How Far Do Simple Transformations Translate Across Text Embedding Models?

Sid Ali Hamideche
Orange
sidali.hamideche@orange.com

Louis Adrien Dufrene
Orange
louisadrien.dufrene@orange.com

Quentin Lampin
Orange
quentin.lampin@orange.com

Guillaume Larue
Orange
guillaume.larue@orange.com

arXiv:2608.05980v1 [cs.LG] 6 Aug 2026

Abstract

We investigate whether simple transformations can translate representations across heterogeneous text embedding models. Understanding how independently trained models organize semantic information is an enabler for AI-to-AI latent communication without decoding into human-readable text. Focusing on lightweight translators such as linear mappings, we test the literature hypothesis of latent universality in a realistic text setting beyond simplified benchmarks. Across nine embedding models differing in architecture, pooling strategy, and training objective, we evaluate compatibility using CKA, downstream transfer, fidelity, and retrieval. Simple translators recover meaningful shared structure and support transfer for some compatible pairs, but fail sharply for others. Compatibility depends jointly on architecture, training objective, pooling, and data distribution. Overall, the results show that heterogeneous embedding spaces are not universally related by simple mappings as often suggested in some literature.

1 Introduction

Text embedding models are increasingly used as interchangeable infrastructure for retrieval, classification, and multimodal systems. This trend raises a fundamental question: when two models process the same text, how similar are their representations, and can a simple transformation (e.g., a linear map) relate them?

This question matters for several reasons. Simple translation between the representations of two models could make pretrained components easier to reuse with other models, enable a more modular ecosystem of embedding-based tools and pipelines, and enable stitching scenarios where a head trained on one model reuses translated features from another [1, 2]. It also probes whether independently trained systems share geometric organization [3, 4], and allows easier interpretation of how models represent concepts and relations. Additionally, one of our target applications is agent-to-agent communication beyond natural language, so understanding translation limits is foundational for a longer-term direction: AI agents communicating via latent spaces rather than decoded text, avoiding information loss from projecting continuous states onto discrete tokens [5–7]. The present study characterizes where simple transformations work and where they fail, challenging the notion of universal latent compatibility often stated in the literature [8–15].

In this work, we first evaluate a constrained regime that admits closed-form solutions and strong interpretability: direct linear maps from small anchor sets [14], relative representations [12], and inverse relative projection [16]. These methods not only state that some shared structure exists but also state that a simple, anchor-based alignment is enough to recover it. We then include a more

Preprint.

flexible linear map trained on a larger paired corpus, as an empirical approximate upper bound on what linear transformations can achieve with more supervision [13].

Our study extends prior work by evaluating a heterogeneous collection of models varying in dimension, tokenizer, pooling, normalization, architecture, and training objective, and by combining complementary diagnostics, CKA, downstream transfer, fidelity, and retrieval, rather than relying on a single metric.

Contributions.

- A systematic empirical study of simple translation approaches across heterogeneous text embedding models.
- Multi-diagnostic evaluation covering geometric similarity, retrieval, and downstream transfer, going beyond simple benchmarks such as classification models trained on MNIST or CIFAR.
- Evidence that simple transformations are useful but strongly pair-dependent, with architectural and training explanations for compatibility patterns.

2 Related Work and Translation Methods

We organize the related literature around the two families of simple translators we evaluate, defining each method in the context of the work that proposed it. Throughout, $X \in \mathbb{R}^{d_A \times m}$ and $Y \in \mathbb{R}^{d_B \times m}$ denote matrices whose columns are paired embeddings of m inputs in models A and B.

Direct linear maps. A long line of work has shown that representations of independently trained models can be related, to a non-trivial extent, by an unconstrained linear map. Lenc and Vedaldi [9] first demonstrated this for image models; Bansal et al. [1] revisited the idea as model stitching, using the linear map itself as a similarity probe. More recently, Lähner and Moeller [13] reported that unconstrained linear transforms outperform isometric ones, and Gorbett and Jana [15], along with Chen et al. [2], showed that affine maps largely preserve downstream performance across LLM representations. Concretely, given paired embeddings we seek $T \in \mathbb{R}^{d_B \times d_A}$ minimizing $\|Y - TX\|_F^2$. The closed-form least-squares solution

$$T^* \;=\; YX^+,$$

which we refer to as *Linear (pinv)*, requires only a small set of paired anchors and underlies the formalization of Maiorca et al. [14]. When more paired data is available, the same objective can be optimized by stochastic gradient descent on mini-batches of pairs, which we call *Linear (SGD)* [13] and use as an empirical upper bound on what an unconstrained linear translator can achieve.

Anchor-based and relative representations. A complementary line of work translates by re-expressing each embedding in terms of its relation to a shared anchor set. Moschella et al. [12] introduced *relative representations* (RR), which replace an embedding x with the vector of cosine similarities to m anchors $\{a_1, \ldots, a_m\}$ in the same model,

$$r(x) \;=\; \big[\, \cos(x, a_1),\; \cos(x, a_2),\; \ldots,\; \cos(x, a_m) \,\big]^\top \;\in\; \mathbb{R}^m,$$

producing a coordinate system whose dimensions carry direct semantic readings (each entry measures similarity to a named anchor concept) and that is approximately invariant to latent isometries. This invariance enables zero-shot stitching, but it constrains the downstream head to be trained on the relative representation rather than on a native embedding space, narrowing the applicability of the method. Inverse Relative Projection (IRP) [16] lifts this constraint by learning a map from the relative space back into a target model's usable embedding space; Cannistraci et al. [17] further incorporate product invariances to strengthen latent communication. Section 4 makes precise the sense in which RR, IRP, and *Linear (pinv)* are variations on the same anchor-based template.

Representational similarity and latent communication. Beyond the translators themselves, a parallel literature quantifies how similar two representation spaces are without explicitly aligning them. CKA [4] and related measures [10] score geometric alignment across models, and the Platonic Representation Hypothesis [3] conjectures that such similarity grows as models scale. Recent work on latent agent communication [5–7] motivates cross-model translation as the substrate for efficient AI-to-AI interfaces, the longer-term application that frames the present study.

3 Experimental Setup

We evaluate nine text embedding models spanning encoder-only, T5-style encoders, and a causal-backbone model (Table 1), intentionally diverse in dimension, tokenizer, pooling, normalization, and training objective.

Table 1: Embedding models included in the study.

