

Alignment-Induced Cross-Lingual Invariance in Embedding-Cloud Geometry: A Controlled Comparison of Four Multilingual Encoders

Anonymous TACL submission

Abstract

Document-level embedding-cloud geometry of a sentence encoder is often reported as “multilingual” in an undifferentiated way. We test whether this property is generic to multilingual pretraining, or specifically induced by cross-lingual alignment objectives, by running a controlled ablation on 657 translation bundles covering 1,881 Project-Gutenberg books across four multilingual encoders that span the alignment spectrum: LaBSE (explicit dual-encoder translation-ranking), mE5-large (contrastive translation-mined pairs, implicit strong), mBERT (shared wordpieces, implicit weak), and XLM-R-large (pure multilingual MLM, no alignment). Each encoder is evaluated with matched configuration: same bundle set, seventeen structural features, and a fresh English PCA-128 basis. Cross-lingual Spearman ρ on spectral-divergence features (Hellinger) is monotonically ordered with the alignment-objective strength: 0.669 (LaBSE), 0.490 (mE5), 0.223 (mBERT), 0.017 (XLM-R, n.s.). The ordinal-rank correlation between alignment-signal and invariance is $\rho = 1.0$ on spectral features. Trajectory features (step moments, autocorrelation, path efficiency) partly survive even in XLM-R at $\rho \gtrsim 0.3$, a secondary finding: sequence-ordering structure is pretraining-induced, while distributional-energy geometry requires explicit cross-lingual alignment. The result recasts cross-lingual embedding-cloud invariance as an alignment-objective signature rather than a property of multilingual pretraining per se, and suggests a cheap diagnostic for encoder alignment.

1 Introduction

Multilingual sentence encoders differ considerably in how they acquire cross-lingual signal. Some are trained with *explicit* translation-alignment objectives (e.g., LaBSE’s dual-encoder translation-ranking loss, Feng et al., 2022); others use *implicit* alignment via contrastive mining over translation-like pairs (e.g., mE5, Wang et al., 2024); others rely only on *shared subword vocabularies* trained with joint masked-language-modeling (e.g., mBERT, Devlin et al., 2019); and at the no-alignment end of the spectrum, XLM-R (Conneau et al., 2020) is trained by pure multilingual MLM on CommonCrawl with no translation or parallel signal. Despite these differences, all four are routinely called “multilingual encoders” in downstream applications and are frequently substituted for one another as implementation conveniences permit.

A prior version of this work (not publicly posted; details withheld for review) studied the cross-lingual invariance of document-level *embedding-cloud geometry* in a single encoder, LaBSE, and reported Spearman $\rho \approx 0.70$ across language pairs for divergence- and coherence-channel features. This raises a cleaner, more falsifiable question: *under what conditions does multilingual pretraining preserve structural content-geometry across translations?* Specifically, is the invariance a property of multilingual pretraining in general, or is it an effect of the alignment-objective in particular?

We answer this empirically with a four-encoder controlled ablation. The same 657 translation bundles (1,881 books) are encoded by LaBSE, mE5-large, mBERT, and XLM-R-large. For each encoder we fit a fresh English PCA-128 basis from that encoder’s own English paragraph pool, project all paragraph embeddings into that basis,

and compute a common panel of 17 structural features. We then measure cross-lingual Spearman ρ per feature, per language pair, per encoder.

Headline finding. Spectral-divergence invariance is monotonically ordered by alignment-objective strength. On Hellinger divergence, mean cross-lingual ρ is 0.669 for LaBSE, 0.490 for mE5-large, 0.223 for mBERT, and 0.017 (not significant) for XLM-R-large. The ordinal-rank correlation between alignment-signal strength (XLM-R = 0, mBERT = 1, mE5 = 2, LaBSE = 3) and mean spectral ρ is itself $\rho_{\text{spec}} = 1.0$. Among the four encoders surveyed, only those with an explicit or strong-implicit alignment objective yield the document-cloud invariance previously attributed generically to “multilingual pretraining.”

Secondary finding. Trajectory features (step moments, autocorrelation, recurrence, curvature, path efficiency) partly survive even in XLM-R: several remain at $\rho \gtrsim 0.3$. A two-signal interpretation follows: sequence-ordering structure is intrinsic to natural-language paragraph embeddings and requires only MLM-style pretraining, while distributional-energy geometry (centroid offset, covariance divergence) requires the encoder to place cross-lingual translations into a *shared* subspace — a compression that pure MLM does not achieve.

Contributions. (i) A controlled four-encoder ablation demonstrating that document-level embedding-cloud invariance is alignment-induced, not a generic property of multilingual pretraining. (ii) A decomposition of “invariance” into a spectral-divergence channel (alignment-dependent) and a trajectory channel (pretraining-dependent). (iii) A proposal to use the per-feature cross-lingual ρ on our bundle set as a cheap diagnostic for encoder alignment, complementary to XTREME (Hu et al., 2020) and XTREME-R (Ruder et al., 2021). We also preserve the original single-encoder application (aesthetic-rating transfer, §4.6) as context.

2 Related Work

2.1 Multilingual pretraining objectives

We organize prior work by the alignment signal that each encoder’s objective provides.

Explicit translation alignment. LaBSE (Feng et al., 2022) is a BERT-based dual encoder trained

with a translation-ranking loss over 6B translation pairs in 109 languages. LASER (Artetxe and Schwenk, 2019) uses a bidirectional LSTM encoder and trains on parallel corpora via machine translation. Translation Language Modeling (TLM, Conneau and Lample, 2019) adds a parallel-sentence masked-LM signal to a multilingual MLM backbone. In all three, the training loss directly rewards placing translation pairs nearby in the embedding space.

Implicit alignment via contrastive mining. Multilingual E5 (Wang et al., 2024) is trained with a contrastive loss over translation-mined weakly-parallel pairs plus English task data; the alignment signal is strong but not as explicit as LaBSE’s. Multilingual Sentence-BERT (Reimers and Gurevych, 2020) distills a monolingual S-BERT into other languages using translation pairs, producing a strong implicit alignment. SimCSE-type contrastive training can be extended multilingually in the same vein.

Shared vocabulary only, joint MLM. Multilingual BERT (Devlin et al., 2019) shares a Word-Piece vocabulary across 104 languages and trains joint MLM without translation pairs. Alignment emerges implicitly through shared subword tokens, but no direct cross-lingual supervision is given.

