neighborhood retention experiment, indicating that bitext mining benefits from preserving the relative neighborhood organization between the two sides of each pair, and that a more linear encoding of translation-pair differences is characteristic of high-performing models. ARCChallenge also shows this connection to a lesser degree, and again benefits from prompting. The main differences between neighborhood retention and peak prominence emerge for RTE3 and WebFAQ, but in opposite directions: Gini yields a stronger relationship for RTE3, whereas correlations for WebFAQ are weaker than in the neighborhood retention results, even when they remain significant and moderately positive in half of the studied languages. The results for ICA^{64} are presented in Appendix B.1, and they support the same conclusions outside ARCChallenge prompting.

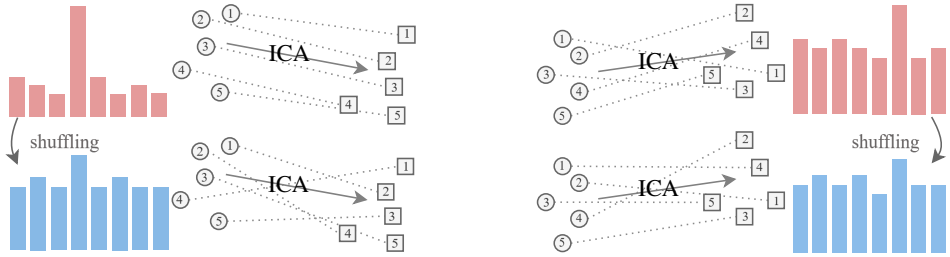
In summary, although the tasks differ in how strongly structure relates to performance, our results show that simple measures computed from paired embeddings can correlate strongly with benchmark scores. Comparing neighborhood retention and ICA, we find that on Tatoeba, both measures achieve very high correlations, consistent with prior work suggesting that language is encoded more linearly than many lower-level concepts. We also show that it is possible to probe this quality using ICA. For ARCChallenge, our most faithful prompt evaluation, prompting improves the correlation, leading us to draw the conclusion that adding a prompt modifies the query with a transformation related to local structure retention between queries and targets. For other datasets, such as WebFAQ, neighborhood retention reliably predicts model ranking, but linear ICA does not capture the relationship consistently across languages. This points to cross-lingual differences in how question–answer structure is encoded in the embedding space. Finally, unlike neighborhood retention, ICA reveals a meaningful relationship for RTE3, which we analyze further in the next section.



(a) Tatoeba:eng-ell (● $0.904^{\ddagger}$ vs. ♦ 0.218)

(b) RTE3:eng (● 0.414^{*} vs. ♦ 0.436^{*})

Figure 3: Comparison between paired (●) and shuffled (♦) Gini-coefficient (horizontal) for selected datasets, with the associated correlation coefficient and significance. The effect varies: in Tatoeba, the difference grows along MTEB performance and shuffling destroys the relationship, while in RTE3, the difference doesn't depend on the performance and the correlation actually increases.



(a) High peak observed, paired (↑) and shuffled (↓)

(b) Low peak observed, paired (↑) and shuffled (↓)

Figure 4: Connection between neighbor retention and ICA peaks: When local structure is retained, the individual differences between paired queries (circles) and targets (squares) align with the vector corresponding to the peak extracted by ICA (arrow), and therefore shuffling destroys the peak. When no local structure is retained between queries and targets, the effects of shuffling are minor.

Shuffling ICA A pertinent question is what, exactly, ICA captures in our experiment. We address this by separating *local* and *global* signals: we randomly shuffle the targets before computing differences, thereby breaking the correspondence between the two texts in each pair. For example, in Tatoeba, shuffling removes the semantic alignment between translations (a local signal) while preserving the overall difference between English and the other language (a global signal). By comparing results from the paired setting (main experiment) to the shuffled setting (this experiment), we can estimate whether ICA primarily captures pair-specific structure (local) or dataset-level differences (global), and how this relates to task performance. In practice, we reuse the ICA models fitted in the main experiment, but shuffle before subtracting the pairs. Figure 3 shows examples for two datasets, Tatoeba:eng-ell and RTE3:eng, which exhibit qualitatively different behavior under shuffling. In both cases, shuffling reduces the prominence of the ICA peak: the Gini coefficient decreases as expected because the shuffled differences align less consistently with any single extracted ICA component. However, the magnitude of this decrease is dataset-dependent, and the extent to which shuffling disrupts the correlation with the scores differs markedly between Tatoeba and RTE3.

For Tatoeba, shuffling largely destroys the relationship between the Gini coefficient and MTEB performance (from $\mathbf{0.904}^{\ddagger}$ to 0.218). After shuffling, all models yield similar Gini values, and the highest-performing models are affected the most. This indicates that ICA primarily captures a *local* signal tied to the correct pairing between each query and its translation—precisely the type of structure that also drives neighborhood retention. This agreement between ICA and neighborhood retention is consistent with prior work suggesting that language-related differences are often encoded in relatively linear directions in embedding and LLM representational spaces. For RTE3, an opposite pattern emerges: the correlation remains essentially unchanged after shuffling (from 0.414^{*} to 0.436^{*}), and all models are affected in broadly similar ways. This suggests that, for RTE3, ICA mainly captures a *global* signal that does not depend on the specific query–target pairing. This aligns with the results of neighborhood retention, where local structure did not predict benchmark rank-

Table 4: Example of our results on the cosine similarity baselines. Full results in Appendix B.6.

ARCChallenge, unprompt.	–	Raw	Max margin	Margin	Ratio	Z-score
corr(MTEB, Ret_k)	**0.890**‡	–	–	–	–	
corr(MTEB, cos)	–	0.131	0.484*	0.344	0.098	**0.863**‡
corr(MTEB, $\text{Ret}_k \mid$ cos)	–	**0.889**‡	**0.879**‡	**0.878**‡	**0.8924**‡	0.510*
corr(MTEB, cos $\mid \text{Ret}_k$)	–	0.112	0.406*	-0.165	-0.161	0.303

ing. In the MTEB evaluation of RTE3, paired instances are labeled based on their relative distance (entailing pairs are close, contradicting pairs are far). Because we evaluate only contradicting pairs (a selection ensuring our experiments measure a variety of non-STS qualities), MTEB effectively rewards models that separate the two texts sufficiently in the embedding space; to receive a high MTEB score, pairwise structure is less critical than achieving a large overall separation, which, in turn, ICA can extract.

We illustrate the likely mechanism in Figure 4. When an embedding space preserves pair-specific local structure between queries and targets, the query–target differences align with a consistent ICA direction, pertaining to a salient peak; shuffling then disrupts this alignment and substantially reduces peak prominence. Conversely, when ICA primarily captures a global dataset-level shift shared across many pairs, the peak is already weaker, and shuffling has little effect. This observation helps reconcile the behaviors of neighborhood retention and the Gini-based ICA measure across tasks. When the query–target relationship is approximately linear, both neighborhood retention and ICA tend to be strong: local neighborhoods are preserved, and the difference concentrates into a small number of ICA directions. When the relationship is strongly non-linear, neighborhood retention may remain high, but ICA captures only part of the structure, leading to weaker or less consistent Gini–performance correlations. Finally, even when neighborhood retention is low (limited local alignment), ICA can still correlate with performance if the benchmark rewards a global separation signal, as in RTE3. The remaining datasets lie between these extremes and exhibit mixed contributions of local and global structure, with selected examples visualized in Appendix Figure 13.

Supporting experiments Our experiments rely on using non-deterministic ICA models. We analyze the stability of ICA and the word-embedding explanations of the ICA peak dimensions in Appendices B.3 and B.4, respectively. These results show that ICA is able to reliably extract the same underlying representations across different initializations, and that the peak dimensions correspond to words associated with the given tasks. Together, these supporting experiments validate that ICA consistently captures the same underlying signals that our main analysis attributes to it. We also present an example of the cosine similarity baseline experiment in Table 4, with full results and in-depth analysis in Appendix B.6. As a summary, these results show that our results do not reduce to underlying cosine similarity: the partial correlation of Ret_k controlled for cosine similarity remains high, while cosine similarity controlled by Ret_k receives values indicating low correlation. The increasing correlation with the increasingly surrounding-aware cosine similarity types also supports our conclusion on the importance of the geometry of embeddings with respect to evaluation.

6 Discussion and Related Work

In this paper, we furthered the understanding of which underlying properties of embeddings lead to high performance. We investigated two methods of probing structural retention: neighborhood retention and ICA. Our results showed that common embedding evaluation datasets display differences in the structural similarities between the query and target side, but that, in general, embedding models that organize their spaces in an analogous geometric fashion exhibit better performance. Taken together, these results indicate that some prevalent tasks reward models whose query–target relationship is well captured by a small number of linear directions. We simultaneously evaluated the degree of linearity the embeddings exhibit in each evaluation dataset. We uncovered connections between ICA and neighborhood retention and showed how linearity and local-global directions in ICA interoperate: the shuffling results show that ICA can reflect either (i) *local* signals tied to correct pairing or (ii) *global* shifts that do not depend on the specific pairing. This helps explain when neighborhood retention and ICA agree (strong local alignment and approximately linear structure)

and when they diverge (non-linear local structure or benchmarks that emphasize global separation). Beyond these structural effects, we also observe the effects of prompting (most notably ARCChallenge), highlighting that evaluation choices that alter the query representation can affect not only absolute benchmark scores but also the relationship between geometry and performance.

Overall, our results provide evidence that embedding space geometry—the consistency of local neighborhoods between paired inputs and the extent to which pairwise differences concentrate into a small number of directions—is closely tied to benchmark performance in several prominent tasks. We found that probes such as neighborhood retention and ICA peakiness can complement downstream evaluation by helping diagnose *how* and *why* models succeed or fail, and offer lightweight signals that may guide future work on training objectives and task- or condition-aware embeddings.