Model Name	Dimension	Tokenizer	Architecture	Pooling	Normalization by default	Max Length
Qwen3-Embedding-4B	2560	Qwen2TokenizerFast	Qwen3ForCausalLM	last token	Yes	40960
e5-large-v2	1024	BertTokenizerFast	BertModel	mean tokens	Yes	512
mxbai-embed-large-v1	1024	BertTokenizerFast	BertModel	cls token	No	512
all-roberta-large-v1	1024	RobertaTokenizerFast	RobertaForMaskedLM	mean tokens	Yes	256
bge-m3	1024	XLMRobertaTokenizerFast	XLMRobertaModel	cls token	Yes	8192
gtr-t5-large	768	T5TokenizerFast	T5EncoderModel	mean tokens	Yes	512
instructor-xl	768	T5TokenizerFast	T5EncoderModel	mean tokens	Yes	512
all-mpnet-base-v2	768	MPNetTokenizerFast	MPNetForMaskedLM	mean tokens	Yes	384
nomic-embed-text-v1.5	768	BertTokenizerFast	NomicBertModel	mean tokens	No	8192

As for the data, for anchor-based methods we use the Basic English 850 word list, compact, human-inspectable, and semantically broad which can be motivated by using human expertise to select anchors. We also use 850 sentences from Wikitext-103, selected by computing the most distant sentences from the others (average across models), in terms of cosine similarity of their embeddings. We found no consistent difference in performance between them, nor significant gains from more elaborate anchor selection, so we keep the pipeline simple. However, we admit that we might not have pushed this step as far as the authors [12, 14, 16] of the original papers did. However, the focus of this study is not anchor selection but rather to show whether there exists a universal transformation between the embeddings of any pair of models. As for *Linear (SGD)* we use Wikitext-103 sentence pairs to train a linear map for each model pair. This can be seen as an approximate upper bound on what a linear map can achieve and compensates for not having further refined the anchor selection step for the anchor-based methods.

Evaluation uses four diagnostics: (1) CKA for geometric similarity, (2) downstream classification transfer, (3) representation fidelity (cosine similarity and normalized error norm on held-out words), and (4) k-NN retrieval. The downstream benchmark includes SST-2, AG News, emotion, CoLA, DBpedia, IMDB, MRPC, and QQP [18, 19], evaluated with a fixed RBF-kernel SVM. Hyperparameter tuning and model selection to maximize per-task performance were not the focus of this work; thus, they were kept minimal. We compare translators under a controlled interface rather than maximize per-model scores.

4 Local Semantic Linear Structure

Before evaluating cross-model transfer, it is worth recalling why a linear translator is a reasonable starting point at all. Word embeddings have long been observed to encode certain semantic relations as approximately linear directions: the canonical *king–queen–man–woman* parallelogram of Mikolov et al. [20, 21], the bilingual lexicon induction of Lample et al. [22], and the formal account of Allen and Hospedales [23] all rest on the same premise: that meaning composes, locally, as linear combinations of a small set of reference points. Replicating their analogy-style reconstruction across our nine heterogeneous text encoders, we observe that this property does carry over: sign patterns and coarse weightings of the decomposition remain stable across architectures, even though precise coefficients depend on each model's normalization and scale. We defer the empirical analysis to Appendix A and use the rest of this section to make explicit why this property is exactly the geometric prerequisite that makes anchor-based linear translation well defined.

Let $X \in \mathbb{R}^{d_A \times m}$ and $Y \in \mathbb{R}^{d_B \times m}$ collect the embeddings of m shared anchors in models A and B. Reconstructing a source vector $x \in \mathbb{R}^{d_A}$ as a least-squares combination of the source-side anchors yields coefficients

$$c^*(x) \;=\; \arg\min_{c \in \mathbb{R}^m} \; \|x - Xc\|_2^2 \;=\; X^+ x,$$

which are nothing more than the multi-anchor generalization of the analogy coefficients above: *king* $\approx \alpha$ *queen* $+ \beta$ *man* $+ \gamma$ *woman* is the same equation written with three anchors instead of m. Transporting the coefficients to the target side, on the assumption that the local linear geometry is

preserved, gives

$$\hat{y} \;=\; Y\,c^{*}(x) \;=\; YX^{+}x \;=\; T_{A\to B}\,x, \qquad T_{A\to B} \;=\; YX^{+},$$

which is exactly the closed-form *Linear (pinv)* estimator of Maiorca et al. [14]. Coefficient-based translation through shared anchors and direct pseudo-inverse alignment are therefore not two different algorithms but two readings of the same operation: one expresses the source point in an anchor basis, the other applies that same expression on the target side.

Two consequences follow. First, *Linear (pinv)* can only succeed to the extent that the anchor-basis decomposition $c^{*}(x)$ is approximately preserved when crossing from A to B, exactly the property the analogy experiment probes, and exactly the property that becomes increasingly fragile as the basis grows or as the two models differ in normalization (Appendix A). Second, the other anchor-based methods we evaluate are variations on the same theme. Relative representations [12] replace the affine coefficients $X^{+}x$ with the cosine-similarity vector to the anchors, trading reconstruction fidelity for invariance to latent isometries; inverse relative projection [16] then learns the analogue of Y from the relative space back to a usable target embedding. All three methods inherit the same implicit assumption, that the local linear geometry around a small set of shared anchors is shared across models, which the rest of the paper puts to the test.

5 Global Alignment and Transfer

5.1 Global Similarity Through CKA

We compute linear CKA [4] over anchor embeddings: $\mathrm{CKA}(X,Y) = \|Y^{\top}X\|_F^2/(\|X^{\top}X\|_F \cdot \|Y^{\top}Y\|_F)$. This metric is often used in literature to compare between the embeddings and is invariant to orthogonal transformations and isotropic scaling.

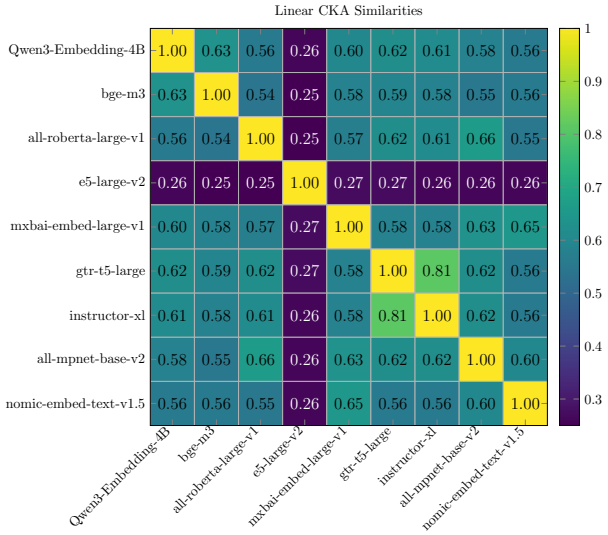


Figure 1: CKA similarity over Basic English anchors.