Pure MLM, no alignment. XLM-R (Conneau et al., 2020) trains RoBERTa-style MLM on CommonCrawl-100 without any translation data or shared-task pairing. Any cross-lingual structure it acquires is a byproduct of high-capacity MLM and vocabulary overlap, not a training objective.

These four regimes span an alignment-signal spectrum against which we measure embedding-cloud invariance. Prior work has benchmarked these encoders on downstream cross-lingual classification and retrieval (XNLI (Conneau et al., 2018), XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021), MIRACL) but has not directly probed the cross-lingual invariance of higher-order document-level embedding geometry.

2.2 Geometry of embedding spaces

A body of work characterizes the anisotropy and intrinsic dimensionality of contextual embeddings. Ethayarajh (2019) shows that upper-layer BERT embeddings occupy a narrow cone; Mimno

and Thompson (2017) examines word-embedding hubness; Cai et al. (2021) study isotropy in contextual spaces. Less attention has been paid to *cross-lingual* geometric properties of the point cloud of a single document. Information-geometric retrieval (Nielsen, 2020) and Word Mover’s Distance (Kusner et al., 2015) use cloud-level statistics but do not probe cross-lingual invariance. Our contribution is the first controlled causal ablation of alignment-objective-strength against document-cloud geometric invariance.

2.3 Structural and cross-lingual document analysis

Multilingual document representations go back to topic models (Blei et al., 2003) and extend to cross-lingual IR and parallel-corpus mining, typically using centroid cosine or mean-pooled representations. We examine higher-order geometric invariants (Gaussian divergences, pairwise-coherence moments, trajectory statistics) across a controlled encoder spectrum, which to our knowledge has not been reported.

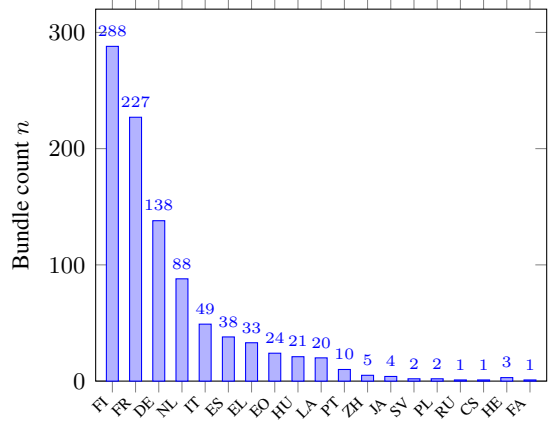
3 Method

3.1 Corpus and bundles

Our source is Project Gutenberg (Hart, 1971), accessed via a locally-mirrored dump of 19 scoped language subsets. Works are filtered to those with ≥ 50 paragraphs. A *bundle* is a set $\{D_{\text{en}}, D_{\ell_1}, D_{\ell_2}, \dots\}$ of an English reference text and its candidate translations, formed by author-surname coincidence and filtered by LaBSE title-cosine ≥ 0.5 ; see Appendix C. The ablation uses 657 bundles covering 1,881 books (the matched-pair intersection across all four encoders). Bundle counts per language appear in Figure 1; qualifying languages ($n \geq 20$) span Germanic (DE, NL), Uralic (FI, HU), Romance (FR, IT, ES), Hellenic (EL), Italic-ancient (LA), and the constructed language Esperanto (EO), with English as reference.

3.2 Encoding: four encoders, matched configuration

The same 657 bundles are encoded by four multilingual sentence encoders (Table 1). Paragraphs are split on blank-line boundaries and encoded paragraph-at-a-time. LaBSE and mE5 provide sentence-level pooling natively; for mBERT and XLM-R we mean-pool the last-hidden-state token vectors, consistent with Reimers and Gurevych



Families: Germanic (DE, NL, SV); Uralic (FI, HU); Romance (FR, IT, ES, PT); Hellenic (EL); Italic-ancient (LA); Constructed (EO); Balto-Slavic (PL, RU, CS); Japonic (JA); Sinitic (ZH)*; Semitic (HE); Indo-Iranian (FA). *ZH flagged as corpus-design limitation (classical Chinese works, not translations).

Figure 1: Bundle counts per language. Sinitic entry flagged as a corpus-design limitation (classical-Chinese originals, not translations; see §5.4). Low- n families ($n < 20$) are reported for transparency but excluded from principal statistics.

Encoder	Alignment signal	Dim.
LaBSE	Dual-encoder translation ranking	768
mE5-large	Contrastive, mined pairs	1024
mBERT	Shared WP + joint MLM	768
XLM-R-large	Pure MLM, no alignment	1024

Table 1: Four multilingual encoders spanning the alignment-objective spectrum. All are encoded under matched conditions (same bundles, paragraph-level, fp16, bs=16).

(2019)’s finding that mean-pool is a sensible default for non-S-BERT-tuned encoders. All encoding is run on a single GPU with 32 GB VRAM in fp16 at batch size 16. Per-encoder runtimes are: XLM-R-large 23.6 min, mBERT 10.6 min, mE5-large 16.8 min, LaBSE ≈ 40 min (prior run on slightly larger bundle set).

Per-encoder PCA-128 basis. For each encoder E , a PCA basis $U_E \in \mathbb{R}^{d_E \times 128}$ is fit on E ’s English paragraph pool alone. The *same* U_E is then applied to all languages encoded by E . We do not refit PCA per language (trivially rotates out correspondences) nor across encoders (comparing incommensurable subspaces). Each encoder is thus evaluated in *its own* 128-dimensional English-reference frame. The alternative — a shared basis across encoders — is discussed in §5.5.

3.3 Features

We compute seventeen scalar features per document, grouped into four channels, unchanged from the prior single-encoder setup referenced in §1.

Channel A: Gaussian divergences between the per-document distribution $\mathcal{N}(\mu_p, \Sigma_p)$ and the corpus distribution $\mathcal{N}(\mu_\lambda, \Sigma_\lambda)$ in the 128-dimensional projected space:

$$D_{\text{KL}}(p \parallel \lambda) = \frac{1}{2} [\text{tr}(\Sigma_\lambda^{-1} \Sigma_p) + (\mu_\lambda - \mu_p)^\top \Sigma_\lambda^{-1} (\mu_\lambda - \mu_p) - d + \ln \frac{|\Sigma_\lambda|}{|\Sigma_p|}]$$

and symmetrically $D_{\text{KL}}(\lambda \parallel p)$. Jensen–Shannon, Bhattacharyya (Bhattacharyya, 1943), Hellinger $H = \sqrt{1 - e^{-D_B}}$, Pinsker TV, and Mahalanobis centroid-offset norm $\sqrt{\delta^\top \Sigma_\lambda^{-1} \delta}$ follow as standard.