Our work relies on other previous work on analyzing embedding models. Studies using ICA to analyze embeddings include: Yamagiwa et al. (2023), Li et al. (2024), and Musil & Mareček (2024) who find interpretable axes in word embeddings using ICA; Huertas-García et al. (2023) who evaluate ICA as a dimensionality reduction technique for contextual sentence embeddings; Aida & Bollegala (2025) who study semantic drift in time with ICA; and Liu et al. (2025) use ICA to identify key neurons affecting LLM performance for pruning. Other notable recent work with LLM representations and ICA includes Liu et al. (2025) and Abe et al. (2025). Another widespread tool to interpret structures in representational spaces is Sparse Auto-Encoders (SAEs, Cunningham et al. (2023)). Papadimitriou et al. (2025) explore the linear structures in the embedding spaces of vision-language models using SAEs, and find a balance between monomodal (i.e., either textual or visual) and modality-agnostic subspaces. Saglam et al. (2025) compare the latent spaces of 11 causal language models and find that high-level semantic information exhibits low-dimensionality and linear separability. Chen et al. (2025) find SAEs to efficiently deconstruct polysemantic neurons, and Lan et al. (2025) find universality in SAE-decompositions across different models. Other ways to study structures include matrix-based entropy (e.g., Skean et al. (2025)), anisotropy (e.g., Timkey & van Schijndel (2021); Mickus et al. (2022)), singular value decomposition (e.g., Chang et al. (2022)), clustering embeddings wrt. topics (e.g., Teimouri et al. (2025)), locally linear embeddings (e.g., Lee et al. (2025)), mapping hidden states with unembedding matrix (e.g., Dar et al. (2023)), and analyzing the effect of removing dimensions (Takeshita et al. (2025)).

LIMITATIONS AND FUTURE WORK

The main limitation of our framework is the requirement of a one-to-one pairing of the datasets. This leaves multiclass classification and clustering out of our experiments, both of which are prominent embedding model tasks. Similarly, we do not answer questions about the actual encoding of concepts themselves, rather the relationship they have with each other: our probes characterize *geometric relationships* between paired representations (e.g., local neighborhood alignment and the concentration of query–target differences), rather than directly explaining how specific concepts are encoded. Embedding spaces likely contain additional non-linear or global phenomena that these probes cannot capture; this may already be reflected in tasks/datasets where correlations are weaker. We leave these relationships for future work. Like any benchmark-driven analysis, our findings could be used for leaderboard-oriented optimization rather than general improvements.

Although we evaluate 25 models across five tasks and multiple languages, our findings may not generalize to all embedding models, datasets, or evaluation setups. Practical constraints (e.g., model access, the feasibility of running certain evaluations) limit coverage for some model–task combinations. Our analysis is also correlational: although our structural measures correlate with benchmark ranking, they do not establish which interventions would *causally* improve a model. Our future work will test causality by training or adapting models with explicit structure-preserving objectives and measuring downstream effects. Similarly, while all tested values of k in Ret_k supported our conclusions, understanding why certain values emerge as better estimators of performance is an open question. Our future work will focus on (i) extending the framework to non-paired tasks, (ii) causal analysis by fine-tuning, (iii) incorporating non-linear probes (e.g., SAEs), and (iv) exploring how structure-based signals can support conditional or customizable embeddings, including properties that are not directly benchmarked.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, code auditing, and as one of the tools for collecting candidate related work. We did not use AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods (outside code auditing), assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Providing critical ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work. In the case of using generative AI for code auditing, suggested fixes were tested and verified for logical soundness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We release the code for replicating our results in an anonymous repository. We use datasets available freely online and (majority of) open models in our analysis. In addition to details provided in the main body of text and in the code base, we detail sampling, prompting, and embedding, and hardware & computational cost in the Appendices A.1 to A.4.

REFERENCES

Ryunosuke Abe, Takatomi Kubo, and Kazushi Ikeda. Towards identification of latent structures in language embeddings. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025. URL https://openreview.net/forum?id=HgRkUfQSa4.

Taichi Aida and Danushka Bollegala. SCDTour: Embedding axis ordering and merging for interpretable semantic change detection. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14775–14785, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.797. URL https://aclanthology.org/2025.findings-emnlp.797/.

Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3197–3203, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1309. URL https://aclanthology.org/P19-1309/.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, Kate Soule, Arafat Sultan, and Radu Florian. Granite embedding models, 2025. URL https://arxiv.org/abs/2502.20204.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks, 2025. URL https://arxiv.org/abs/2511.07025.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.9. URL https://aclanthology.org/2022.emnlp-main.9/.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

Yuheng Chen, Pengfei Cao, Kang Liu, and Jun Zhao. The knowledge microscope: Features as better analytical lenses than neurons. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10493–10515, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.516. URL https://aclanthology.org/2025.acl-long.516/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893/.

Michael Dinzinger, Laura Caspari, Kanishka Ghosh Dastidar, Jelena Mitrović, and Michael Granitzer. WebFAQ: A multilingual collection of natural Q&A datasets for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 3802–3811, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3731934. URL https://doi.org/10.1145/3726302.3731934.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal A Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Shalin Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Validad Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zl3pfz4VCV.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding, 2022. URL https://arxiv.org/abs/2007.01852.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9, Prague, jun 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-1401.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao. Jina embeddings: A novel set of high-performance sentence embedding models, 2023a.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents, 2023b.

Johan Himberg, Aapo Hyvärinen, and Fabrizio Esposito. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage*, 22(3):1214–1222, 2004. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2004.03.027. URL https://www.sciencedirect.com/science/article/pii/S1053811904001661.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. KaLM-Embedding: Superior training data brings a stronger embedding model, 2025. URL https://arxiv.org/abs/2501.01028.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 2553–2561, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403305. URL https://doi.org/10.1145/3394486.3403305.

Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. Exploring dimensionality reduction techniques in multilingual transformers. *Cognitive Computation*, 15: 590–612, 2023. doi: https://doi.org/10.1007/s12559-022-10066-8.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. doi: 10.1109/72.761722.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders, 2025. URL https://arxiv.org/abs/2410.06981.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=aJDykpJAYF.

Rongzhi Li, Takeru Matsuda, and Hitomi Yanaka. Exploring intra and inter-language consistency in embeddings with ICA. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19104–19111, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1065. URL https://aclanthology.org/2024.emnlp-main.1065/.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Lucy H. Lin and Noah A. Smith. Situating sentence embedders with nearest neighbor overlap, 2019. URL https://arxiv.org/abs/1909.10724.

Yiheng Liu, Junhao Ning, Sichen Xia, Xiaohui Gao, Ning Qiang, Bao Ge, Junwei Han, and Xintao Hu. Pruning large language models by identifying and preserving functional networks, 2025. URL https://arxiv.org/abs/2508.05239.

Timothee Mickus, Denis Paperno, and Mathieu Constant. How to dissect a Muppet: The structure of transformer embedding spaces. *Transactions of the Association for Computational Linguistics*, 10:981–996, 2022. doi: 10.1162/tacl_a_00501. URL https://aclanthology.org/2022.tacl-1.57/.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Tomáš Musil and David Mareček. Exploring interpretability of independent components of word embeddings with automated word intruder test. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6922–6928, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.605/.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

oprogramador. most-common-words-by-language, 2017. URL https://github.com/oprogramador/most-common-words-by-language/tree/master.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham M. Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=qPsmGjpq1j.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November 2011. ISSN 1532-4435.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Yuxin Ren, Qipeng Guo, Zhijing Jin, Shauli Ravfogel, Mrinmaya Sachan, Bernhard Schölkopf, and Ryan Cotterell. All roads lead to Rome? Exploring the invariance of transformers' representations, 2023. URL https://arxiv.org/abs/2305.14555.

Baturay Saglam, Paul Kassianik, Blaine Nelson, Sajana Weerawardhena, Yaron Singer, and Amin Karbasi. Large language models encode semantics and alignment in linearly separable representations. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2282–2303, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.124. URL https://aclanthology.org/2025.ijcnlp-long.124/.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27705–27726, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1410. URL https://aclanthology.org/2025.emnlp-main.1410/.

Tatoeba community. Tatoeba: Collection of sentences and translations, 2021.

Maryam Teimouri, Jenna Kanerva, and Filip Ginter. A deep dive into multi-head attention and multi-aspect embedding. In Galia Angelova, Maria Kunilovskaya, Marie Escribe, and Ruslan Mitkov (eds.), *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pp. 1263–1270, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2025.ranlp-1.146/.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Stephan Tulkens and Thomas van Dongen. Model2vec: Fast state-of-the-art static embeddings, 2024. URL https://github.com/MinishLab/model2vec.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral,

Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. EmbeddingGemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

Bin Wang, C.-C. Jay Kuo, and Haizhou Li. Just rank: Rethinking evaluation with word and sentence similarities. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6060–6077, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.419. URL https://aclanthology.org/2022.acl-long.419/.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

Suhang Wang, Jiliang Tang, Charu Aggarwal, and Huan Liu. Linked document embedding for classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pp. 115–124, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983755. URL https://doi.org/10.1145/2983323.2983755.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general Chinese embedding, 2023.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 478–487, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/xieb16.html.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Discovering universal geometry in embeddings with ICA. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4647–4675, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.283. URL https://aclanthology.org/2023.emnlp-main.283/.

Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. Axis tour: Word tour determines the order of axes in ICA-transformed embeddings. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 477–506, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.28. URL https://aclanthology.org/2024.findings-emnlp.28/.

Jinsung Yoon and Sercan O Arik. Embedding-Converter: A unified framework for cross-model embedding transformation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25464–25482, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1237. URL https://aclanthology.org/2025.acl-long.1237/.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. Arctic-embed 2.0: Multilingual retrieval without compromise, 2024. URL https://arxiv.org/abs/2412.04506.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1393–1412, 2024.

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. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Methodological Details

In this section, we describe details relating to our methodology.

A.1 Data Sampling

We use the validation/development or test sets to ensure the models we use in the experiments are not trained on our data. We randomly sample up to 5000 documents from the datasets for all of our experiments to ensure manageable resource usage for our experiments: for example, the original WebFAQ dataset contains approximately 100k question-answer pairs for each language, hence using the full dataset would drastically increase the resources taken by embedding as well as time for fitting the ICA model. On the other hand, selecting 5000 examples is not possible in all cases; for instance, the SummEval dataset only contains 100 examples, of which we use 80 and reserve 20 as a test set for future work.