Figure 1 reveals a structured, non-uniform compatibility pattern. `gtr-t5-large` and `instructor-xl` form the most similar pair: both use T5 encoder backbones with mean pooling, and contrastive fine-tuning on similar datasets, their near-identical training pipeline yields similar geometry. `all-mpnet-base-v2` and `all-roberta-large-v1` show moderate CKA with each other despite different base dimension. `e5-large-v2` is a pronounced outlier, despite using BERT with mean pooling, properties shared by several high-CKA models. Its pretraining, as stated by its authors [24], produces a geometry that diverges from the rest (cosine similarity in the range [0.7, 1] instead of [-1, 1], affecting a lot of angle-sensitive metrics like CKA). This observation already suggests that the universality up to a simple transformation might be too strong. It only needs one model to diverge in one aspect to question the universality. Note, however, that CKA is not invariant to arbitrary linear transformations.

5.2 Downstream Task Transfer

Table 2 and Figure 2 together reveal both the aggregate patterns and the pairwise structure of cross-model transfer. Three observations stand out. Table 2 shows both the average of cross-model accuracy (9x9) and the average of the same-model accuracy (the 9 models where the translation is $T_{i \to i}$). Note that this is different from not applying the translation at all; even for the same model we applied the lossy translation (That is why the results are different according to the translation method, in the table).

Linear (SGD) achieves the highest cross-model accuracy, which is expected given that it is not dependent on selecting the most optimal anchors. Among anchor-based methods, IRP consistently outperforms *Linear (pinv)* and RR, despite the methods being really close in their logic. This suggests that the anchor pruning step presented in IRP model is effective which could have been applied to the 2 other methods as well, but we tried to apply the methods as presented in their original papers, with the only difference in selecting the initial anchor set. This confirms that the choice of anchors can be improved, however as we also trained *Linear (SGD)* with a large paired corpus, we see approximately what other methods can achieve.

Table 2: Average accuracy and macro-F1: cross-model pairs / same-model $T_{i \to i}$. *Linear (SGD)* uses more paired training data than anchor-based methods and serves as a linear approximate upper bound, not a direct comparison.

	Linear-alignment		RelativeRepresentations		IRP		Learned-Linear	
Task	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
sst2	70.16/88.75	66.39/88.56	66.73/**91.73**	60.29/**91.67**	74.64/91.09	72.05/91.04	**84.97**/90.92	**84.49**/90.83
AG News	52.45/85.15	46.82/84.60	33.37/77.58	23.97/76.51	57.95/87.34	53.65/86.91	**83.48/89.63**	**83.06/89.63**
emotion	28.52/57.21	24.22/56.10	24.28/52.54	16.02/50.34	32.36/58.19	29.35/58.32	**43.74/59.72**	**42.27/58.67**
cola	63.69/73.93	59.12/72.16	55.33/57.69	41.46/54.13	64.07/75.08	60.38/72.16	**68.04/75.71**	**64.69/75.04**
dbpedia	44.02/88.19	40.88/87.24	17.62/73.51	12.84/72.36	50.90/94.59	48.16/94.25	**89.11/96.97**	**88.72/96.89**
imdb	60.46/80.12	52.66/79.16	53.89/75.90	43.83/74.11	63.65/**81.39**	58.36/**81.02**	**73.72**/80.67	**71.67**/79.87
mrpc	51.74/57.04	43.76/54.36	52.62/56.24	39.89/52.59	51.93/57.36	45.63/54.33	**55.08/58.09**	**50.27/55.72**
qqp	52.88/69.69	42.87/67.54	53.82/69.86	43.56/67.23	52.55/64.07	44.81/61.24	**54.47/72.55**	**46.85/70.26**

The pairwise matrices in Figure 2 show that performance is not uniform. Compatible clusters align with shared architectural and training choices: `gtr-t5-large`–`instructor-xl` transfer well bidirectionally due to matching T5 backbones. `e5-large-v2` is consistently a poor source and target, but not necessarily always the least compatible, as for some tasks/some translation methods, there are worse performing pairs. `Qwen3-Embedding-4B`, despite its causal decoder backbone, which might be the most different from the rest, still shows moderate compatibility. This suggests that architectural difference alone is not a sufficient condition for compatibility.

Overall, the cross-model performance (outside the diagonal) is often lower than the same-model performance (diagonal), and sometimes much lower for certain pairs/tasks. This questions the universality hypothesis, and the suggestion that the representations of different models, regardless of their architecture, training objective, pooling strategy, etc., are similar up to a simple transformation (e.g., linear).

5.3 Representation Fidelity After Translation

We compare translated and original target embeddings for 1000 common English words held out from the anchor set, reporting cosine similarity and normalized error norm for *Linear (pinv)* and *Linear (SGD)*. Figure 3 presents the mean similarity/distance across the 1000 words for each model pair.

The fidelity matrices largely mirror the CKA and downstream transfer patterns. High cosine similarity and low normalized error appear for the T5-family pair; low fidelity appears for pairs involving `e5-large-v2` (note that the cosine similarity target in `e5-large-v2` column is [0.7, 1], so it might appear higher than the others, but it is not, and some translations are even out of that range, making the translation further different from the original).

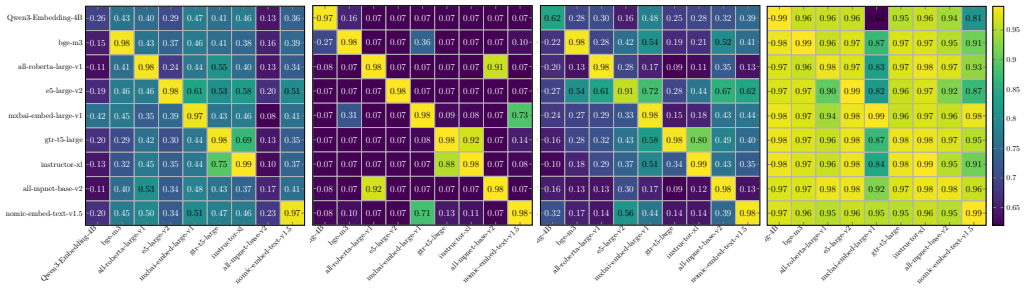
The correspondence between fidelity and downstream transfer exists but does not explain everything. Indeed, evaluating fidelity might be more important, as it shows the imperfection even more. As for some of the evaluated tasks, they might not need the exact translation and can handle the small

(a) AG News



Linear (pinv) RR IRP Linear (SGD)

(b) DBpedia



Linear (pinv) RR IRP Linear (SGD)

Figure 2: Task-transfer accuracy matrices for AG News and DBpedia (rows: source models; columns: target models).

error margin, but for other tasks, this error might have a bigger impact. For example for high prompt adherent image generation tasks, or text-to-text generation tasks that need an exact answer not an approximation. Here, even with the linear map trained via SGD, the translation is far from perfect. This further calls the universality hypothesis into question.