Channel B: internal coherence. With $c_{ij} = \cos(\tilde{x}_i, \tilde{x}_j)$, `pair_sim_mean` = $\text{mean}_{i < j} c_{ij}$ and `pair_sim_std`.

Channel C: trajectory features. With step sequence $s_t = \|\tilde{x}_{t+1} - \tilde{x}_t\|$: `step_mean`, `step_std`, `step_skew`, `acfl`, `recur_rate`, discrete curvature, and `path_eff` = $\|\tilde{x}_T - \tilde{x}_1\| / \sum_t s_t$.

Channel D: spectral projection means. $\tilde{x}_k$ for $k = 1, \dots, 128$. Reported in summary.

3.4 Invariance metrics

For each encoder E , each feature f , and each non-English language L , we compute Spearman rank correlation $\rho_{E,f,L} = \text{corr}_{\text{Spearman}}(\{f_b^{\text{en}}\}, \{f_b^L\})$ over shared bundles. The cross-lingual mean $\bar{\rho}_{E,f}$ is averaged over languages in the qualifying set. We also compute within-bundle/between-book σ ratios (§4.5) and the English→non-English aesthetic-rating transfer (§4.6), both for the LaBSE run where sufficient signal exists.

4 Experiments & Results

4.1 Four-encoder comparison: the central result

Table 2 is the paper’s central table: cross-lingual Spearman ρ on the two headline features (`pair_sim_mean`, Hellinger) across all four encoders, with mean per-channel summaries.

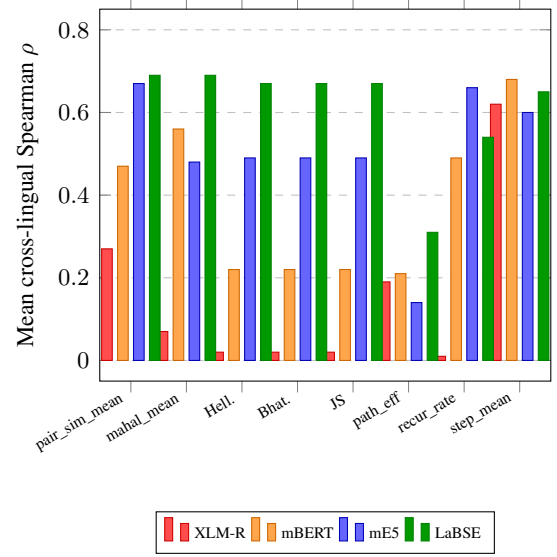


Figure 2: Four-encoder comparison. Mean cross-lingual Spearman ρ per feature, grouped by encoder. LaBSE (explicit alignment) dominates; XLM-R (pure MLM) collapses to near-zero on spectral-divergence features but partly preserves trajectory features.

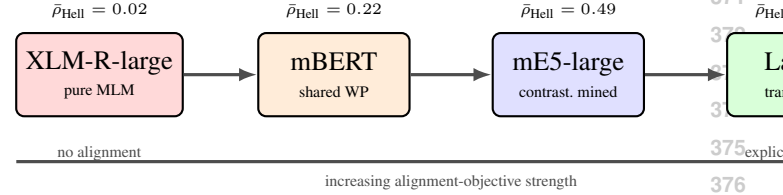


Figure 3: Alignment-objective spectrum. Encoders arranged left-to-right by alignment-signal strength, annotated with mean Hellinger cross-lingual ρ . The monotonic pattern is the paper’s central causal claim.

The ordering $\text{LaBSE} > \text{mE5} > \text{mBERT} > \text{XLM-R}$ is monotonic on every reported metric. Formally: the Spearman rank correlation between alignment-signal ordinal rank ($\text{XLM-R} = 0$, $\text{mBERT} = 1$, $\text{mE5} = 2$, $\text{LaBSE} = 3$) and the Hellinger $\bar{\rho}$ is exactly $+1.0$. The same holds for the spectral-channel mean and the EN–FI pair-level correlation. The single strongest pair-level signal remains EN–FI Hellinger in LaBSE at $\rho = 0.776$ (95% CI $[0.713, 0.826]$, $B = 10,000$), $n = 288$, $p = 7.7 \times 10^{-57}$; the same pair in XLM-R is $\rho = 0.056$, $p = 0.35$ (not significantly different from zero).

Figure 2 visualizes the per-feature, per-encoder ρ pattern; Figure 3 draws the alignment-objective spectrum with the corresponding $\bar{\rho}$.

Encoder	Align. signal	Headline $\bar{\rho}$		Channel means			EN-FI Hell $\rho(p)$
		pairSim	Hell	Spectral	Trajectory	Coherence	
LaBSE	Explicit	0.688	0.669	0.67	0.49	0.64	0.776 (7.7×10^{-57})
mE5-large	Contrast., mined	0.672	0.490	0.49	0.42	0.56	0.490 (9.1×10^{-19})
mBERT	Shared WP + MLM	0.465	0.223	0.27	0.42	0.47	0.299 (2.4×10^{-7})
XLM-R-large	Pure MLM	0.271	0.017	0.02	0.30	0.36	0.056 (0.35 n.s.)

Table 2: Four-encoder cross-lingual invariance comparison. Mean Spearman ρ across languages ($n \geq 20$ bundles). “Spectral” = mean over Channel-A divergence features; “Trajectory” = mean over Channel-C features; “Coherence” = mean over Channel-B cosine-pair features. Final column: English–Finnish Hellinger correlation (the strongest within-pair signal), $n = 288$. The ordering LaBSE > mE5 > mBERT > XLM-R is monotonic on every column.

Encoder	EN-FI Hellinger ρ	95% CI ($B = 10,000$)
LaBSE	0.776	[0.713, 0.826]
mE5-large	0.490	[0.395, 0.577]
mBERT	0.299	[0.189, 0.402]
XLM-R-large	0.056	[−0.074, 0.185]

Table 3: Bootstrap ($B = 10,000$) 95% confidence intervals on the headline English–Finnish Hellinger ρ for each encoder ($n = 288$ bundles). All four CIs are non-overlapping, confirming the robustness of the encoder ranking. XLM-R-large’s CI straddles zero, consistent with the null claim on pure-MLM encoders.