Secondly, some of the datasets contain multiple types of relationships between the texts or multiple options for pair forming. SummEval dataset contains one document and multiple options for summary and associated quality scores, out of which we choose the one with the highest score. The availability of multiple summary options could facilitate using more than the best summary; however, this only increases the number of targets, not queries, and therefore, we opted to only include the best summaries. RTE3 datasets contain both entailing and contradicting passages, out of which we only consider the contradicting examples and discard the rest. We select the contradicting passages because using a mix of labels would be against our assumption of a clear, uniform signal across pairs, and using the entailing pairs would effectively turn RTE3-multi into an STS task.

Datasets are licensed under CC-BY-SA-4.0 (ARCChallenge), CC-BY-4.0 (WebFAQ, RTE3), MIT (SummEval), and CC-BY-2.0 (Tatoeba).

A.2 Embedding models and Prompting

A.2.1 Prompting

Prompts used in our experiments are presented in Table 5. The prompts were selected based on (1) MTEB-given prompt used in the evaluation, if available (ARCChallenge) (2) Prompts given by the dataset-developer, if available, or following examples of most typical prompts given by model developers or best practices in the field (others). While prompt sensitivity is a known feature of LLMs, the best choice would be to use a prompt optimized for each model and dataset separately. However, we leave these additional experiments for future work and therefore adopt the described approach. All prompts are inserted into the query with the template

Table 5: Prompts used in our experiments.

Dataset	Prompt
ARCChallenge	*Retrieve the answer to the question.*
SummEval	*Given a news article, retrieve semantically similar summaries.*
RTE3-multi	*Retrieve contradicting passages.*
Tatoeba	*Retrieve parallel sentences.*
WebFAQ	*Given a web search query, retrieve relevant passages that answer the query.*

Table 6: Embedding models used in this study, number of parameters (Par.) in units of billion, and average score (Sc.) on the MTEB multilingual leaderboard. Models are categorized by provider or base model from which the model is fine-tuned. Models marked with a star (*) are not included in the WebFAQ evaluation, and models with no parameter counts are closed-weight models.

Model	Par. (B)	Sc. ↑	Model	Par. (B)	Sc. ↑
bge-m3	0.568	59.56	granite-107m-multilingual	0.107	51.81
bge-base-en-v1.5	0.033	43.76	granite-125m-english	0.125	44.04
jina-b-en-v1	0.11	40.91	gemini-embedding-001*	-	68.37
jina-v2-small-en	0.033	35.22	EmbeddingGemma	0.308	61.15
multilingual-e5-large-inst.	0.56	63.22	LaBSE	0.471	52.07
e5-mistral-7b-instruct*	7.111	60.25	Qwen3-Embedding-8B*	7.567	70.58
multilingual-e5-large	0.118	58.62	Qwen3-Embedding-0.6B*	0.596	64.34
e5-small	0.033	43.46	gte-multilingual-base	0.305	58.34
text-embedding-3-large*	-	58.96	KaLM-multilingual-mini	0.494	57.05
text-embedding-3-small *	-	54.01	gte-small	0.033	44.1
potion-multilingual-128M	0.128	47.23	llama-embed-nemotron-8b*	7.505	69.46
potion-base-8M	0.008	38.56	snowflake-arctic-l-v2.0	0.568	57.03
			snowflake-arctic-l	0.335	43.33

Instruct: `<prompt>`
Query: `<query>`

while the target is encoded as is.

A.2.2 Embedding Models

In Table 6, we present the models used in this study with parameter counts and average score on MTEB evaluation. Below, we list the licenses for each embedding model.

MIT: BAAI/bge-m3 (Chen et al., 2024), BAAI/bge-base-en-v1.5 (Xiao et al., 2023), intfloat/-multiligual-e5-large-instruct & intfloat/multilingual-e5-large (Wang et al., 2024), intfloat/e5-mistral-instruct & intfloat/e5-small (Wang et al., 2022b; 2023), minishlab/potion-multilingual-128M & minishlab/potion-base-8M (Tulkens & van Dongen, 2024), HIT-TMG/KaLM-embedding-multilingual-mini-v1 (Hu et al., 2025), thenlper/gte-small (Li et al., 2023).

Apache-2.0: Jinaai/jina-b-en-v1 (Günther et al., 2023a), Jinaai/jina-v2-small-en (Günther et al., 2023b), IBM-granite/granite-197-multilingual & IBM-granite/granite-125M-english (Awasthy et al., 2025), Google/LaBSE (Feng et al., 2022), Qwen/Qwen3-Embedding-8B & Qwen/Qwen3-Embedding-0.6B (Zhang et al., 2025), AliBaBaNLP/gte-multilingual-base (Zhang et al., 2024), Snowflake/snowflake-arctic-l-v2.0 & Snowflake/snowflake-arctic-l (Yu et al., 2024).

CC-BY-4.0: Google/gemini-embedding-001 (Gemini Team, 2025), Google/EmbeddingGemma (Vera et al., 2025)

Other permissive: NVIDIA/llama-embed-nemotron-8b (Babakhin et al., 2025)

A.2.3 Data Contamination

To account for the possibility of data contamination affecting our results, we investigate the fine-tuning datasets reported by model developers in MTEB. We use the `mteb`-library to search for possible overlaps between our datasets and the models' finetuning data. We find that out of the 22 open models we use, 21 report this information, and we find no overlap with the datasets used in this study. In the case of thenlper/gte-small, the one model missing this information, we refer to the associated publication (Li et al. (2023), Appendix A), which states that there is again no overlap with our datasets. We conclude that while we are not able to verify these results for the closed models (by OpenAI and Google), we have ruled out data contamination in almost all models. We also note

that explicitly using MTEB datasets for finetuning and not reporting them in the finetuning data is against the conventions of the field.

A.3 Independent Component Analysis

In this section, we explain ICA and provide an illustrative example of its use. Similarly to PCA, the objective of ICA is to fit a matrix A, known as *unmixing matrix*, with a given set of constraints; instead of variance preservation, the target is the independence of resulting components S. Compared to PCA, which uses the covariance of the data to construct A, ICA also includes higher order moments, skewness and kurtosis, to estimate the unmixing matrix A (Yamagiwa et al., 2023). The independent components produced by ICA can be calculated with

$$S = XA,$$

$A \in \mathbb{R}^{d \times d_{ICA}}$ is the fitted unmixing matrix, $X \in \mathbb{R}^{N \times d}$ is the observed signals, and $S \in \mathbb{R}^{N \times d_{ICA}}$ is the resulting independent components.

The motivation behind ICA can be understood with a simple example: Assume N microphones in a room with M sound sources; each microphone catches the sound of multiple sources with different volumes. The recording of each microphone can be presented as a linear combination of all the sources $x_n = b_{1n}s_1 + b_{2n}s_2 + \cdots + b_{nm}s_m$. Thus, the process of recoding the observed values X can be represented as matrix multiplication $X = SB$. Furthermore, X can be decoded into the sources S again with the Moore-Penrose pseudo-inverse of $B^+ = A$ as $S = XA$. If $M = N$ and all sources are independent as assumed (i.e., $d = d_{ICA}$), the pseudo-inverse reduces to the regular matrix inverse. However, in a typical use case, only the observed matrix X is known, and hence the matrix A is estimated from observed values X and the assumption of independence of sources.

Using ICA on language model output embeddings may call into question the use of embeddings as the source signal: the most canonical examples of applying ICA are on time-series or otherwise "continuous" data, like images, with embeddings lacking this structure. Despite this, ICA has been successfully applied to embeddings in previous work, as seen in our Background and Related Work sections. Another obstacle we face with ICA is the cardinality of our datasets. To fit an ICA model of dimension $\dim_{ICA} = N'$, we need at least N' instances of embeddings; another reason we run our experiments with low-dimensional ICA.

A.4 Computational Details

Our experiments cover 25 models, 29 datasets, and two prompting options (with and without prompt); altogether, we run experiments for 1450 combinations. Additionally, we run the MTEB evaluation for a subset of the combinations. We use AMD MI250x GPUs and AMD EPYC 7763 CPUs on *an anonymous HPC environment* to run all experiments. All values reported below are top-end approximations.

On average, the experiments required 6-10 GB of memory, while some outliers, like the WebFAQ evaluation, required up to 20 GB. GPU usage varies based on the dataset size, both in time and in the number of GPUs used. We used 1-2 GPUs to embed the datasets and 2-8 GPUs to run MTEB evaluation. The total GPU computational cost of the embedding is 1650 GPU hours. The MTEB evaluation of ARCChallenge, Tatoeba, and SummEval required approximately 20 GPU hours, all with 2 GPUs, while the cost was significantly higher for WebFAQ, for which we estimate 1440 GPU hours, caused by the large size of the full evaluation set and the need to use 8 GPUs for some models. Calculating word embeddings for ICA explanations required approximately 10 GPU hours. All other experiments (calculation of neighborhood retention, fitting ICA models, calculating statistics, stability analysis) required a comparatively smaller amount of CPU hours: neighborhood retention experiment approximately 10 minutes/experiment, fitting ICA approximately 30 minutes (with some rare outliers taking more than 6 hours), and all analyses less than 5 minutes. In total, the computational cost for our experiment was 3120 GPU hours and 1100 CPU hours, or on average 2.15 hours of GPU and 45 minutes of CPU usage per dataset, model, and prompt combination.

Outside of the cost required to run the experiments in the paper, our preliminary experiments also required computational resources. For example, we experimented with values of d_{ICA}, which carries an associated CPU cost comparable to the CPU usage of the full experiment. However, as the

Table 7: Spearman correlation (↑) for Gini-coefficient calculated from ICA^{64} with MTEB scores, with bolded values for strong relationship (≥ 0.7). Star ($*$) indicates significance at level $p < 0.05$, dagger (†) at level $p < 0.01$, and double dagger (‡) at level $p < 0.001$.