5.4 k-NN Retrieval

We evaluate top-k retrieval ($k \in \{1, 5, 10\}$) on 1000 common English words using Basic English word anchors (Figure 4); sentence-anchor results for Wikitext-103 are provided in Appendix B.

The retrieval matrices echo the downstream transfer pattern: T5-family retrieve well; `e5-large-v2` remain outliers. Retrieval and transfer do not always agree: k-NN measures average neighborhood geometry while classification depends on preserving decision boundaries. Overall, retrieval with k-NN shows similar patterns to other downstream tasks in terms of model compatibility. However, k-NN retrieval might not be the best way to evaluate translation quality as it might suffer from the hubness problem, where some points (hubs) appear as nearest neighbors very frequently to other points [25] (refer to Appendix D for deep hubness analysis in the embeddings of the tested models). Nevertheless, it is still worth evaluating to understand which models are compatible and which architectural or training differences make them incompatible.

6 Model Compatibility

To summarize pairwise behavior, we define a symmetric compatibility score $\mathrm{Compat}(A, B) = \frac{1}{2}(\mathrm{Acc}_{A\to B} + \mathrm{Acc}_{B\to A})$, the average accuracy in both translation directions across all methods and tasks between models A and B.

Figure 5 crystallizes the compatibility structure. The two tightest clusters, `gtr-t5-large/instructor-xl` and `all-mpnet-base-v2/all-roberta-large-v1`, are

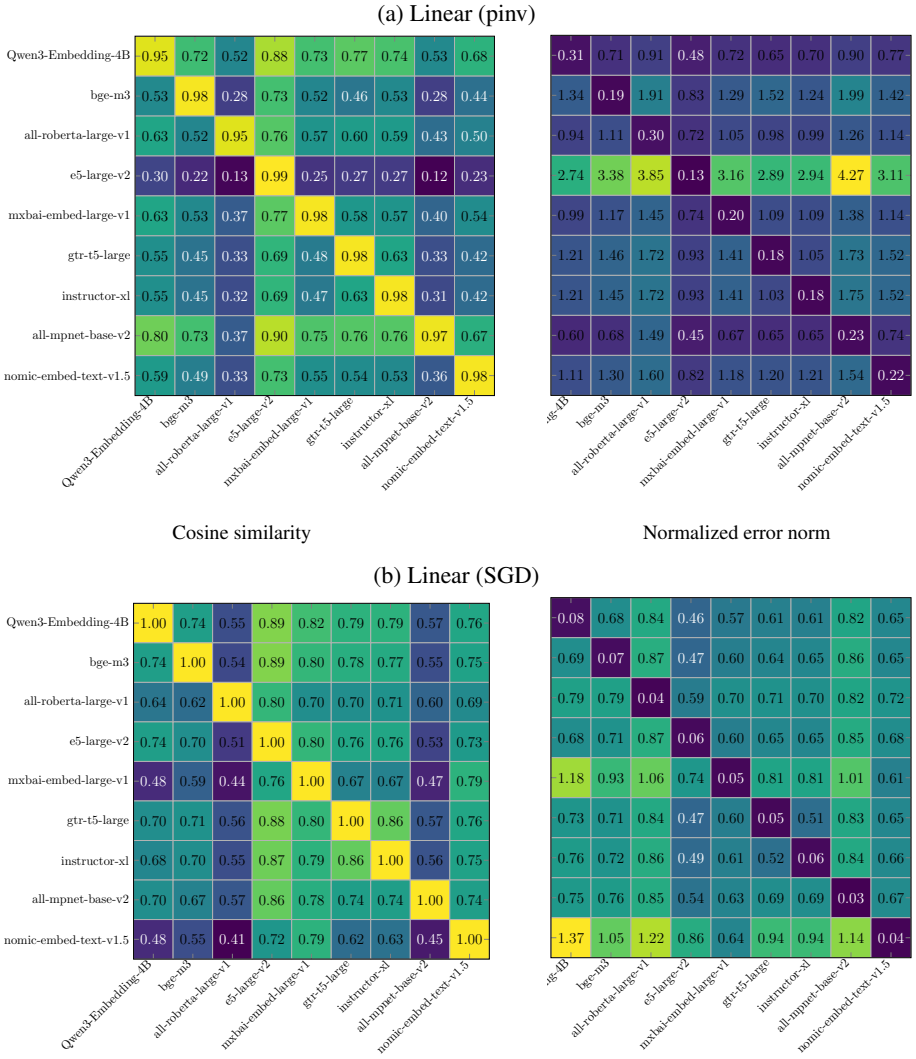


Figure 3: Embedding fidelity on held-out words. T5-family pairs show the highest cosine similarity.

explained by shared inductive biases: same encoder family, same pooling (mean), and similar contrastive training. Models sharing two or more of these properties cluster together; those differing on training objective, pooling, or normalization cluster farther apart.

Two exceptions are informative. `Qwen3-Embedding-4B` (causal decoder, last-token pooling, 2560-dim) clusters moderately close to the encoder group, closer than `e5-large-v2`, a standard BERT model. Its contrastive fine-tuning produces a semantically organized space broadly compatible with encoder-style models: training objective overrides architectural differences. `e5-large-v2`, conversely, shares architecture, pooling, and normalization with `all-roberta-large-v1` yet is the most isolated node, due to its E5 pretraining specificities [24]. Compatibility is shaped by the combination of architecture, training objective, data distribution, and pooling strategy.

The fact that some models are more compatible with each other than with others, and that this compatibility varies with the task and translation method, further emphasizes that the representations of different models are not as similar as suggested.

(a) RR



Top-1 Top-5 Top-10

(b) Linear (pinv)



Top-1 Top-5 Top-10

Figure 4: k-NN retrieval on 1000 common words (Basic English anchors). The compatibility clusters visible in downstream transfer, T5-family, BERT mean-pooling, recur here. `e5-large-v2` and `mxbai-embed-large-v1` show degraded retrieval as both source and target.

