MEASURING THE LANGUAGE SHARE OF MULTILINGUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual language models can represent the same content in different languages, but the balance between language and sentence differences inside these models remains difficult to measure. We introduce Language Factor Share (LFS), which uses translations of the same sentences to measure average language differences relative to sentence differences at each layer, without choosing a reference language. Across 33 open base models, language share is lowest at an intermediate layer and higher at the input and output. This pattern repeats on a second collection of translations for 17 models, and the decrease varies widely among similarly sized models. Lower language share is also associated with higher average accuracy on two multilingual benchmarks. In a preregistered study of 19 models, lower language share goes with greater benefit from relevant context in another language when predicting English text, after accounting for a proxy for language data availability, language family, script, and tokenization. Reporting the two parts the share is built from shows which one changed, or whether both did. Our training experiments did not establish that lowering the share improves performance. LFS therefore helps researchers examine how multilingual representations are organized and interpret their changes alongside direct tests of model performance.

1 INTRODUCTION

Large language models are trained mostly on English and a few other high-resource languages, and they often perform worse in lower-resource languages (Shani & Basirat, 2025; Hu et al., 2025; Kocmi et al., 2025). One common explanation is that such a model represents the same content differently in different languages. Part of the variation across translations of the same sentences belongs to the language, part to the sentence, and part to the interaction of the two. We ask how the first two parts, language and sentence, compare at each layer of a model. This structural question is distinct from whether translations retrieve one another or whether the model answers a task correctly. Prior work locates language-specific structure near the input and output layers and shared structure in between (Shani & Basirat, 2025; Tezuka & Inoue, 2025), finds that the shared layers are not simply English (Bafna et al., 2025), and shows that word-embedding spaces (Conneau et al., 2018) can be aligned by simple linear maps and that contextual representations can also be mapped linearly, though less cleanly than static embeddings (Peng & Søgaard, 2024; Xu & Koehn, 2021). The measures used in that work are pairwise similarities, retrieval scores, or neuron selections. We add a measure of the relative size of the language and sentence effects, defined at every layer and comparable across models measured the same way, and we test where it works and where it fails rather than assume it works (Ding & Koehn, 2021).

Take N sentences translated into L languages, so that the translations of one sentence share a sentence identity; embed each of the $L \times N$ strings at one layer of a model; and ask how the systematic variation among the resulting vectors splits between *which language* a vector came from and *which sentence* it encodes. Language Factor Share (LFS) is the language part of that split as a fraction of the two systematic parts. It lies between 0 and 1, has a reference value under the noise model of §3 that depends on the numbers of languages and sentences, and can be computed at every layer. Unlike retrieval scores such as MEXA (Kargaran et al., 2025), it needs no reference language or nearest-neighbor search and summarizes all languages and sentences at once.

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

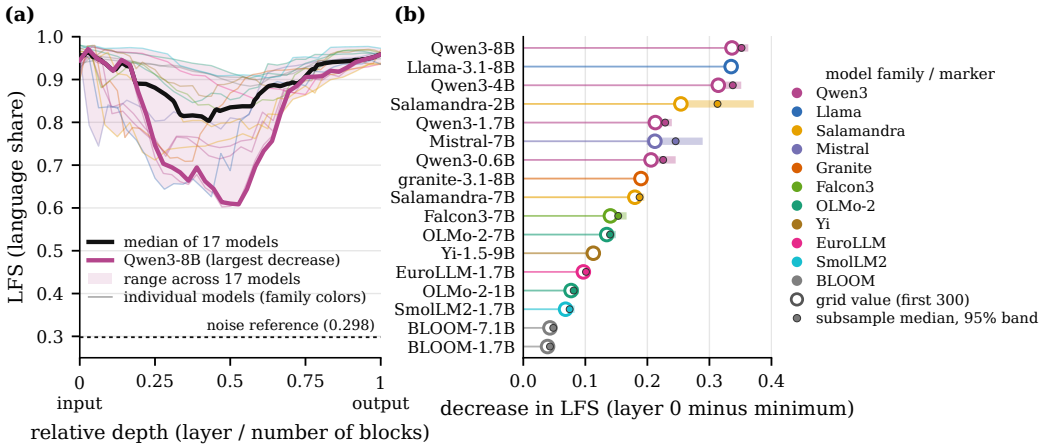

Figure 1: (a) LFS by relative depth (layer index over the number of transformer blocks) for the 17 primary-set models (128 languages, 300 sentences): faint lines are the individual models, the band their range, the thick line the median, Qwen3-8B (largest decrease) is drawn separately, and the dashed line is the noise reference (0.298). (b) Decrease in LFS from layer 0 to the minimum layer for all 17 models, colored by family and sorted. Open markers: the first-300-sentence grid values used throughout; filled marker and bar: median and 95 percent interval over 2,000 subsets of 300 of the 1,500 stored sentences (14 models; Table 8), a sensitivity to sentence selection rather than an interval around the grid value. Standard estimator; per-model values in Tables 4 and 8.