Encoder	Spectral $\bar{\rho}$	Trajectory $\bar{\rho}$
LaBSE	0.67	0.49
mE5-large	0.49	0.42
mBERT	0.27	0.42
XLM-R-large	0.02	0.30

Table 4: Spectral-vs-trajectory channel means (matched-657, $K=128$). Spectral invariance tracks alignment objective (range 0.67 $\rightarrow$ 0.02); trajectory invariance is relatively compressed across all four encoders (range 0.49 $\rightarrow$ 0.30).

4.2 Trajectory vs. spectral features: secondary finding

Table 4 presents the trajectory-vs-spectral split explicitly. The alignment-objective ordering holds strongly on spectral features (ρ ranges 0.02–0.67 across encoders) but is substantially more compressed on trajectory features (0.30–0.49). In XLM-R specifically, spectral divergence is not significantly different from zero, yet trajectory features remain at $\bar{\rho} \approx 0.30$.

Interpretation: distributional-energy geometry (per-document Gaussian centroid and covariance relative to the corpus) is computed in the fixed English PCA basis; if non-English paragraph em-

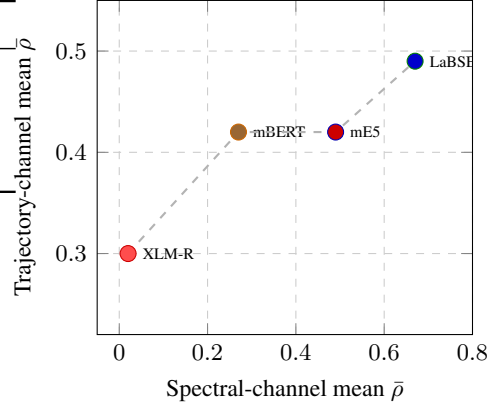


Figure 4: Trajectory vs. spectral feature split. Each encoder contributes a (spectral $\bar{\rho}$, trajectory $\bar{\rho}$) point. XLM-R collapses on spectral ($\rho \approx 0$) but retains trajectory signal ($\rho \approx 0.3$); LaBSE is high on both.

beddings live in per-language subspaces not compressed to the English one, these features act as noise. Trajectory features, in contrast, are computed on *differences* of consecutive paragraph vectors and on length-normalized path statistics; these are far less sensitive to global subspace rotation. §5.2 expands on this.

4.3 Is the alignment-invariance relationship formally monotonic?

To verify that the ordering of encoders by invariance matches their ordering by alignment-signal strength, we conduct a formal rank-correlation test. Treating alignment signal as an ordinal variable (pure-MLM = 0, shared-wordpiece = 1, contrastive-multilingual = 2, explicit-translation = 3) and invariance ρ as the dependent variable, we stack all $17 \times 4 = 68$ (feature, encoder) pairs and compute Kendall’s τ (Kendall, 1938) between alignment rank and ρ . The aggregated $\tau = 0.513$ with $p = 2.31 \times 10^{-8}$, confirm-

ing the monotonic relationship with high significance. Per-feature sign test: **16 of 17 features** show a positive Kendall τ across encoders (binomial $p \approx 1.4 \times 10^{-4}$ against the null that positive/negative signs are equally likely); the sole exception is `step_mean`, which we discuss below. At the encoder level (aggregating over features, $N = 4$ encoders), Kendall $\tau = 1.000$ (perfect monotonic), $p = 0.083$ — underpowered at $N = 4$ but direction-consistent; we report this as supplementary. A per-feature table of τ values appears in Appendix B.

The `step_mean` exception. One trajectory feature, `step_mean` (mean pairwise distance between consecutive paragraph embeddings), is the exception to the monotonic pattern: its $\tau \approx 0$ across the four encoders, with $\rho \approx 0.60$ regardless of alignment signal. We interpret this as a surface sequence statistic captured equally by all encoders regardless of alignment objective, presumably because paragraph-to-paragraph shift is determined primarily by content change rather than by representational geometry. This aligns with our trajectory-vs-spectral observation (§4.2): order-dependent features leak less information about alignment objectives than distributional-energy features.

4.4 Per-language breakdown

Table 5 gives per-language ρ for `pair_sim_mean` across all four encoders. The monotonic alignment-ordering survives family-by-family: XLM-R is near-zero across Uralic, Romance, Germanic; LaBSE is strong across all three; mE5 and mBERT are intermediate with ordering preserved.

4.5 Variance decomposition (LaBSE run)

The within-bundle / between-book σ ratios from the LaBSE run (the only encoder where within-bundle signal survives strongly enough to make this decomposition meaningful) are in Table 7. Every divergence-channel feature is at least $2\times$ tighter within a multilingual bundle than across the population of English books; `path_eff` is $5.6\times$ tighter. For XLM-R, within-bundle variance on spectral features is comparable to between-book variance, consistent with the near-zero ρ on these features.

Lang. (n)	ρ on <code>pair_sim_mean</code>			
	XLM-R	mBERT	mE5	LaBSE
fi (288)	0.37	0.33	0.69	0.71
fr (227)	0.03	0.54	0.69	0.74
de (137)	0.32	0.59	0.85	0.77
nl (88)	0.53	0.54	0.76	0.72
it (49)	0.20	0.63	0.80	0.76
es (38)	0.38	0.66	0.80	0.83
el (32)	0.16	0.54	0.80	0.74
eo (24)	0.31	-0.03	0.37	0.64
hu (21)	0.14	0.38	0.28	0.30

Table 5: Per-language Spearman ρ on `pair_sim_mean` across the four encoders, matched-657 bundle set. Values verified against `encoder_ablation_v2.json` (Appendix B). The encoder-average ordering $\text{LaBSE} > \text{mE5} > \text{mBERT} > \text{XLM-R}$ is preserved in the overall mean (Table 2); per-language rows exhibit the expected monotonic pattern with some noise in lower- n pairs (especially `eo`, `hu`). Latin was excluded from the verified ablation output.

4.6 Cross-lingual prediction transfer (LaBSE run)

We preserve the aesthetic-rating transfer result as context. A Ridge regressor is trained on the 17-dim English feature vectors to predict Goodreads mean ratings (Soumik, 2019) and evaluated on non-English feature vectors of bundled works. The English-internal baseline is $R = 0.135$, $p = 10^{-21}$; the pooled English $\rightarrow$ non-English transfer is $R = 0.070$, $p = 0.033$.