7 Discussion

Simple transformations explain a meaningful part of the relationship between text embedding spaces: they recover shared local structure, reveal compatibility clusters, and enable non-trivial downstream transfer. The analysis identifies three levels at which compatibility breaks down. At the *geometric* level, models trained on different data distributions or with different pooling strategies produce spaces that can be related by a single linear map, but only to a certain extent. At the *task* level, even geometrically similar models can fail on fine-grained tasks.

Furthermore, the relatively good cross-model performance of *Linear (SGD)* on certain tasks confirms that the universality hypothesis is not completely untrue. Maybe sticking to very simple translation methods of the linear family is too much of a limitation, and allowing even small non-linearities would allow bridging the gap between more models, and lend more support to the universality hypothesis.

8 Limitations and Broader Impact

Our work has the following limitations: (1) we evaluate only simple transformation families over a finite model set with classification-oriented tasks; (2) the fixed SVM setup keeps comparisons controlled but may understate absolute performance. No uncertainty estimates or compute accounting are reported.

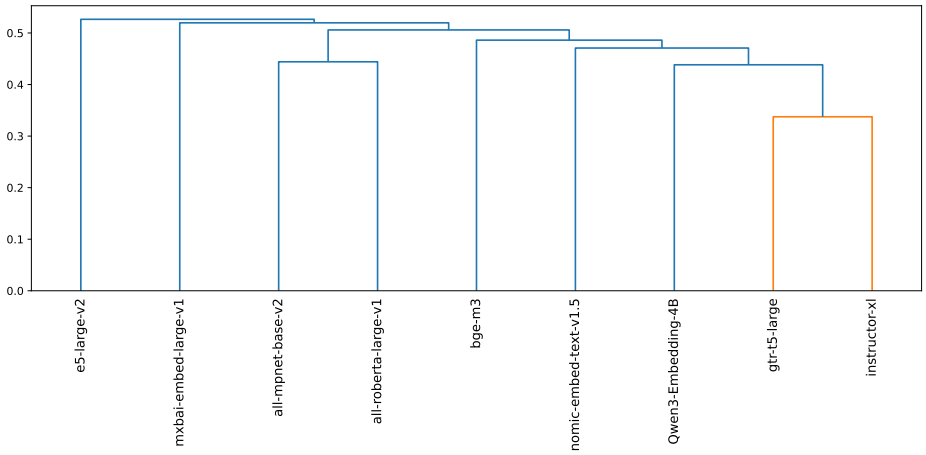


Figure 5: Compatibility dendrogram from bidirectional IRP transfer accuracy.

A more fundamental limitation is the scope of the translation problem itself. All methods evaluated here operate on fixed-length sentence or word embeddings: the map is $\mathbb{R}^{d_A} \to \mathbb{R}^{d_B}$, a single-vector-to-single-vector problem. Many practically important tasks, cross-model attention, token-level generation, sequence-to-sequence reuse, require translating full contextual representations of shape $\mathbb{R}^{l_A \times d_A} \to \mathbb{R}^{l_B \times d_B}$, where l_A and l_B are sequence lengths. When the two models use different tokenizers, $l_A \neq l_B$ in general for the same input, creating a structural misalignment that a position-wise linear map cannot resolve. Even with the same tokenizer, cross-attention-based decoders or self-attention-based text-to-text decoders depend on the full key-value sequence from the encoder rather than on a pooled summary; the geometric structure exploited by our translators simply does not apply in such cases. These sequence-level challenges represent a qualitatively harder translation problem that linear methods are ill-equipped to address, and that future work will need to confront directly.

9 Conclusion

The main conclusion of this work is that simple transformations work, but only in some cases, and the failures are as telling as the successes. Across our nine-model evaluation, cross-model transfer achieves competitive performance for a handful of compatible pairs, those sharing architecture family, training objective, pooling strategy, and data distribution. However, for many pairs—especially those crossing architectural or training boundaries—linear translation degrades substantially. The downstream accuracy and macro-F1 averages are often markedly worse than in-model performance, and the confusion matrices tell a sharper story: off-diagonal degradation is not uniform noise but structured failure, entire class boundaries collapse under translation for incompatible pairs, while staying intact for compatible ones.

These results push back against strong universality hypotheses in the literature. Work on relative representations [12], latent space translation [14, 16], and the Platonic Representation Hypothesis [3] has suggested that independently trained models may converge toward shared representational geometry. Our findings qualify this view: in the controlled setting of sentence and word embeddings, with models deliberately chosen to span diverse architectures and training regimes, the geometry is *not* universally shared. The pairs that transfer well are exceptions explained by specific shared design choices, not evidence of a general convergence. Our results show that, beyond simple settings like MNIST or CIFAR classifiers where universality results are strongest, the hypothesis may not extend to heterogeneous, large-scale text embedding models trained on vastly different corpora.

The conclusion for future work is that more complex translators are needed. Linear maps are useful as diagnostic baselines precisely because they make the cost of incompatibility visible; but bridging genuinely heterogeneous spaces will require nonlinear or learned alignment mechanisms that go beyond what a pseudo-inverse over a small word anchor set can provide.

References

[1] Y. Bansal, P. Nakkiran, and B. Barak. *Revisiting Model Stitching to Compare Neural Representations*. 14, 2021. arXiv: 2106.07682 [cs]. URL: http://arxiv.org/abs/2106.07682 (visited on 05/04/2026).

[2] A. Chen, J. Merullo, A. Stolfo, and E. Pavlick. *Transferring Linear Features Across Language Models With Model Stitching*. 1, 2025. arXiv: 2506.06609 [cs]. URL: http://arxiv.org/abs/2506.06609 (visited on 04/29/2026).

[3] M. Huh, B. Cheung, T. Wang, and P. Isola. *The Platonic Representation Hypothesis*. 25, 2024. arXiv: 2405.07987 [cs]. URL: http://arxiv.org/abs/2405.07987 (visited on 04/29/2026).

[4] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. "Similarity of Neural Network Representations Revisited". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 24, 2019, pp. 3519–3529.

[5] J. Zou et al. *Latent Collaboration in Multi-Agent Systems*. 8, 2025. arXiv: 2511.20639 [cs]. URL: http://arxiv.org/abs/2511.20639 (visited on 05/04/2026).

[6] Z. Du et al. *Enabling Agents to Communicate Entirely in Latent Space*. 16, 2026. arXiv: 2511.09149 [cs]. URL: http://arxiv.org/abs/2511.09149 (visited on 05/04/2026).

[7] R.-J. Zhu et al. *A Survey on Latent Reasoning*. 10, 2025. arXiv: 2507.06203 [cs]. URL: http://arxiv.org/abs/2507.06203 (visited on 05/04/2026).