Dataset	Unpromted	Prompted	Dataset	Unpromted	Prompted
ARCChallenge	**0.732**‡	0.669‡	SummEval	0.278	0.275
RTE3:eng	0.268	0.238	WebFAQ:bul	-0.478*	-0.406
RTE3:fra	0.608†	0.596†	WebFAQ:deu	0.595†	**0.703**†
RTE3:ita	0.561†	0.669‡	WebFAQ:ell	-0.300	-0.026
RTE3:deu	0.417*	0.499*	WebFAQ:eng	0.290	0.447
Tatoeba:deu-eng	**0.838**‡	**0.945**‡	WebFAQ:fas	-0.552*	-0.581*
Tatoeba:bul-eng	**0.917**‡	**0.895**‡	WebFAQ:fin	**0.707**†	0.562*
Tatoeba:ukr-eng	**0.937**‡	**0.935**‡	WebFAQ:fra	0.659†	0.321
Tatoeba:ita-eng	**0.920**‡	**0.898**‡	WebFAQ:hin	-0.294	-0.323
Tatoeba:fra-eng	**0.849**‡	**0.879**‡	WebFAQ:ita	**0.748**‡	0.688†
Tatoeba:vie-eng	**0.874**‡	**0.884**‡	WebFAQ:jpn	0.659†	0.688†
Tatoeba:ell-eng	**0.872**‡	**0.853**‡	WebFAQ:ukr	-0.026	-0.123
Tatoeba:jpn-eng	**0.854**‡	**0.844**‡	WebFAQ:vie	0.490*	0.614†
Tatoeba:hin-eng	**0.892**‡	**0.885**‡	WebFAQ:zho	0.298	0.249
Tatoeba:fin-eng	**0.888**‡	**0.868**‡			

embeddings could be saved and reused in all experiments, the preliminary experiments do not affect GPU hours substantially.

B Complementary Results

In this section, we present the results for ICA^{64}, neighborhood retention results for suboptimal number of neighbors, analysis on the stability of ICA, and other supplementary examples and figures.

B.1 Results for ICA (dim=64)

Table 7 presents the our results for ICA^{64}. Similar conclusions hold: ARCChallenge and Tatoeba receive high scores, while the scores of WebFAQ vary between languages. RTE3 dataset again shows increased scores compared to the neighborhood retention experiment.

B.2 Neighborhood Retention for Suboptimal Number of Neighbors

In the main body of the paper, we present the neighborhood retention experiment results for $k = 10$, which was found to be in the top 3 values for all datasets. We remark that all tested values of k support the conclusions. In Table 8 we justify this claim by presenting the results for one of the universally worst values of $k = 0.1N$ (and $k = 5$ for SummEval), which was analogously found to be among the bottom 3 values. These results show that again, 4/5 datasets show significant strong to medium correlation, with RTE3 being the outlier. ARCChallenge shows noticeably smaller values for $k = 0.1N = 94$ than the optimal $k = 10$ results, displaying that very local information between queries and targets has a lot of predictive power in this task. On the other hand, in Tatoeba and WebFAQ, some languages (like Greek and Finnish) show higher correlation when analyzed through the larger neighborhood ($k = 0.1N = 500$). We conclude that despite the differences, the overall conclusions we draw are still justified even with the worst value of k.

B.3 Stability of ICA

As the FastICA algorithm is non-deterministic, analysis of the stability of the results we acquire from these methods is necessary. To evaluate this, we use 4 models (LaBSE, Qwen3-Embedding-0.6B, bge-m3, multilingual-e5-large). We opted to use Tatoeba and these models for this analysis, as analyzing the stability of the peak requires the presence of a sustained peak in the results. We

Table 8: Spearman correlation (↑) for neighborhood retention with suboptimal $k = 0.1N$ with MTEB scores, with bolded values for strong relationship (≥ 0.7). Star ($*$) indicates significance at level $p < 0.05$, dagger (†) at $p < 0.01$, and double dagger (‡) at $p < 0.001$.

Dataset	Unprompted	Prompted	Dataset	Unprompted	Prompted
ARCChallenge	**0.781**‡	**0.866**‡	SummEval	0.587†	0.615†
RTE3:eng	-0.304	-0.048	WebFAQ:bul	**0.903**‡	**0.915**‡
RTE3:fra	0.149	0.268	WebFAQ:deu	**0.843**‡	**0.841**‡
RTE3:ita	0.344	0.495*	WebFAQ:ell	**0.977**‡	**0.915**‡
RTE3:deu	0.009	0.180	WebFAQ:eng	0.538*	0.453
Tatoeba:deu-eng	**0.853**‡	**0.798**‡	WebFAQ:fas	**0.940**‡	**0.882**‡
Tatoeba:bul-eng	**0.964**‡	**0.897**‡	WebFAQ:fin	**0.946**‡	**0.872**‡
Tatoeba:ukr-eng	**0.922**‡	**0.876**‡	WebFAQ:fra	**0.891**‡	**0.831**‡
Tatoeba:ita-eng	**0.916**‡	**0.852**‡	WebFAQ:hin	**0.880**‡	**0.847**‡
Tatoeba:fra-eng	**0.878**‡	**0.829**‡	WebFAQ:ita	**0.893**‡	**0.775**‡
Tatoeba:vie-eng	**0.953**‡	**0.870**‡	WebFAQ:jpn	**0.874**‡	**0.837**‡
Tatoeba:ell-eng	**0.962**‡	**0.875**‡	WebFAQ:ukr	**0.940**‡	**0.862**‡
Tatoeba:jpn-eng	**0.952**‡	**0.888**‡	WebFAQ:vie	**0.891**‡	**0.812**‡
Tatoeba:hin-eng	**0.930**‡	**0.810**‡	WebFAQ:zho	**0.841**‡	**0.849**‡
Tatoeba:fin-eng	**0.975**‡	**0.914**‡			

fit 8 differently initialized ICA[32] models per dataset-model combination and evaluate the stability of the Gini-coefficient measure and the peak-related direction extracted by ICA. In this experiment, we use only the unprompted embeddings to limit the number of models needed to fit in this step and because we are evaluating the stability of our methodology, not the effect of prompts.

We first evaluate the stability of the Gini coefficient. We find that among the different initializations, the Gini coefficient is stable, measured with the strict coefficient of variation at threshold 0.01, in all model-dataset combinations except for Qwen-embedding-0.6B and Finnish (and even in this case, the standard deviation across runs is just above the threshold). This overwhelming stability led to our usage of Gini over other metrics (peak-to-mean-ratio, coefficient of variation).

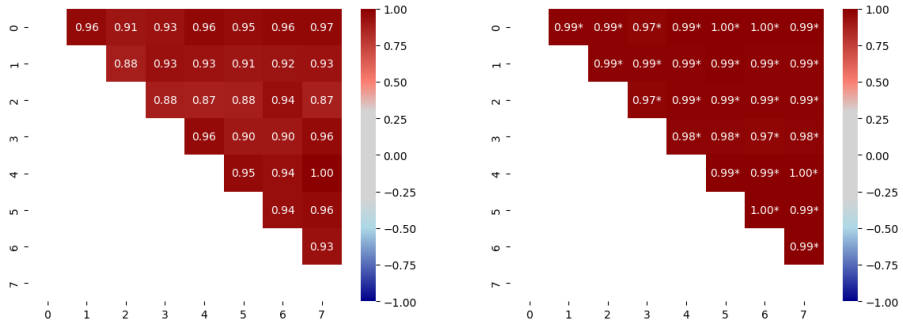
The other question on the stability of ICA is how consistent the fitted model is. The Icasso method (Himberg et al., 2004) has been suggested to evaluate the fluctuations of results produced by ICA. We modify this setup slightly and incorporate greedy dimension matching inspired by the work of Yamagiwa et al. (2023), as we are mostly interested in the stability of the peak dimension, which we assign with $\arg\max$-function for this step. Following Li et al. (2024), we measure stability with

$$\sigma_{ij} = \left| \frac{\mathrm{Cov}(s_i, s'_j)}{\sqrt{\frac{1}{d}\sum_k s_{ik}^2}\sqrt{\frac{1}{d}\sum_k s'^2_{jk}}} \right|,$$

where s_i and s'_j are rows from the unmixing matrices from different runs. Our results also show that ICA can reliably produce equivalent results: the row values of the unmixing matrices corresponding to the peak correlate strongly across runs, meaning ICA is able to latch onto the *same representation in the embeddings*, an example of which is shown in Figure 5. Out of the ten languages tested, LaBSE and multilingual-e5-large display stability comparable to the exemplified figures. For Qwen3-Embedding-0.6B and bge-m3, we reach similarly comparable stability for 9/10 and 7/10 languages. Even in the negative cases, most comparisons reach correlations comparable to the example: For example, on bge-m3 and the Ukrainian split, 6/8 ICA models from the 8 runs agree on the peak, but the remaining 2/8 only correlate with each other. We conclude that, on average, we reach high stability on the fitted ICA models.

B.4 Interpretability through ICA

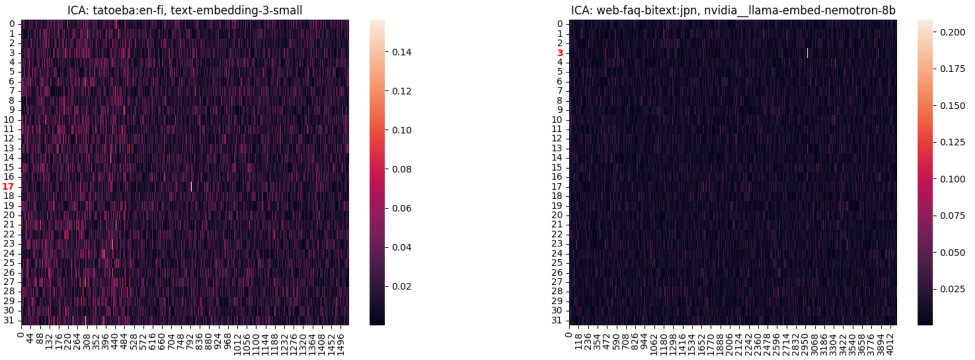
We follow Li et al. (2024) and Yamagiwa et al. (2024) and explain the results of the fitted ICA models by using word embeddings of the top 10k words per language (oprogramador, 2017), and select the words that obtain the highest and lowest values for the peak dimension. Again, we use



(a) Cross-run stability of ICA

(b) Cross-run stability of ICA peaks

Figure 5: (a) Aggregated similarity between the components of 8 differently initialized ICA models (b) Similarity between peak components across runs. Both heatmaps are on LaBSE and Tatoeba:eng-deu. Star (*) indicates significance at level $p < 0.05$.



(a) text-embedding-3-small, Tatoeba:fin-eng

(b) llama-embed-nemotron-8b, WebFAQ:jpn

Figure 6: Two unmixing matrices displaying one or two embedding model dimensions that strongly contribute to the peak dimension (bolded and red) in ICA.