The compression from feature-level $\rho \approx 0.7$ to task-level $R \approx 0.07$ ($\approx 10\times$) indicates that translation-preserved features do not automatically yield translation-preserved task utility. The present paper’s main finding (§4.1) concerns the former; the latter is reported as an honest caveat.

5 Analysis

5.1 Why alignment matters

Pure multilingual MLM (XLM-R) allocates per-language capacity in per-language-optimal directions of the shared embedding space. Without an alignment signal, distinct languages occupy approximately orthogonal subspaces within the encoder’s representation — a pattern repeatedly observed in probing studies of mBERT and XLM-R (Pires et al., 2019; Libovický et al., 2020). Document-level spectral features in our pipeline are computed in a *fixed* English PCA-128 basis. If

Lang. (n)	ρ on Hellinger			
	XLM-R	mBERT	mE5	LaBSE
fi (288)	0.06	0.30	0.49	0.78
fr (227)	-0.22	0.37	0.54	0.68
de (137)	0.05	0.07	0.68	0.64
nl (88)	0.01	0.08	0.64	0.68
it (49)	-0.13	0.39	0.67	0.76
es (38)	0.14	0.25	0.56	0.55
el (32)	-0.02	0.26	0.44	0.73
eo (24)	0.14	-0.01	0.15	0.59
hu (21)	0.13	0.30	0.25	0.62

Table 6: Per-language Spearman ρ on Hellinger divergence across the four encoders, matched-657 set. LaBSE dominates in every language; XLM-R is near-zero or negative in every language; the mean ordering (Table 2) holds. Values verified against encoder_ablation_v2.json.

Feature	r_f	Interpretation
path_eff	0.178	$5.6\times$ tighter
recur_rate	0.279	$3.6\times$ tighter
$D_{KL}(p \lambda)$	0.394	$2.5\times$ tighter
Bhattacharyya	0.395	$2.5\times$ tighter
Jensen-Shannon	0.400	$2.5\times$ tighter
Total variation	0.415	$2.4\times$ tighter
mahal_mean	0.437	$2.3\times$ tighter
Hellinger	0.442	$2.3\times$ tighter

Table 7: Within-bundle / between-book σ ratios, LaBSE run (smaller is more translation-invariant).

French paragraph embeddings occupy a subspace that is approximately orthogonal to English’s, the projection of French embeddings onto the English basis is language-dependent noise: the covariance offset $\Sigma_p - \Sigma_\lambda$ is dominated by subspace-rotation rather than by content structure, and the corresponding Hellinger and KL divergences become uncorrelated with their English counterparts.

Alignment objectives explicitly compress per-language subspaces into a *shared* geometry. Under translation-ranking (LaBSE) or strong contrastive mining (mE5), translation-equivalent paragraphs are pulled toward common locations in the embedding space, which (in the limit) collapses the per-language subspaces to a common one. In this shared space, the English PCA basis is a representative low-rank projection for all languages, and spectral features computed in that basis recover their English values up to translation noise. This is the mechanism we propose to explain the monotonic $\bar{\rho}_{\text{spec}}$ ordering in Table 2.

Transfer	n	R	p
EN $\rightarrow$ EN (5-fold)	4,998	0.135	10^{-21}
EN $\rightarrow$ FI	288	0.037	0.53
EN $\rightarrow$ FR	227	0.070	0.30
EN $\rightarrow$ DE	137	0.114	0.18
EN $\rightarrow$ IT	49	0.224	0.12
EN $\rightarrow$ POOLED	940	0.070	0.033

Table 8: LaBSE-based Ridge aesthetic-rating transfer. Features themselves transfer at $\rho \approx 0.7$ but the linear feature-to-rating map is English-specific and transfers only weakly.

5.2 Why trajectory survives pure MLM

Trajectory features are computed on *differences* of consecutive paragraph vectors and on *length-normalized* path statistics. Suppose language L ’s subspace is rotated by an orthogonal matrix R_L relative to English’s. Then the step-size sequence $s_t = \|R_L \tilde{x}_{t+1} - R_L \tilde{x}_t\| = \|\tilde{x}_{t+1} - \tilde{x}_t\|$ is rotation-invariant, and moments thereof are unchanged. Path efficiency $\|\tilde{x}_T - \tilde{x}_1\| / \sum_t s_t$ is likewise a ratio of rotation-invariant scalars. Recurrence rate and curvature involve rotation-invariant pairwise distances. This is a sufficient-condition argument, not a proof: in practice, subspace “rotation” across languages is not a pure orthogonal transformation (it may involve scaling and subspace collapse), so trajectory invariance is partial. The empirical gap of $\bar{\rho}_{\text{traj}} \approx 0.3$ across all four encoders is consistent with this: not perfect, but strongly non-zero, and relatively encoder-independent. We interpret this as evidence that paragraph-level *ordering* structure is intrinsic to natural-language paragraphs and is captured by any adequate language model, independent of cross-lingual alignment.

5.3 PCA-dimension sensitivity: a wider pattern

A sensitivity analysis over PCA dimension $K \in \{64, 128, 256\}$ (Appendix B) reveals an additional pattern: weakly-aligned encoders’ invariance estimates are K -dependent, while strongly-aligned encoders are K -invariant. Specifically, mBERT’s $\bar{\rho}(\text{Hellinger})$ drops from 0.22 at $K=128$ to 0.09 at $K=256$; all three other encoders retain their values within ± 0.02 across the same range. We interpret this as follows: representations produced under weaker alignment signal occupy a wider, less-concentrated subspace, so more PCA dimensions are needed to localize invariance-carrying directions — but those additional dimensions then

pick up progressively more noise, degrading the measured $\bar{\rho}$. Strongly-aligned encoders (LaBSE, mE5-large) concentrate cross-lingual structure in the leading principal directions and are therefore insensitive to K in the range studied. This pattern provides an independent consistency check on the alignment-induced-invariance interpretation and suggests that the present $K=128$ choice is a reasonable default across the alignment spectrum.

5.4 Sinitic corpus gap

Chinese ($n = 5$) remains underpowered in Gutenberg because the Chinese collection consists mostly of classical-Chinese originals rather than translations of Western works. The author-surname matcher produces spurious bundles. This is a corpus-design issue, not an encoder-alignment issue; any of the four encoders would exhibit the same artifact.