[8] M. Huh, P. Agrawal, and A. A. Efros. *What Makes ImageNet Good for Transfer Learning?* 10, 2016. arXiv: 1608.08614 [cs]. URL: http://arxiv.org/abs/1608.08614 (visited on 04/29/2026).

[9] K. Lenc and A. Vedaldi. "Understanding Image Representations by Measuring Their Equivariance and Equivalence". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 991–999.

[10] A. Csiszárik, P. K\ Horösi-Szabó, Á. Matszangosz, G. Papp, and D. Varga. "Similarity and Matching of Neural Network Representations". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 5656–5668.

[11] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. *Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax*. 13, 2017. arXiv: 1702.03859 [cs]. URL: http://arxiv.org/abs/1702.03859 (visited on 04/29/2026).

[12] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà. *Relative Representations Enable Zero-Shot Latent Space Communication*. 7, 2023. arXiv: 2209.15430 [cs]. URL: http://arxiv.org/abs/2209.15430 (visited on 04/29/2026).

[13] Z. Lähner and M. Moeller. "On the Direct Alignment of Latent Spaces". In: *Proceedings of UniReps: The First Workshop on Unifying Representations in Neural Models*. Proceedings of UniReps: The First Workshop on Unifying Representations in Neural Models. PMLR, 14, 2024, pp. 158–169.

[14] V. Maiorca, L. Moschella, A. Norelli, M. Fumero, F. Locatello, and E. Rodolà. "Latent Space Translation via Semantic Alignment". In: *Advances in Neural Information Processing Systems* 36 (15, 2023), pp. 55394–55414.

[15] M. Gorbett and S. Jana. *Characterizing Linear Alignment Across Language Models*. Version 3. 26, 2026. arXiv: 2603.18908 [cs]. URL: http://arxiv.org/abs/2603.18908 (visited on 04/29/2026).

[16] V. Maiorca, L. Moschella, M. Fumero, F. Locatello, and E. Rodolà. *Latent Space Translation via Inverse Relative Projection*. 21, 2024. arXiv: 2406.15057 [cs]. URL: http://arxiv.org/abs/2406.15057 (visited on 04/29/2026).

[17] I. Cannistraci, L. Moschella, M. Fumero, V. Maiorca, and E. Rodolà. *From Bricks to Bridges: Product of Invariances to Enhance Latent Space Communication*. 20, 2024. arXiv: 2310.01211 [cs]. URL: http://arxiv.org/abs/2310.01211 (visited on 04/29/2026).

[18] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. "MTEB: Massive Text Embedding Benchmark". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2023. Ed. by A. Vlachos and I. Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 2014–2037.

[19] K. Enevoldsen et al. *MMTEB: Massive Multilingual Text Embedding Benchmark*. 13, 2025. arXiv: 2502.13595 [cs]. URL: http://arxiv.org/abs/2502.13595 (visited on 04/29/2026).

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. 7, 2013. arXiv: 1301.3781 [cs]. URL: http://arxiv.org/abs/1301.3781 (visited on 04/29/2026).

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013.

[22] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou. "Word Translation without Parallel Data". In: International Conference on Learning Representations. 15, 2018.

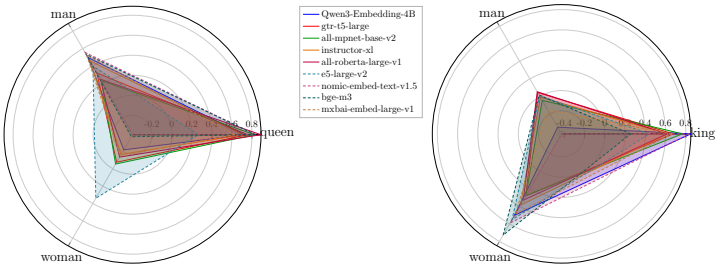
[23] C. Allen and T. Hospedales. "Analogies Explained: Towards Understanding Word Embeddings". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 24, 2019, pp. 223–231.

[24] L. Wang et al. *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. 22, 2024. arXiv: 2212.03533 [cs]. URL: http://arxiv.org/abs/2212.03533 (visited on 04/30/2026).

[25] M. Radovanovi&#263, A. Nanopoulos, and M. Ivanovi. "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data". In: *Journal of Machine Learning Research* 11.86 (2010), pp. 2487–2531.

A Analogy-Style Reconstruction Across Heterogeneous Encoders

This appendix reports the empirical analogy-reconstruction analysis referenced in Section 4. For each model we fit coefficients α, β, γ to minimize $\|\text{king} - (\alpha\,\text{queen} + \beta\,\text{man} + \gamma\,\text{woman})\|_2^2$, and repeat for *queen*. The goal is not to establish identical coefficients but to test whether sign patterns and coarse weightings remain stable across heterogeneous architectures, that is, whether the local linear geometry that Section 4 relies on actually carries over to modern sentence-level encoders.



(a) king $= \alpha$ queen $+ \beta$ man $+ \gamma$ woman. (b) queen $= \alpha$ king $+ \beta$ man $+ \gamma$ woman.

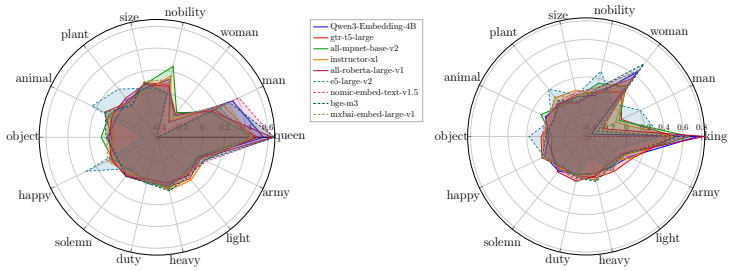
Figure 6: Analogy-style decompositions (3-concept basis). Most models agree on sign pattern and coarse weighting.

Figure 6 shows broad qualitative agreement: *king* receives positive contributions from *queen* and *man* and a negative one from *woman*; *queen* shows the mirrored pattern. This consistency holds across models, suggesting the royalty-gender subspace is robustly linear across architectures, exactly the prerequisite invoked in Section 4 for anchor-based translation to be meaningful.

Enlarging to a 14-concept basis (Figure 7) shows the coarse pattern survives, but coefficient-level agreement becomes less stable, particularly for models with different normalization conventions (`mxbai-embed-large-v1` and `nomic-embed-text-v1.5` versus the normalized majority). Precise coefficient values are sensitive to each model's scale and normalization, foreshadowing the role these factors play in the cross-model compatibility patterns observed in the main paper.