Table 9: Examples of top and bottom 10 most salient words for the peak-dimension in ICA[32], with translations. The emergence of "Finnish translation", "American", "Notre-Dame" clearly indicates the presence of translation, language, and location-related signals, while units, shapes, and animals indicate common answers to questions in the ARCChallenge dataset.

Model & Dataset	Peak top 10	Peak bottom 10
llama-embed-nemotron-8b Tatoeba:eng-fin	quotations, quotes, quote, maybe, perhaps, memories, somehow, remember, sometimes, remembered	do-you-use,(käytätkö) use (impr.),(käyttäkää) translated-to-Finnish,(suomentanut) Finnish-translation,(suomennos) I-send,(lähtetän) can-you,(voitteko) could-you,(voisitteko) do-you-want,(tahdotko) would-you-take,(ottaisitko) you-use,(käytät)
gte-multilingual-base Tatoeba:eng-fra	travelling, drinking, USA, World, living, eating, speech, American, America, speaking	I love,(J'aime) calculated,(calculée) Notre-Dame, It was,(C'était) it was,(c'était) calculated,(calculé) I,(Je) time,(l'heure) Yet,(Pourtant) love,(l'amour)
e5-mistral-instruct ARCChallenge	rope, worm, needle, meters, lightning, dried, snake, cylinder, meters, frog	education, opinions, justice, educators, stakeholders, politics, policies, attitudes, teachers

only the unprompted embeddings (the prompt would likely affect the keywords for each peak). For Tatoeba, we use both English and non-English word lists, and for WebFAQ and RTE-multi, we use the appropriate languages. These results allow us to see which words each ICA dimension corresponds to the strongest.

Examples of results are given in Table 9. From the selection of words, it is clear which side corresponds to the query and which to the target. Although not perfect, the presence of Finnish translation-related vocabulary on the Tatoeba:eng-fin task, and the American and French related vocabulary on Tatoeba:eng-fra task, suggests that the peak seen in our experiments does correspond to a language-specific signal. Similarly, in the ARCChallenge task, the top 10 words are associated with answers (animals, shapes, and units), as the bottom words can be seen as educational, opinionated, and questioning, corresponding to the contents of the ARCChallenge dataset. We concede that the perceived quality for the obtained word explanations is heavily model and dataset-dependent; for example, SummEval or RTE3 do not produce a clearly categorizable set of words.

Another way of using ICA for interpretation is to investigate the fitted unmixing matrix. Figure 6 presents the matrix for llama-embed-nemotron-8b on Tatoeba:eng-fin task and llama-embed-nemotron-8b on WebFAQ:jpn, with vertical values indicating dimensions of the ICA model and horizontal values corresponding to the dimensions of the embedding model. In the first example, the dimension corresponding to the peak, 17, shows one distinctive high value, highlighting the dimension of the original model that most corresponds to the peak. The same can be seen on the second example, where two dimensions are highlighted as most contributing to the peak. This means that Finnish translation and Japanese question-answering rely on only one or two dimensions in the respective models.

B.5 Complementary figures

In this section, we present supplementary figures. Figures 7–12 display visualizations of MTEB-scores against Ret_k with annotated model names and parameter counts, and Figure 13 presents shuffling effect visualization for ARCChallenge and SummEval. Each caption contains a short analysis of the most important takeaways.

B.6 Comparison to Cosine Similarity Baselines

Neighborhood retention operates on relatively low-complexity statistics of embeddings. To evaluate whether our previously presented results actually reduce to pure cosine similarity, in this section, we investigate a comparison of Ret_k to a collection of 5 cosine similarity variations. For a paired

dataset $D = (t, t')_k$ with N paired instances, define

$$\text{raw} = \frac{1}{N} \sum_i \cos(t_i, t'_i)$$

$$\text{margin} = \frac{1}{N} \sum_i \cos(t_i, t'_i) - \text{mean}_j \cos(t_i, t'_j)$$

$$\text{max-margin} = \frac{1}{N} \sum_i \cos(t_i, t'_i) - \max_j \cos(t_i, t'_j)$$

$$\text{ratio} = \frac{1}{N} \sum_i \frac{\cos(t_i, t'_i)}{\text{mean}_j \cos(t_i, t'_j)}$$

$$\text{z-score} = \frac{1}{N} \sum_i \frac{\cos(t_i, t'_i) - \text{mean}_j \cos(t_i, t'_j)}{\text{std}_j \cos(t_i, t'_j)}.$$

These cosine similarity baselines, motivated in part by Artetxe & Schwenk (2019), represent different levels of measuring similarity: absolute closeness (raw), simple retrieval (max-margin, the sign of which reduces to Recall@1), and increasingly structure-aware metrics (margin, ratio, z-score). Similar to Ret_k, we calculate these 5 scores for our datasets and correlate them with MTEB performance. Additionally, we report the partial correlations for Ret_k ($k = 10$) and the different types of cosine baseline, controlling for each in respective cases. The former results show the variation that different cosine similarity types have and the latter results specifically aim to answer if our main results incidentally measured only cosine similarity.

The results are presented in Tables 10–14, each with short individual analysis in the caption. The results show clearly that in each cosine similarity type, dataset, and prompt option, the correlation values obtained from Ret_k are generally higher, which supports the conclusion that our work does not reduce to cosine similarity. Especially remarkable is that the partial correlation of Ret_k controlled with cosine similarity ($\text{Ret}_k|\text{cos}$) is consistently higher than the partial correlation of cosine similarity controlled with Ret_k ($\text{cos}|\text{Ret}_k$), which in most cases falls in or close to -0.3 to 0.3 range. The highest correlations between cosine similarity and MTEB performance are achieved on max-margin, corresponding to retrieval, and z-score cosine similarity, corresponding to more structural awareness. Taken together, we draw the conclusion that our Ret_k metric does not reduce to cosine similarity, and note that the on average better results on z-score cosine similarity hint even more at the relation between embedding space structure and embedding model performance. The RTE3 datasets again show no correlation, for reasons which were discussed in Section 5.

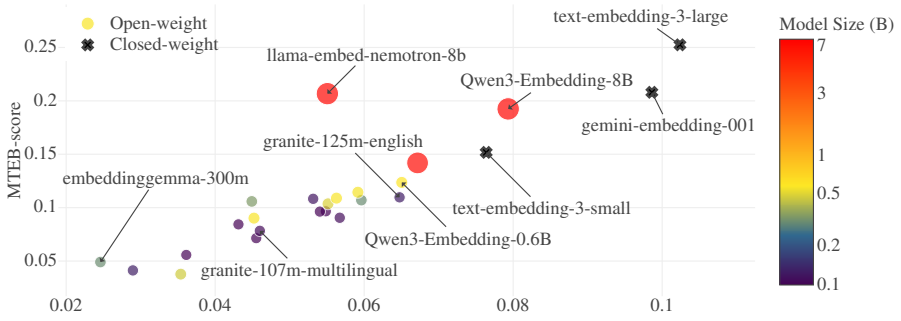


Figure 7: ARCChallenge: Neighborhood retention (horizontal) vs. MTEB-score (Spearman r **0.890**‡). This figure shows that outside the clear outlier, llama-embed-nemotron-8b, the relationship between neighborhood retention and MTEB-score is strong. The largest models receive both the highest MTEB performance and retention scores. Among the smaller models, size shows more variation.

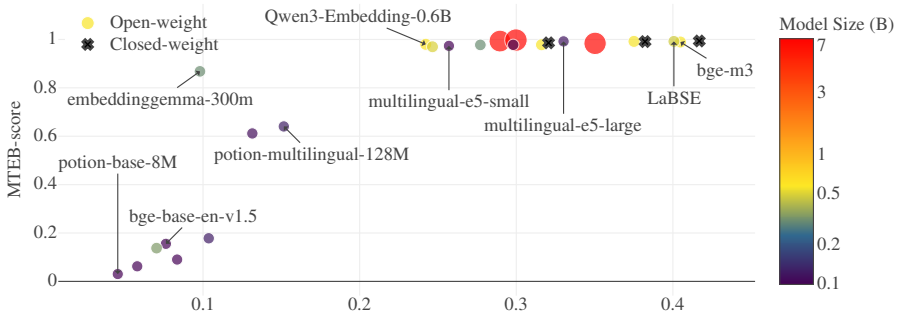


Figure 8: Tatoeba:deu-eng: Neighborhood retention (horizontal) vs. MTEB-score (Spearman r **0.888**‡). German is among the least correlated Tatoeba languages in our experiment. Clearly, one possible cause for this is the fact that so many models receive close to perfect scores on this task, reflecting the overall good German capabilities of the tested models. Slight outlier behavior can be seen from embeddinggemma-300m. Otherwise, similar conclusions hold as for the other Tatoeba tasks, namely that multilingual versions outperform the monolingual versions within model families.

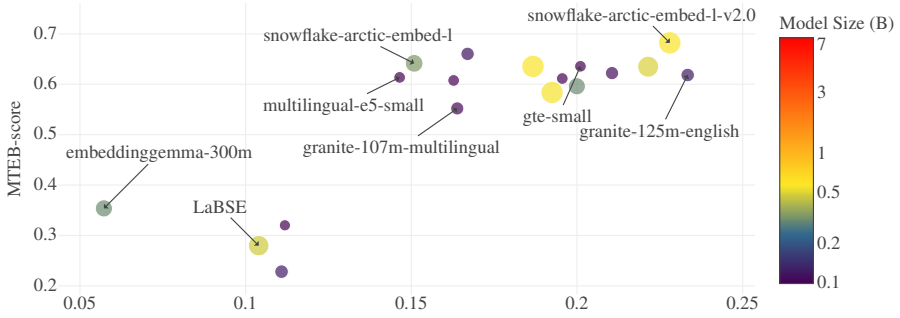


Figure 9: WebFAQ:eng: Neighborhood retention (horizontal) vs. MTEB-score (Spearman r 0.620†). The English subset of WebFAQ was a clear outlier in our results, with a low correlation coefficient. Compared to Figure 10, which shows Greek results (among the best), it is clear where the difference emerges: on the English WebFAQ task, almost all models reach similar performance, and only 4 models struggle on the task. This is in line with the fact that English is the most used language for language model training, and leads to a situation where, unlike with other languages, models with middle-range scores of 0.4–0.5 do not exist, and therefore they are arranged into two clusters–those that solve the task and those that struggle–instead of the relationship seen in the other languages.