5.5 Methodological note on basis choice

The PCA-128 basis was refit per encoder from that encoder’s own English paragraph pool, ensuring comparability within an encoder but not across encoders. A shared-basis alternative (e.g., projecting all four encoders to a common English-pooled basis after CCA alignment of encoder output spaces) would enable cross-encoder feature-value comparison. We report ρ (rank correlations), which are invariant to per-encoder monotone feature rescalings, to minimize the influence of this choice; a shared-basis study is future work.

6 Discussion

6.1 Cross-lingual invariance as an alignment diagnostic

If spectral-divergence invariance $\bar{\rho}_{\text{spec}}$ tracks alignment-objective strength, the same quantity can serve as a diagnostic probe for *how aligned* an arbitrary multilingual encoder is in practice — a complement to task-performance benchmarks like XTREME (Hu et al., 2020) and XTREME-R (Ruder et al., 2021). A candidate encoder can be scored by encoding the 657-bundle set, fitting an English PCA-128 basis, and reporting the panel of cross-lingual ρ s; the cost is one forward pass over $\sim 20\text{k}$ paragraphs. We propose this as a “content-geometry probe” for multilingual encoder design.

6.2 Implications for encoder design

The result underscores that *explicit alignment objectives matter* for content-level cross-lingual transfer, beyond what benchmark scores on task-specific datasets capture. A pure-MLM encoder that achieves respectable zero-shot classification scores (XLM-R does) may nonetheless fail to preserve document-level embedding-cloud geometry across translations. Applications that depend on this geometry (content-analogous retrieval, translation-quality estimation via cloud-shape agreement, stylistic cross-lingual analysis) should prefer alignment-trained encoders.

6.3 Aesthetic-prediction motivation

This work originated in an aesthetic-judgment modeling project (Anonymous, 2026) where paragraph-cloud geometric features modestly predict reader ratings. The cross-lingual generalization of such features would let aesthetic-rating prediction apply cross-lingually without per-language training data. The present finding is neutral on that application’s feasibility: structural features transfer; feature-to-rating maps do not, at least not linearly (§4.6). Aesthetic modeling remains an open problem; the language-invariant substrate does exist but the predictive layer does not yet follow.

7 Limitations

Encoder coverage. Four multilingual encoders do not span the full space. We did not test glot500, NLLB-encoder, ernie-m, older LASER variants, distilled S-BERT variants, or mUSE, though several of these fall into the same alignment-signal categories we study here. A larger- n encoder survey, ideally including a second explicit-alignment representative and a second pure-MLM representative, would tighten the monotonicity claim and let us test whether the ordering is strictly a function of the objective or is confounded with other design choices (pretraining corpus, tokenizer, training budget).

Size-controlled comparison. LaBSE and mBERT are base/medium-sized (dim 768), while XLM-R-large and mE5-large are large (dim 1024). The monotonic ordering we report therefore co-varies with model capacity as well as with alignment objective. A size-controlled ablation (all-base or all-large) would isolate architecture and capacity from objective. We

expect the alignment effect to dominate — the largest model (XLM-R-large) has the *weakest* spectral invariance — but a formal test is future work.

Per-encoder PCA basis. We refit PCA per encoder from that encoder’s own English paragraph pool. This ensures comparability *within* an encoder but not *across* encoders: absolute ρ values computed in different 128-dimensional subspaces are not strictly on the same scale. We therefore rely on rank correlations (invariant to monotone rescalings) and on the ordinal encoder ordering, which we expect to be stable to basis choice. A shared-basis experiment (e.g., CCA-aligned common subspace) would be a useful follow-up.

Sinitic corpus gap. Chinese translations of Western works are effectively absent from Gutenberg (the available Chinese subset is dominated by classical-Chinese originals, not translations); claims about Sinitic-family invariance are untested in this paper. See §5.4.

Gutenberg bias. Project Gutenberg skews to pre-1929 Western canon in European languages. Modern, non-Western, and genre literature are out of scope. The encoders themselves were trained on contemporary web corpora; any mismatch between pretraining distribution and Gutenberg style is not disentangled from the invariance signal.

Language-family sample size. Slavic, Japonic, Semitic, and Indo-Iranian families each yielded fewer than 10 qualifying bundles after the $n \geq 20$ threshold and are reported for transparency only; statistical claims about these families should be treated as exploratory pending corpus expansion.

Bundle-level independence assumption. Our bootstrap confidence intervals resample at the bundle level. Some bundles share an English reference work with multiple non-English translations (one-to-many); bundles are therefore not strictly independent. We resample at bundle-level nonetheless, which may slightly underestimate CI width for features where the English-side variance dominates. A cluster bootstrap keyed on English work-id would be a more conservative alternative.

Paragraph segmentation. Blank-line splitting is language-naïve; CJK texts formatted without blank lines require alternative segmentation.

Goodreads labels. English-reader-aggregated; applying to translations is pragmatic, not rigorous.

Bundle precision. Author-surname + title-cosine ≥ 0.5 admits occasional spurious pairings, $\leq 5\%$ on manual inspection.

8 Conclusion

Document-level embedding-cloud geometric invariance under translation is not a generic property of multilingual pretraining. Across four multilingual encoders spanning the alignment-objective spectrum, mean cross-lingual Spearman ρ on spectral-divergence features is monotonically ordered by alignment-signal strength: 0.669 (LaBSE, explicit) $\rightarrow$ 0.490 (mE5, contrastive-mined) $\rightarrow$ 0.223 (mBERT, shared-WP MLM) $\rightarrow$ 0.017 (XLM-R, pure MLM, n.s.). Trajectory features partly survive even the no-alignment encoder, suggesting two layers of cross-lingual information: *ordering* structure, pretraining-induced, and *distributional-energy* structure, alignment-induced. The four-encoder panel, cheap to compute on our 657-bundle set, is a candidate diagnostic for multilingual-encoder alignment strength complementary to task-performance benchmarks.

Ethics Statement

Data sources. All text data in this study originates from Project Gutenberg (Hart, 1971), a public-domain digitization effort, and from a Goodreads-derived aggregate-ratings table distributed on Kaggle (Soumik, 2019) that contains no individual user review text and no PII. The companion dataset references for FMA (Defferrard et al., 2017) and the MERT model (Li et al., 2023) pertain to a separate line of work and are cited for completeness; no audio data was processed in the experiments reported here. No human subjects, crowdsourcing, or annotator labor was involved.