B k-NN Retrieval: Wikitext-103 Sentence Anchors

We report k-NN retrieval results on Wikitext-103 sentence anchors for RR, *Linear (pinv)*, and *Linear (SGD)* in Figures 8–10. The compatibility structure is consistent with the word-anchor results in

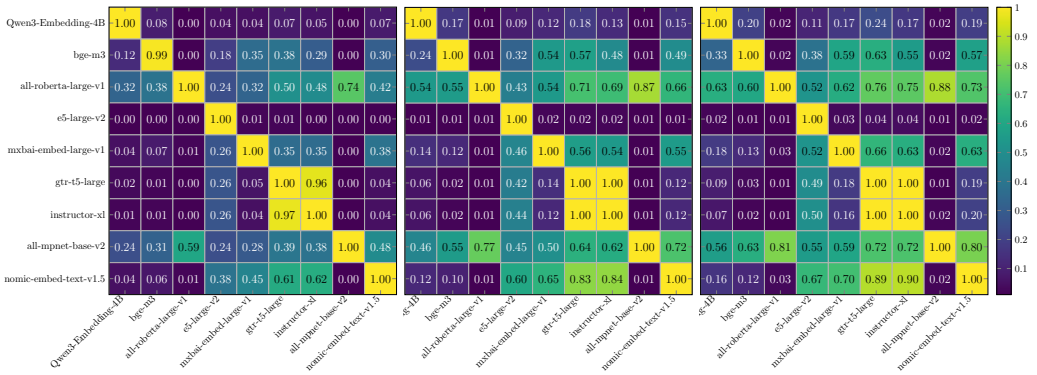


(a) Expanded-basis coefficients for *king*.

(b) Expanded-basis coefficients for *queen*.

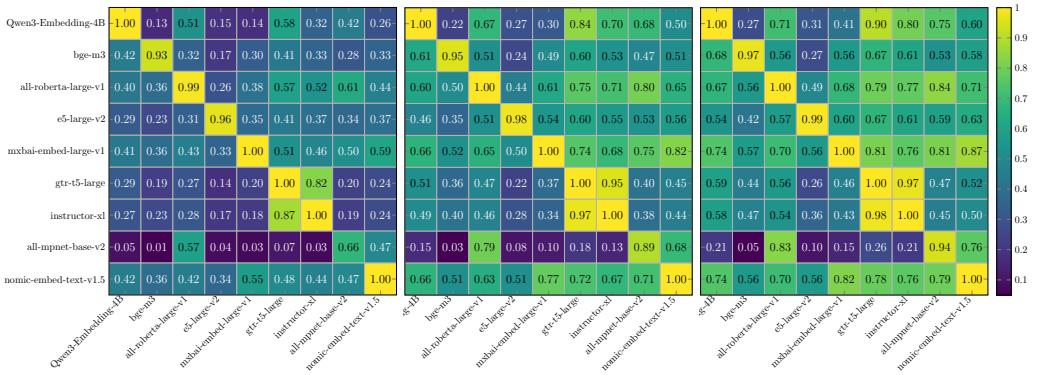
Figure 7: Analogy-style decompositions with a 14-concept basis. Core pattern survives; secondary coefficients become less stable, especially for unnormalized models.

the main paper: T5-family retrieve well; `e5-large-v2` remain outliers. *Linear (SGD)* provides substantially better neighborhood preservation, confirming the upper-bound interpretation, but also suffers in a similar way from model incompatibility.



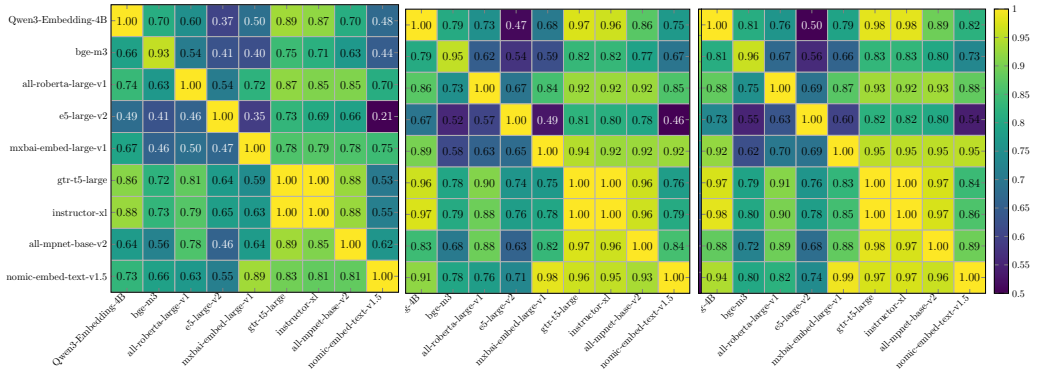
Top-1 Top-5 Top-10

Figure 8: k-NN retrieval: RR, Wikitext-103 sentence anchors.



Top-1 Top-5 Top-10

Figure 9: k-NN retrieval: Linear (pinv), Wikitext-103 sentence anchors.

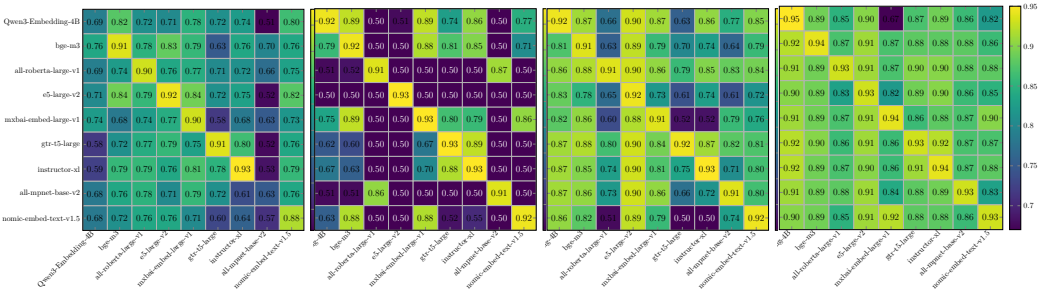


Top-1 Top-5 Top-10

Figure 10: k-NN retrieval: Linear (SGD), Wikitext-103 sentence anchors.

C All Pairwise Accuracy Matrices

Here we report the full set of task-transfer accuracy matrices for all methods and tasks. The main paper shows a subset of these matrices for AG News and DBpedia, but the patterns are consistent across tasks: compatible pairs show high accuracy; incompatible pairs show structured degradation. Furthermore, for some tasks, the performance of some pairs degrades even more. This suggests that these tasks are more sensitive to the translation fidelity, and that the error margin achieved by the translation methods is not sufficient for these tasks, a finding that again calls the universality hypothesis into question.