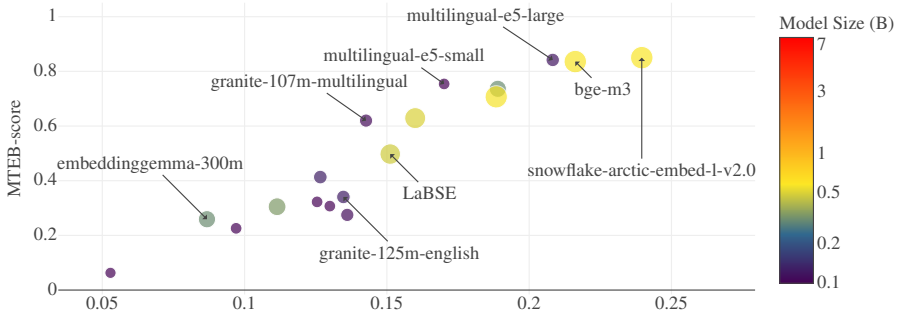


Figure 10: WebFAQ:ell: Neighborhood retention (horizontal) vs. MTEB-score (Spearman r **0.946**‡). This Greek dataset received one of the highest correlation scores in WebFAQ experiments. Compare to Figure 9.

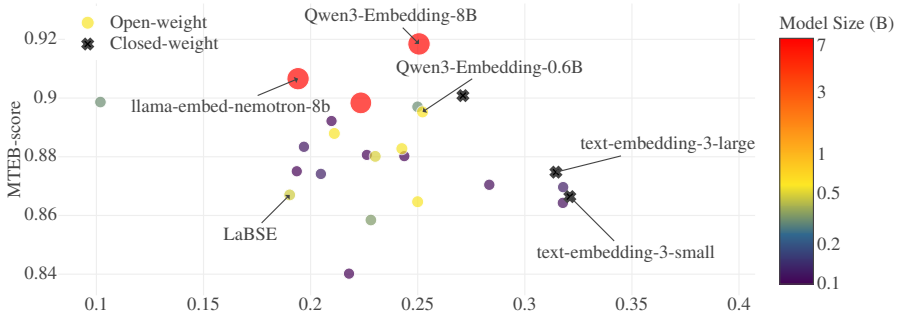


Figure 11: RTE3:eng: Neighborhood retention (horizontal) vs. MTEB-score (Spearman r -0.230). As seen in Section 5, RTE3-multi shows almost no relationship between retention and evaluation score, which we tied to the way MTEB evaluates contradiction in Section 5: instances close to each other are classified as entailing, while instances far apart are classified as contradicting.

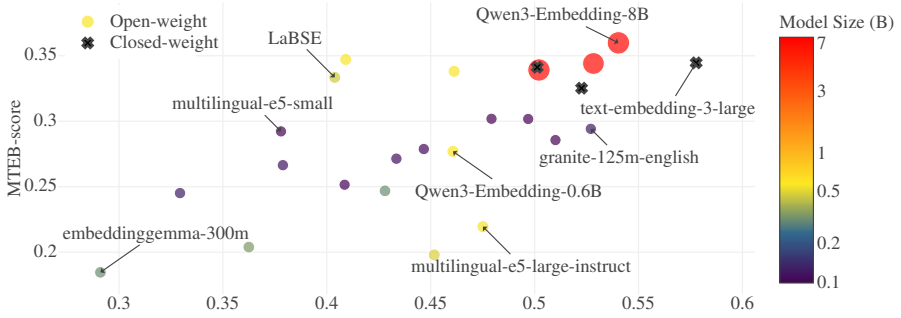
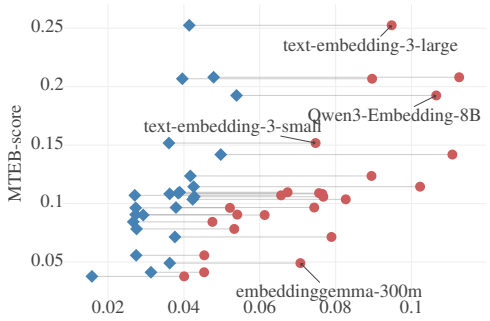
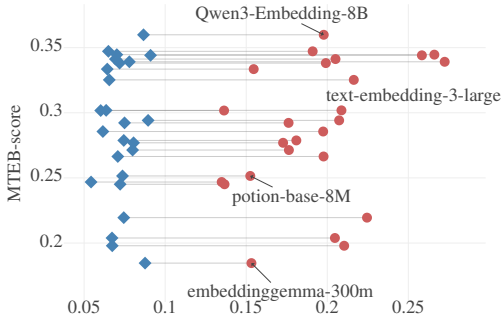


Figure 12: SummEval: Neighborhood retention (horizontal) vs. MTEB-score (Spearman r $0.643^{\ddagger}$). These results display a lot of variance in the middle range of retention, while the largest models are clustered closely together at the highest end. Small models, like granite-125M-english, reach comparable results and retention to the largest models.



(a) ARCChallenge (● $0.810^{\ddagger}$, ◆ $0.689^{\ddagger}$)



(b) SummEval (with prompt, ● 0.428^{*}, ◆ 0.220)

Figure 13: Shuffling experiment visualisation for two additional datasets. (a) Shuffling affects ARCChallenge, but the relationship is still noticeable and significant after shuffling. This means that while neighborhood retention correlated even more strongly with performance on ARCChallenge, global information is still enough to distinguish model performance differences. (b) The effect on prompted SummEval: Shuffling affects the correlation heavily. SummEval saw medium correlation in neighborhood retention, which leads us to conclude that summarization is encoded non-linearly and locally on average in the models we test.

Table 10: Results for cosine similarity type "raw": correlation with MTEB-evaluation scores with our neighborhood retention (Ret_k), cosine similarity (cos), partial correlation of Ret_k controlled with cosine similarity ($\text{Ret}_k|\text{cos}$), and partial correlation of cosine similarity controlled with Ret_k ($\text{cos}|\text{Ret}_k$) for both unprompted and prompted setups. These results show that Ret_k consistently outperforms raw cosine similarity, and raw cosine similarity does not account for the high correlation values achieved.

cos: raw	Unprompted				Prompted			
Dataset	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$
ARCChallenge	**0.890**‡	0.131	**0.889**‡	0.112	**0.945**‡	0.136	**0.944**‡	0.012
SummEval	0.643‡	0.231	0.621†	-0.092	0.588†	0.087	0.601†	-0.177
RTE3-multi:en	-0.230	-0.063	-0.223	-0.026	-0.058	0.052	-0.064	0.059
RTE3-multi:fr	0.149	-0.475*	0.109	-0.466*	0.268	-0.238	0.243	-0.209
RTE3-multi:it	0.389	-0.303	0.386	-0.299	0.488*	-0.138	0.487*	-0.136
RTE3-multi:de	0.025	-0.438*	0.041	-0.439*	0.162	-0.099	0.158	-0.092
Tatoeba:deu-eng	**0.888**‡	0.487*	**0.856**‡	0.198	**0.904**‡	0.289	**0.895**‡	0.006
Tatoeba:bul-eng	**0.962**‡	0.570†	**0.944**‡	0.003	**0.928**‡	0.307	**0.922**‡	-0.166
Tatoeba:ukr-eng	**0.939**‡	0.570†	**0.910**‡	0.109	**0.918**‡	0.370	**0.906**‡	-0.131
Tatoeba:ita-eng	**0.917**‡	0.478*	**0.895**‡	0.199	**0.884**‡	0.304	**0.872**‡	-0.067
Tatoeba:fra-eng	**0.925**‡	0.397*	**0.911**‡	-0.068	**0.895**‡	0.217	**0.891**‡	-0.139
Tatoeba:vie-eng	**0.945**‡	0.574†	**0.919**‡	0.115	**0.916**‡	0.337	**0.906**‡	-0.092
Tatoeba:ell-eng	**0.968**‡	0.616†	**0.948**‡	-0.139	**0.922**‡	0.276	**0.919**‡	-0.216
Tatoeba:jpn-eng	**0.968**‡	0.532†	**0.957**‡	0.245	**0.939**‡	0.355	**0.934**‡	0.236
Tatoeba:hin-eng	**0.931**‡	0.650‡	**0.877**‡	0.011	**0.876**‡	0.371	**0.857**‡	-0.130
Tatoeba:fin-eng	**0.968**‡	0.486*	**0.958**‡	0.068	**0.959**‡	0.242	**0.957**‡	-0.133
WebFAQ:bul	**0.926**‡	-0.381	**0.913**‡	0.040	**0.866**‡	-0.015	**0.876**‡	-0.267
WebFAQ:deu	**0.889**‡	0.319	**0.888**‡	0.308	**0.856**‡	0.302	**0.840**‡	0.067
WebFAQ:ell	**0.946**‡	-0.304	**0.941**‡	0.023	**0.930**‡	0.005	**0.934**‡	-0.235
WebFAQ:eng	0.620†	0.381	0.549*	0.171	0.544*	0.271	0.490*	-0.012
WebFAQ:fas	**0.934**‡	-0.342	**0.930**‡	0.262	**0.893**‡	0.049	**0.892**‡	0.024
WebFAQ:fin	**0.953**‡	0.368	**0.963**‡	0.567*	**0.913**‡	0.342	**0.903**‡	0.145
WebFAQ:fra	**0.878**‡	0.224	**0.885**‡	0.312	**0.843**‡	0.232	**0.834**‡	0.033
WebFAQ:hin	**0.909**‡	-0.321	**0.908**‡	0.305	**0.891**‡	0.092	**0.895**‡	0.208
WebFAQ:ita	**0.907**‡	0.366	**0.907**‡	0.367	**0.816**‡	0.342	**0.792**‡	0.107
WebFAQ:jpn	**0.864**‡	-0.011	**0.866**‡	0.132	**0.827**‡	0.139	**0.823**‡	0.007
WebFAQ:ukr	**0.942**‡	-0.432	**0.937**‡	-0.345	**0.909**‡	0.044	**0.909**‡	-0.063
WebFAQ:vie	**0.936**‡	0.036	**0.936**‡	-0.043	**0.903**‡	0.278	**0.895**‡	0.052
WebFAQ:zho	**0.891**‡	-0.042	**0.890**‡	0.002	**0.841**‡	0.224	**0.832**‡	0.015