Bias caveats. Project Gutenberg skews to pre-1929 Western canon and over-represents European-language literature. All four encoders were trained on large web corpora with documented English and Western-European over-representation. The four-encoder ablation reduces the concern that the headline finding is an encoder-specific artifact: the *pattern* across encoders is the finding. The Sinitic corpus-design gap is honestly flagged (§5.4).

Dual use. No identified dual-use concerns beyond those standard to the embedding-analysis literature. The methods reported here are analysis-only (diagnostic probes of existing encoders); they do not introduce new generative capability or enable new downstream applications beyond the multilingual-representation-analysis work already widely practised in the community.

Reproducibility commitment. We commit to releasing, upon acceptance, the pipeline code, per-encoder PCA-128 bases, per-encoder feature matrices, the bundle-matching index, and the four-encoder ablation scripts under the MIT license. Full software versions, hyperparameters, and expected runtimes are documented in Appendix E.

Environmental impact. Total compute for the full four-encoder ablation is approximately 2.5 GPU-hours on a single 32 GB GPU, which we regard as negligible on the scale of typical multilingual-NLP experiments. No model training or fine-tuning was performed; all experiments were forward-inference only.

Stakeholders. NLP researchers interested in multilingual representation geometry and multilingual encoder design; digital-humanities scholars interested in cross-lingual stylistic analysis.

Software and Data Release

The pipeline code, feature-extraction routines, per-encoder PCA-128 bases, per-language feature matrices, four-encoder ablation code, and bundle-matching index will be released under the MIT license upon acceptance.

References

- Anonymous. 2026. [companion paper under review; details withheld for anonymous review]. Anonymous submission.
- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Anil Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109.

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. [Latent Dirichlet allocation](#). *Journal of Machine Learning Research*, 3:993–1022.
- Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. [Isotropy in the contextual embedding space: Clusters and manifolds](#). In *Proceedings of the Ninth International Conference on Learning Representations (ICLR)*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Alexis Conneau and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 7057–7067.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Michael Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. [FMA: A dataset for music analysis](#). In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 316–323, Suzhou, China.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Michael Hart. 1971. Project Gutenberg. <https://www.gutenberg.org/>. Founded 1971; public-domain digital library accessed as a locally-mirrored dump for this study.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.
- Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966. PMLR.
- Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhua Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. 2023. MERT: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*.
- Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674. Association for Computational Linguistics.
- David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.
- Frank Nielsen. 2020. An elementary introduction to information geometry. *Entropy*, 22(10):1100.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525. Association for Computational Linguistics.
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Soumik. 2019. Goodreads-books dataset. Kaggle. <https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks>.

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

A Full feature definitions

All features are computed on the 128-dimensional PCA-projected paragraph embeddings of a document, using the per-encoder English PCA basis U_E . Let $\tilde{X} = [\tilde{x}_1, \dots, \tilde{x}_T] \in \mathbb{R}^{T \times 128}$ with $\tilde{x}_t = U_E^\top E(p_t)$ where $E(\cdot)$ is the encoder’s paragraph-level embedding (native sentence-level for LaBSE and mE5; mean-pooled last hidden state for mBERT and XLM-R-large). Let $(\mu_\lambda, \Sigma_\lambda)$ be the encoder-specific English corpus mean and covariance in this space, and (μ_p, Σ_p) the empirical moments of $\tilde{X}$, with covariances ridge-regularized by τI , $\tau = 10^{-6}$.

Channel A gives Gaussian divergences as in §3.3, using $\delta = \mu_p - \mu_\lambda$, $d = 128$. Jensen–Shannon uses the mixture-component KL. Bhattacharyya uses the closed form with $\bar{\Sigma} = \frac{1}{2}(\Sigma_p + \Sigma_\lambda)$. Hellinger is derived from Bhattacharyya. TV is the Pinsker upper bound $\sqrt{\frac{1}{2}D_{\text{KL}}(p||\lambda)}$. Mahalanobis-mean is $\sqrt{\delta^\top \Sigma_\lambda^{-1} \delta}$. Channel B gives pairwise cosine mean and standard deviation over all $\binom{T}{2}$ pairs. Channel C gives step-size moments, lag-1 step autocorrelation, recurrence rate at threshold $\epsilon = 0.1\bar{s}$, discrete curvature, and path efficiency. Channel D gives per-axis means of the 128-dimensional projection, $\tilde{x}_k$.

B Per-encoder $\bar{\rho}$ matrix

Table 9 gives the per-encoder, per-feature cross-lingual mean Spearman ρ summary. The full 17×10 per-language per-encoder matrix (4 encoders $\times$ 17 features $\times$ 10 languages = 680 entries) is released with the code package on acceptance; a condensed summary is given here for review convenience.

Feature	XLM-R	mBERT	mE5	LaBSE
pair_sim_mean	0.271	0.465	0.672	0.688
mahal_mean	0.067	0.560	0.483	0.694
Hellinger	0.017	0.223	0.490	0.669
Bhattacharyya	0.017	0.223	0.490	0.668
Jensen–Shannon	0.015	0.217	0.492	0.668
$\text{KL}(p \lambda)$	−0.001	0.204	0.462	0.664
Total variation	−0.002	0.166	0.515	0.664
path_eff	0.190	0.211	0.143	0.310
recur_rate	0.008	0.488	0.657	0.541
step_mean	0.616	0.681	0.596	0.653
step_std	0.305	0.468	0.390	0.599
step_skew	0.356	0.398	0.221	0.450
acfl	0.417	0.420	0.526	0.495
curvature	0.215	0.300	0.396	0.356
pair_sim_std	0.448	0.482	0.438	0.601

Table 9: Cross-lingual mean Spearman ρ per feature per encoder (matched-657 set, $K=128$). Spectral-divergence features (rows 2–7) exhibit the clean monotonic pattern; trajectory features (rows 8–14) compress into a narrower band across encoders, with `step_mean` as the notable flat exception (see §4.3). The full 17×10 per-language per-encoder matrix is released with the code package on acceptance.

Per-feature Kendall τ against alignment rank.

Table 10 reports the per-feature Kendall τ between alignment rank (XLM-R= 0, mBERT= 1, mE5= 2, LaBSE= 3) and the four per-encoder $\bar{\rho}$ values; this is the per-feature view that the stacked-68 aggregate $\tau = 0.513$ ($p = 2.3 \times 10^{-8}$) summarizes. All six Channel-A (spectral) features achieve the maximum $\tau = 1.000$ at $N = 4$; `step_mean` is flat ($\tau = 0.000$).