Linear (pinv) RR IRP Linear (SGD)

Figure 11: Task-transfer accuracy matrices for SST-2 (rows: source models; columns: target models).



Linear (pinv) RR IRP Linear (SGD)

Figure 12: Task-transfer accuracy matrices for Emotion (rows: source models; columns: target models).



Linear (pinv) RR IRP Linear (SGD)

Figure 13: Task-transfer accuracy matrices for Cola (rows: source models; columns: target models).



Linear (pinv) RR IRP Linear (SGD)

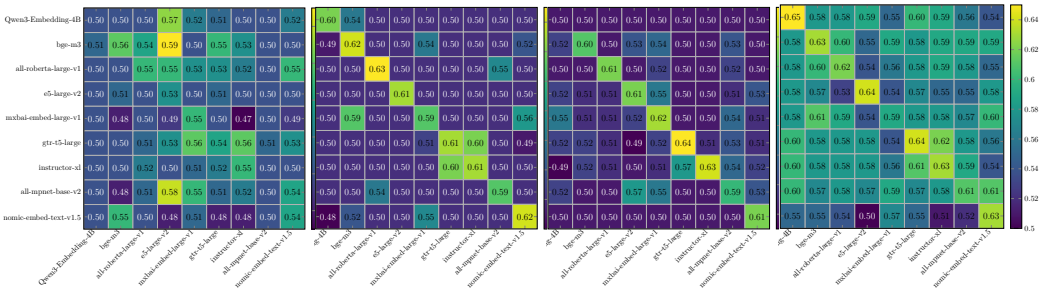
Figure 14: Task-transfer accuracy matrices for IMDB (rows: source models; columns: target models).

D Hubness analysis

Tables 3, 4, and 5 report hubness statistics for the embeddings of 1000 of the most frequent English words. The tables show how many times each word appears as a Top-1/5/10 neighbor of other words for each model. We compute the 95th percentile, which indicates high hubness, and the 10th percentile, which indicates antihubness if low (points that are rarely retrieved as neighbors). We observe that the models with high hubness (high 95th percentile, high skewness, and high variance) are the least performing models as targets for retrieval in k-NN. In particular, `e5-large-v2` has one of the highest hubness indications, and also one of the highest antihubness indications (low 10th percentile and low median), this is a strong indication of why `e5-large-v2` performs poorly in k-NN retrieval. Curiously, `Qwen3-Embedding-4B` also has a high hubness indication, which is consistent with its performance in k-NN as a target model. However, since it did well as a source model, this suggests that hubness cannot explain everything about cross-model compatibility.

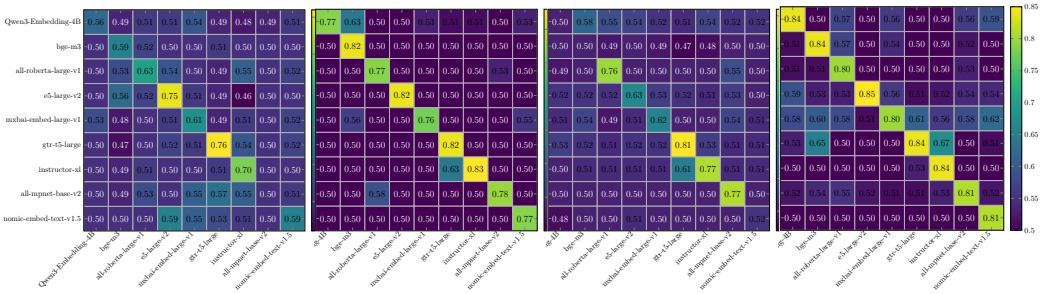
Model	Percentil 95	Percentil 10	Max	Median	Skewness	Variance
Qwen3-Embedding-4B	1.00	1.00	2.0	1.0	0.00	0.03
all-roberta-large-v1	1.00	1.00	3.0	1.0	3.23	0.02
e5-large-v2	1.00	1.00	5.0	1.0	9.66	0.04
bge-m3	1.00	1.00	1.0	1.0	0.00	0.00
mxbai-embed-large-v1	1.00	1.00	1.0	1.0	0.00	0.00
gtr-t5-large	1.00	1.00	1.0	1.0	0.00	0.00
all-mpnet-base-v2	1.00	1.00	1.0	1.0	0.00	0.00
instructor-xl	1.00	1.00	1.0	1.0	0.00	0.00
nomic-embed-text-v1.5	1.00	1.00	1.0	1.0	0.00	0.00

Table 3: Hubness Statistics for Top 1



Linear (pinv) RR IRP Linear (SGD)

Figure 15: Task-transfer accuracy matrices for MRPC (rows: source models; columns: target models).



Linear (pinv) RR IRP Linear (SGD)

Figure 16: Task-transfer accuracy matrices for QQP (rows: source models; columns: target models).

Model	Percentil 95	Percentil 10	Max	Median	Skewness	Variance
Qwen3-Embedding-4B	14.05	1.00	125.0	3.0	7.36	51.24
all-roberta-large-v1	11.00	2.00	108.0	4.0	10.07	29.58
e5-large-v2	18.05	1.00	181.0	2.0	7.64	148.64
bge-m3	12.00	1.00	38.0	4.0	2.88	19.22
mxbai-embed-large-v1	10.00	1.00	329.0	4.0	18.75	161.95
gtr-t5-large	11.00	2.00	33.0	4.0	2.16	11.12
all-mpnet-base-v2	10.00	2.00	23.0	4.0	1.30	8.50
instructor-xl	12.00	2.00	33.0	4.0	2.04	12.27
nomic-embed-text-v1.5	10.00	2.00	111.0	4.0	13.43	20.39

Table 4: Hubness Statistics for Top 5

Model	Percentil 95	Percentil 10	Max	Median	Skewness	Variance
Qwen3-Embedding-4B	31.05	2.00	318.0	5.0	8.41	311.17
all-roberta-large-v1	22.00	3.00	224.0	7.0	8.93	191.79
e5-large-v2	46.00	1.00	324.0	3.0	6.80	702.73
bge-m3	27.00	2.00	124.0	7.0	4.05	123.09
mxbai-embed-large-v1	20.05	2.00	683.0	6.0	13.70	1025.52
gtr-t5-large	23.00	3.00	90.0	8.0	3.41	59.63
all-mpnet-base-v2	20.00	4.00	41.0	9.0	1.46	33.85
instructor-xl	24.00	3.00	102.0	8.0	3.22	68.72
nomic-embed-text-v1.5	22.00	3.00	371.0	8.0	19.95	179.15

Table 5: Hubness Statistics for Top 10