Table 11: Results for cosine similarity type "max-margin": correlation with MTEB-evaluation scores with our neighborhood retention (Ret_k), cosine similarity (cos), partial correlation of Ret_k controlled with cosine similarity ($\text{Ret}_k|\text{cos}$), and partial correlation of cosine similarity controlled with Ret_k ($\text{cos}|\text{Ret}_k$) for both unprompted and prompted setups. In this case, while we achieve high correlation values for cosine similarity, Ret_k still consistently outperforms max-margin cosine similarity. The partial correlations also support the conclusion that max-margin cosine similarity does not account for the high correlation values achieved in Ret_k.

cos: max-margin	Unprompted				Prompted			
Dataset	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$
ARCChallenge	**0.890**‡	0.484*	**0.879**‡	0.406*	**0.945**‡	0.566†	**0.921**‡	0.198
SummEval	0.643‡	0.406*	0.582†	0.243	0.588†	0.428*	0.525†	0.311
RTE3-multi:en	-0.230	-0.134	-0.194	-0.044	-0.058	0.034	-0.069	0.051
RTE3-multi:fr	0.149	0.325	-0.153	0.327	0.268	0.483*	-0.059	0.421*
RTE3-multi:it	0.389	0.395	0.153	0.170	0.488*	0.575†	0.141	0.374
RTE3-multi:de	0.025	0.217	-0.162	0.267	0.162	0.388	-0.099	0.369
Tatoeba:deu-eng	**0.888**‡	**0.802**‡	0.659‡	0.220	**0.904**‡	**0.778**‡	**0.733**‡	0.062
Tatoeba:bul-eng	**0.962**‡	**0.855**‡	**0.856**‡	0.160	**0.928**‡	**0.815**‡	**0.770**‡	0.126
Tatoeba:ukr-eng	**0.939**‡	**0.829**‡	**0.791**‡	0.012	**0.918**‡	**0.811**‡	**0.741**‡	0.121
Tatoeba:ita-eng	**0.917**‡	**0.808**‡	**0.741**‡	0.136	**0.884**‡	**0.786**‡	0.668‡	0.169
Tatoeba:fra-eng	**0.925**‡	**0.849**‡	**0.775**‡	0.474*	**0.895**‡	**0.822**‡	0.658‡	0.278
Tatoeba:vie-eng	**0.945**‡	**0.865**‡	**0.776**‡	0.244	**0.916**‡	**0.813**‡	**0.727**‡	0.074
Tatoeba:ell-eng	**0.968**‡	**0.882**‡	**0.847**‡	0.103	**0.922**‡	**0.838**‡	**0.704**‡	0.029
Tatoeba:jpn-eng	**0.968**‡	**0.836**‡	**0.889**‡	-0.077	**0.939**‡	**0.825**‡	**0.795**‡	-0.007
Tatoeba:hin-eng	**0.931**‡	**0.858**‡	**0.741**‡	0.328	**0.876**‡	**0.808**‡	0.586†	0.136
Tatoeba:fin-eng	**0.968**‡	**0.892**‡	**0.839**‡	0.224	**0.959**‡	**0.871**‡	**0.819**‡	0.043
WebFAQ:bul	**0.926**‡	**0.866**‡	0.671†	0.194	**0.866**‡	**0.783**‡	0.641†	0.302
WebFAQ:deu	**0.889**‡	**0.785**‡	0.675†	0.092	**0.856**‡	**0.771**‡	0.619†	0.256
WebFAQ:ell	**0.946**‡	**0.884**‡	**0.805**‡	0.514*	**0.930**‡	**0.736**‡	**0.840**‡	0.065
WebFAQ:eng	0.620†	0.401	0.523*	-0.101	0.544*	0.441	0.370	-0.111
WebFAQ:fas	**0.934**‡	**0.775**‡	**0.835**‡	0.227	**0.893**‡	**0.761**‡	**0.746**‡	0.284
WebFAQ:fin	**0.953**‡	**0.920**‡	0.674†	0.302	**0.913**‡	**0.789**‡	**0.757**‡	0.172
WebFAQ:fra	**0.878**‡	**0.847**‡	0.485*	0.239	**0.843**‡	**0.806**‡	0.505*	0.313
WebFAQ:hin	**0.909**‡	**0.839**‡	**0.733**‡	0.458	**0.891**‡	**0.837**‡	**0.717**†	0.545*
WebFAQ:ita	**0.907**‡	**0.856**‡	0.584*	0.055	**0.816**‡	**0.792**‡	0.446	0.322
WebFAQ:jpn	**0.864**‡	**0.845**‡	0.466	0.346	**0.827**‡	**0.800**‡	0.472	0.340
WebFAQ:ukr	**0.942**‡	**0.866**‡	**0.746**‡	-0.100	**0.909**‡	0.682†	**0.825**‡	0.128
WebFAQ:vie	**0.936**‡	**0.845**‡	**0.753**‡	-0.000	**0.903**‡	**0.802**‡	**0.730**‡	0.309
WebFAQ:zho	**0.891**‡	**0.851**‡	0.570*	0.318	**0.841**‡	**0.835**‡	0.451	0.419

Table 12: Results for cosine similarity type "margin": correlation with MTEB-evaluation scores with our neighborhood retention (Ret_k), cosine similarity (cos), partial correlation of Ret_k controlled with cosine similarity ($\text{Ret}_k|\text{cos}$), and partial correlation of cosine similarity controlled with Ret_k ($\text{cos}|\text{Ret}_k$) for both unprompted and prompted setups. Margin cosine similarity achieves some high and significant correlation coefficients, but in general, Ret_k achieves higher results. Again, $\text{Ret}_k|\text{cos}$ values are high and $\text{cos}|\text{Ret}_k$ are low, meaning the observed correlation in Ret_k is not attributable to margin cosine similarity.

cos: margin	Unprompted				Prompted			
Dataset	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$
ARCChallenge	**0.890**‡	0.344	**0.878**‡	-0.165	**0.945**‡	0.529†	**0.922**‡	0.046
SummEval	0.643‡	0.303	0.620†	0.217	0.588†	0.292	0.575†	0.255
RTE3-multi:en	-0.230	-0.157	-0.185	-0.075	-0.058	-0.100	-0.032	-0.088
RTE3-multi:fr	0.149	0.310	-0.071	0.283	0.268	0.393	0.051	0.303
RTE3-multi:it	0.389	0.344	0.236	0.136	0.488*	0.411*	0.326	0.160
RTE3-multi:de	0.025	0.215	-0.106	0.238	0.162	0.261	0.017	0.207
Tatoeba:deu-eng	**0.888**‡	0.671‡	**0.793**‡	0.197	**0.904**‡	0.696‡	**0.804**‡	0.069
Tatoeba:bul-eng	**0.962**‡	**0.740**‡	**0.918**‡	0.183	**0.928**‡	**0.758**‡	**0.833**‡	0.262
Tatoeba:ukr-eng	**0.939**‡	0.686‡	**0.884**‡	-0.119	**0.918**‡	**0.721**‡	**0.821**‡	0.034
Tatoeba:ita-eng	**0.917**‡	0.648‡	**0.852**‡	-0.034	**0.884**‡	0.682‡	**0.774**‡	0.122
Tatoeba:fra-eng	**0.925**‡	**0.719**‡	**0.862**‡	0.372	**0.895**‡	**0.748**‡	**0.753**‡	0.213
Tatoeba:vie-eng	**0.945**‡	**0.770**‡	**0.867**‡	0.224	**0.916**‡	**0.747**‡	**0.802**‡	0.139
Tatoeba:ell-eng	**0.968**‡	**0.805**‡	**0.917**‡	0.337	**0.922**‡	**0.782**‡	**0.797**‡	0.238
Tatoeba:jpn-eng	**0.968**‡	**0.715**‡	**0.933**‡	-0.022	**0.939**‡	**0.744**‡	**0.859**‡	-0.066
Tatoeba:hin-eng	**0.931**‡	**0.730**‡	**0.869**‡	0.372	**0.876**‡	**0.723**‡	**0.728**‡	0.183
Tatoeba:fin-eng	**0.968**‡	**0.779**‡	**0.916**‡	0.048	**0.959**‡	**0.772**‡	**0.897**‡	0.112
WebFAQ:bul	**0.926**‡	0.635†	**0.872**‡	-0.023	**0.866**‡	**0.808**‡	0.665†	0.475
WebFAQ:deu	**0.889**‡	0.236	**0.883**‡	-0.125	**0.856**‡	0.550*	**0.788**‡	-0.124
WebFAQ:ell	**0.946**‡	0.560*	**0.921**‡	-0.020	**0.930**‡	**0.703**†	**0.869**‡	0.293
WebFAQ:eng	0.620†	0.003	0.645†	-0.224	0.544*	0.055	0.563*	-0.184
WebFAQ:fas	**0.934**‡	0.682†	**0.880**‡	-0.232	**0.893**‡	**0.794**‡	0.682†	0.162
WebFAQ:fin	**0.953**‡	0.449	**0.950**‡	-0.404	**0.913**‡	0.641†	**0.852**‡	-0.157
WebFAQ:fra	**0.878**‡	0.296	**0.875**‡	-0.257	**0.843**‡	0.583*	**0.750**‡	-0.011
WebFAQ:hin	**0.909**‡	0.616†	**0.863**‡	-0.295	**0.891**‡	**0.792**‡	0.687†	0.216
WebFAQ:ita	**0.907**‡	0.201	**0.907**‡	-0.192	**0.816**‡	0.492*	**0.751**‡	-0.106
WebFAQ:jpn	**0.864**‡	0.558*	**0.795**‡	0.048	**0.827**‡	0.666†	0.664†	0.126
WebFAQ:ukr	**0.942**‡	**0.738**‡	**0.870**‡	0.106	**0.909**‡	**0.777**‡	**0.753**‡	0.109
WebFAQ:vie	**0.936**‡	0.397	**0.926**‡	-0.161	**0.903**‡	0.637†	**0.831**‡	-0.056
WebFAQ:zho	**0.891**‡	0.550*	**0.839**‡	-0.065	**0.841**‡	0.686†	0.671†	0.072