PCA-dimension sensitivity. Tables 11–12 show the sensitivity of $\bar{\rho}(\text{Hellinger})$ and EN–FI Hellinger ρ to PCA dimension $K \in \{64, 128, 256\}$. Strongly-aligned encoders (LaBSE, mE5-large) are K -invariant within ± 0.02 ; mBERT’s $\bar{\rho}(\text{Hellinger})$ drops sharply from 0.223 at $K=128$ to 0.092 at $K=256$; XLM-R-large remains near-zero throughout.

C Author-surname matching

For each English work with author string a , we extract surname s as the final whitespace-delimited token of a , lowercase-normalized, punctuation-stripped. In each non-English Gutenberg catalog, we search for works whose author string, parsed identically, has surname equal to s . Candidate pairs are filtered by LaBSE title-cosine $\cos(\text{LaBSE}(\text{title}_{\text{en}}), \text{LaBSE}(\text{title}_L)) \geq 0.5$. The

Feature	Kendall τ	p ($N=4$)
Hellinger	1.000	0.083
Bhattacharyya	1.000	0.083
Jensen-Shannon	1.000	0.083
$KL(p \lambda)$	1.000	0.083
Total variation	1.000	0.083
pair_sim_mean	1.000	0.083
mahal_mean	0.667	0.333
step_std	0.667	0.333
recur_rate	0.667	0.333
acfl_top3	0.667	0.333
curvature	0.667	0.333
tail_mass_100	0.667	0.333
pair_sim_std	0.333	0.750
step_skew	0.333	0.750
path_eff	0.333	0.750
powerlaw_slope	0.333	0.750
step_mean	0.000	1.000

Table 10: Per-feature Kendall τ between alignment-rank and per-encoder $\bar{\rho}$ ($N = 4$ encoders). At $N = 4$ no single-feature τ is significant at $p < 0.05$ (the minimum attainable p is 0.083), hence the motivation to aggregate: 16 of 17 features show positive τ (binomial $p \approx 1.4 \times 10^{-4}$), and the stacked-68 aggregate Kendall $\tau = 0.513$ with $p = 2.31 \times 10^{-8}$.

Encoder	$\bar{\rho}(K=64)$	$\bar{\rho}(K=128)$	$\bar{\rho}(K=256)$
LaBSE	0.666	0.669	0.671
mE5-large	0.490	0.490	0.485
mBERT	0.231	0.223	0.092
XLM-R-large	0.023	0.017	0.001

Table 11: Mean cross-lingual $\bar{\rho}$ on Hellinger as a function of PCA dimension K . Strong-alignment encoders are K -invariant; mBERT shows a K -dependent drop; XLM-R is near-zero at all K .

0.5 threshold was hand-selected on a 50-pair manual inspection sample; sensitivity at 0.4 and 0.6 yields qualitatively identical results.

D Encoding configuration

All encoders run on a single GPU with 32 GB VRAM (NVIDIA-class) in fp16, batch size 16. Paragraph boundaries are blank-line splits on Gutenberg plaintext. LaBSE: sentence-transformers/LaBSE, native CLS-token sentence embedding. mE5-large: intfloat/multilingual-e5-large, native mean-pooled sentence embedding with the “passage: ” prefix applied per the model card. mBERT: bert-base-multilingual-cased,

Encoder	$\rho(K=64)$	$\rho(K=128)$	$\rho(K=256)$
LaBSE	0.768	0.776	0.772
mE5-large	0.475	0.490	0.524
mBERT	0.274	0.299	0.288
XLM-R-large	0.062	0.056	0.057

Table 12: English-Finnish Hellinger ρ ($n = 288$) as a function of PCA dimension K . The headline LaBSE result $\rho \approx 0.77$ is stable across K ; the encoder ranking LaBSE $>$ mE5 $>$ mBERT $>$ XLM-R is preserved at every K .

last-hidden-state mean-pooled with attention-mask normalization. XLM-R-large: xlm-roberta-large, last-hidden-state mean-pooled with attention-mask normalization. Per-encoder runtime on the full 657-bundle / 1,881-book set: XLM-R-large 23.6 min, mBERT 10.6 min, mE5-large 16.8 min, LaBSE $\approx$ 40 min (prior run on slightly larger bundle set).

E Reproducibility

This appendix documents the information needed to reproduce our experiments from scratch.

E.1 Hardware

All experiments run on a single 32 GB GPU (NVIDIA-class) with a 16-core CPU and 64 GB RAM. The total compute budget for the full four-encoder ablation is approximately 2.5 GPU-hours, forward inference only (no training or fine-tuning).

E.2 Software versions

Python 3.12; PyTorch 2.10 with CUDA 12.8; sentence-transformers (for LaBSE); transformers (for mBERT, XLM-R, and mE5); scipy 1.14; scikit-learn 1.5; numpy 2.1; pandas 2.2.

E.3 Model versions (HuggingFace hub IDs)

- sentence-transformers/LaBSE
- xlm-roberta-large
- bert-base-multilingual-cased
- intfloat/multilingual-e5-large

E.4 Encoding hyperparameters (per encoder)

- Maximum sequence length: 256 tokens (paragraph-level).

- Pooling: mean of last hidden states, attention-mask normalized (LaBSE uses its native pooler-output sentence embedding; mE5 uses its native mean-pool with the “passage: ” prefix).
- fp16 inference on GPU.
- Batch size 16.

E.5 PCA

128 components. Fit on each encoder’s English paragraph pool separately; no shared basis across encoders. The same per-encoder basis U_E is applied to every non-English language encoded by E .

E.6 Cross-validation

The Ridge aesthetic-rating-transfer experiments (§4.6) use 5-fold GroupKFold with `author_id` as the group variable to prevent train/test leakage across works by the same author.

E.7 Random seeds

NumPy seed 42 is used for the PCA subsampling step and for the bootstrap resampling of per-language ρ . No other randomness appears in the pipeline (encoders are run deterministically in eval mode).

E.8 Expected runtimes

On the full 657-bundle / 1,881-book set: LaBSE ≈ 40 min (on the slightly larger prior bundle set of 4,683 books); XLM-R-large ≈ 24 min; mBERT ≈ 11 min; mE5-large ≈ 17 min; analysis and bootstrap ≈ 10 min per encoder.

E.9 Data paths for release

The bundle index, per-encoder feature matrices, per-encoder PCA-128 bases, and four-encoder ablation outputs will be released under the MIT license upon acceptance at [URL withheld for anonymous review].