Table 13: Results for cosine similarity type "ratio": correlation with MTEB-evaluation scores with our neighborhood retention (Ret_k), cosine similarity (cos), partial correlation of Ret_k controlled with cosine similarity ($\text{Ret}_k|\text{cos}$), and partial correlation of cosine similarity controlled with Ret_k ($\text{cos}|\text{Ret}_k$) for both unprompted and prompted setups. Ratio cosine similarity achieves low to medium correlation, but in general, Ret_k achieves higher results. Again, $\text{Ret}_k|\text{cos}$ values are high and $\text{cos}|\text{Ret}_k$ are low, meaning the observed correlation in Ret_k is not attributable to margin cosine similarity.

cos: ratio	Unprompted				Prompted			
Dataset	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$
ARCChallenge	**0.890**‡	0.098	**0.892**‡	-0.161	**0.945**‡	0.230	**0.942**‡	0.044
SummEval	0.643‡	0.231	0.633‡	0.182	0.588†	0.273	0.595†	0.296
RTE3-multi:en	-0.230	-0.112	-0.209	-0.053	-0.058	-0.082	-0.046	-0.074
RTE3-multi:fr	0.149	0.323	-0.063	0.296	0.268	0.350	0.129	0.266
RTE3-multi:it	0.389	0.312	0.276	0.129	0.488*	0.354	0.391	0.167
RTE3-multi:de	0.025	0.239	-0.091	0.254	0.162	0.282	0.048	0.239
Tatoeba:deu-eng	**0.888**‡	0.385	**0.867**‡	0.050	**0.904**‡	0.431*	**0.881**‡	0.062
Tatoeba:bul-eng	**0.962**‡	0.615†	**0.940**‡	0.148	**0.928**‡	0.655‡	**0.878**‡	0.245
Tatoeba:ukr-eng	**0.939**‡	0.570†	**0.909**‡	-0.036	**0.918**‡	0.677‡	**0.853**‡	0.246
Tatoeba:ita-eng	**0.917**‡	0.285	**0.911**‡	-0.131	**0.884**‡	0.480*	**0.851**‡	0.166
Tatoeba:fra-eng	**0.925**‡	0.313	**0.918**‡	0.121	**0.895**‡	0.645‡	**0.827**‡	0.276
Tatoeba:vie-eng	**0.945**‡	0.595†	**0.915**‡	0.121	**0.916**‡	0.518†	**0.886**‡	0.138
Tatoeba:ell-eng	**0.968**‡	0.663‡	**0.945**‡	0.226	**0.922**‡	**0.745**‡	**0.831**‡	0.297
Tatoeba:jpn-eng	**0.968**‡	0.455*	**0.959**‡	-0.089	**0.939**‡	0.605†	**0.903**‡	-0.048
Tatoeba:hin-eng	**0.931**‡	0.498*	**0.909**‡	0.130	**0.876**‡	0.564†	**0.819**‡	0.181
Tatoeba:fin-eng	**0.968**‡	**0.724**‡	**0.931**‡	0.050	**0.959**‡	0.525†	**0.945**‡	0.160
WebFAQ:bul	**0.926**‡	0.589*	**0.884**‡	-0.040	**0.866**‡	0.672†	**0.795**‡	0.440
WebFAQ:deu	**0.889**‡	0.158	**0.889**‡	-0.167	**0.856**‡	0.201	**0.852**‡	-0.146
WebFAQ:ell	**0.946**‡	0.513*	**0.926**‡	-0.028	**0.930**‡	0.591†	**0.898**‡	0.272
WebFAQ:eng	0.620†	0.119	0.615†	-0.066	0.544*	-0.127	0.547*	-0.145
WebFAQ:fas	**0.934**‡	0.552*	**0.916**‡	-0.358	**0.893**‡	0.585*	**0.831**‡	0.017
WebFAQ:fin	**0.953**‡	0.267	**0.959**‡	-0.435	**0.913**‡	0.366	**0.904**‡	-0.218
WebFAQ:fra	**0.878**‡	0.243	**0.879**‡	-0.256	**0.843**‡	0.273	**0.830**‡	-0.071
WebFAQ:hin	**0.909**‡	0.540*	**0.893**‡	-0.410	**0.891**‡	0.616†	**0.817**‡	0.020
WebFAQ:ita	**0.907**‡	0.096	**0.911**‡	-0.229	**0.816**‡	0.148	**0.818**‡	-0.168
WebFAQ:jpn	**0.864**‡	0.424	**0.832**‡	-0.073	**0.827**‡	0.449	**0.777**‡	0.017
WebFAQ:ukr	**0.942**‡	0.593†	**0.913**‡	0.202	**0.909**‡	0.680†	**0.824**‡	0.078
WebFAQ:vie	**0.936**‡	0.364	**0.929**‡	-0.190	**0.903**‡	0.416	**0.882**‡	-0.061
WebFAQ:zho	**0.891**‡	0.401	**0.872**‡	-0.172	**0.841**‡	0.426	**0.802**‡	-0.038

Table 14: Results for cosine similarity type "z-score": correlation with MTEB-evaluation scores with our neighborhood retention (Ret_k), cosine similarity (cos), partial correlation of Ret_k controlled with cosine similarity ($\text{Ret}_k|\text{cos}$), and partial correlation of cosine similarity controlled with Ret_k ($\text{cos}|\text{Ret}_k$) for both unprompted and prompted setups. Z-score cosine similarity, our most structural cosine similarity baseline, achieves high and significant correlation coefficients, comparable to those with Ret_k. In this case, we can see dataset-specific variations in the partial correlations.

cos: z-score	Unprompted				Prompted			
Dataset	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$	Ret_k	cos	$\text{Ret}_k\|\text{cos}$	$\text{cos}\|\text{Ret}_k$
ARCChallenge	**0.890**‡	**0.863**‡	0.510*	0.303	**0.945**‡	**0.922**‡	0.607†	0.340
SummEval	0.643‡	0.588†	0.455*	0.341	0.588†	0.574†	0.388	0.361
RTE3-multi:en	-0.230	0.216	-0.476*	0.471*	-0.058	0.441*	-0.393	0.562†
RTE3-multi:fr	0.149	0.432*	-0.327	0.507*	0.268	0.652‡	-0.238	0.645‡
RTE3-multi:it	0.389	0.533†	-0.136	0.415*	0.488*	**0.761**‡	-0.287	**0.702**‡
RTE3-multi:de	0.025	0.339	-0.411*	0.514*	0.162	0.558†	-0.421*	0.647‡
Tatoeba:deu-eng	**0.888**‡	**0.834**‡	0.572†	0.191	**0.904**‡	**0.804**‡	**0.707**‡	-0.191
Tatoeba:bul-eng	**0.962**‡	**0.869**‡	**0.843**‡	0.214	**0.928**‡	**0.825**‡	**0.750**‡	0.011
Tatoeba:ukr-eng	**0.939**‡	**0.887**‡	0.672‡	0.094	**0.918**‡	**0.863**‡	0.645‡	0.221
Tatoeba:ita-eng	**0.917**‡	**0.897**‡	0.585†	0.446*	**0.884**‡	**0.869**‡	0.482*	0.373
Tatoeba:fra-eng	**0.925**‡	**0.857**‡	0.679‡	0.047	**0.895**‡	**0.857**‡	0.520†	0.172
Tatoeba:vie-eng	**0.945**‡	**0.870**‡	**0.759**‡	0.175	**0.916**‡	**0.814**‡	**0.724**‡	-0.024
Tatoeba:ell-eng	**0.968**‡	**0.893**‡	**0.829**‡	0.079	**0.922**‡	**0.826**‡	**0.728**‡	-0.106
Tatoeba:jpn-eng	**0.968**‡	**0.867**‡	**0.867**‡	-0.162	**0.939**‡	**0.837**‡	**0.782**‡	-0.103
Tatoeba:hin-eng	**0.931**‡	**0.867**‡	0.692‡	0.158	**0.876**‡	**0.803**‡	0.587†	0.017
Tatoeba:fin-eng	**0.968**‡	**0.954**‡	**0.774**‡	0.657‡	**0.959**‡	**0.925**‡	**0.729**‡	0.385
WebFAQ:bul	**0.926**‡	**0.973**‡	0.160	**0.800**‡	**0.866**‡	**0.903**‡	0.156	0.530*
WebFAQ:deu	**0.889**‡	**0.831**‡	0.628†	0.331	**0.856**‡	**0.779**‡	0.574*	-0.130
WebFAQ:ell	**0.946**‡	**0.948**‡	0.445	0.478	**0.930**‡	**0.915**‡	0.430	0.158
WebFAQ:eng	0.620†	0.517*	0.415	-0.118	0.544*	0.360	0.508*	-0.288
WebFAQ:fas	**0.934**‡	**0.934**‡	0.349	0.349	**0.893**‡	**0.856**‡	0.496*	0.074
WebFAQ:fin	**0.953**‡	**0.924**‡	0.637†	0.241	**0.913**‡	**0.833**‡	0.681†	-0.096
WebFAQ:fra	**0.878**‡	**0.827**‡	0.550*	0.184	**0.843**‡	**0.812**‡	0.397	0.091
WebFAQ:hin	**0.909**‡	**0.897**‡	0.398	0.223	**0.891**‡	**0.889**‡	0.329	0.304
WebFAQ:ita	**0.907**‡	**0.858**‡	0.575*	0.033	**0.816**‡	**0.765**‡	0.457	-0.122
WebFAQ:jpn	**0.864**‡	**0.870**‡	0.248	0.318	**0.827**‡	**0.812**‡	0.275	0.080
WebFAQ:ukr	**0.942**‡	**0.926**‡	0.479	0.134	**0.909**‡	**0.856**‡	0.595*	-0.036
WebFAQ:vie	**0.936**‡	**0.882**‡	0.676†	-0.174	**0.903**‡	**0.845**‡	0.623†	0.232
WebFAQ:zho	**0.891**‡	**0.876**‡	0.424	0.280	**0.841**‡	**0.812**‡	0.376	-0.036