We define LFS and what it cannot see (§3). In 33 open base models, language share is lowest at an intermediate layer, and how far it falls varies widely among models of similar size (§5). A lower share goes with higher average benchmark accuracy and more use of relevant context in another language; we also compare LFS with cosine similarity and MEXA (§6). Lowering the share in training did not improve the tested outcomes, so a lower share does not by itself mean a better model (§7). The replication, context-transfer, and training analyses followed protocols fixed before their runs, as did the seven models added after development, and every failed prediction is reported; the across-model benchmark, cosine, MEXA, and gain-beyond-retrieval comparisons were planned later, on outcomes we had already seen.

2 Related work

Language dominance and layerwise sharing. Logit-lens decoding (nostalgebraist, 2020) of non-English input often yields English tokens at intermediate layers, which Wendler et al. (2024) interpret as a concept space biased toward English; two cautions apply, that intermediate states are far from every output embedding (Appendix M) and that hubness makes a few points appear among the nearest neighbors of many representations (Radovanović et al., 2010; Conneau et al., 2018). Later work reads the same layers as a shared multilingual space rather than an internal translation into English (Shani & Basirat, 2025; Bafna et al., 2025), identifies transfer neurons between shared and language-specific spaces (Tezuka & Inoue, 2025), and locates language-specific neurons mostly in the first and last layers (Kojima et al., 2024; Tang et al., 2024; Zhang et al., 2026). Geometry-based and decoding-based probes disagree about where languages mix (Bayazit et al., 2026); LFS is geometry-based and reports a share rather than naming one internal language; in related layerwise work, Ben Tamo et al. (2026) track per-layer language probabilities and cross-lingual similarity on two models and five non-English languages, and LFS summarizes that pattern as one number per layer, with a noise reference, on 33 models and up to 128 languages, without fine-tuning. Word-embedding spaces are often aligned using linear or orthogonal maps (Conneau et al., 2018; Marchisio et al., 2021; 2022), building on earlier work on lexicon induction from monolingual corpora (Koehn & Knight, 2002), and contextual last-layer representations can also be mapped linearly, though linear fits are less accurate than for static embeddings (Peng & Søgaard, 2024); language-specific information occupies a low-dimensional linear subspace that can be located per layer or projected away (Liang

et al., 2021; Xie et al., 2022; Zhao et al., 2024), language identity admits largely separable linear directions (Marinov et al., 2026), and our mapping analysis asks the same question observationally with a sequence of increasingly flexible maps.

Intrinsic multilinguality metrics. MEXA (Kargaran et al., 2025) scores whether parallel sentences retrieve each other and reports an average Pearson correlation of 0.90 with downstream task performance; Information Parity (Tsvetkov & Kipnis, 2024) compares per-language compression to English; the semantic hub hypothesis reports a shared space across languages and modalities (Wu et al., 2025). Intrinsic scores can conflate surface form with meaning (Poelman & de Lhoneux, 2026) and relate only moderately, and inconsistently across models and layers, to transfer performance (Shah et al., 2024). Similarity indices (Kornblith et al., 2019; Kudugunta et al., 2019), layer ablations (Muller et al., 2021), and analyses of language means (Chang et al., 2022) describe shared structure; centering each language on its own mean makes encoder representations more language-neutral (Libovický et al., 2020), and that additive picture is what LFS quantifies. Performance-prediction and transfer-language selection work models task outcomes from language and data features (Lin et al., 2019; Xia et al., 2020), DEPART decomposes multilingual performance variance into model, language, and interaction components (Uppadhyay et al., 2026), and Hu et al. (2025) quantify language disparities using a mixed-effects model of benchmark scores; we instead adjust associations between representation measures and task outcomes for measured language characteristics. Related methodological work includes Zhao et al. (2026), who split neuron activation probabilities in audio-language models additively into language, emotion, interaction, and noise; LFS applies the classical two-way variance decomposition (Fisher, 1925) to whole hidden-state vectors of text models, per layer, on 128 languages, with a stated noise reference. Our contribution is to define a language share directly over whole sentence representations and test its layerwise patterns, relation to task outcomes, and sensitivity to measurement choices.

3 Language Factor Share

Data and grid. We take N sentences $\{s_c\}_{c=1}^{N}$ and their translations $s_{\ell c}$ into L languages, $\ell = 1, \ldots, L$, so that at each layer of a model we have one representation for every sentence in every language: the grid is complete (balanced), and the translations of one sentence share a sentence identity. We run each $s_{\ell c}$ through the model, average the hidden states of its tokens (excluding padding) at the chosen layer in 32-bit precision, and obtain one vector $\boldsymbol{h}_{\ell c} \in \mathbb{R}^D$ per cell, where D is the hidden-state dimension. We then standardize each coordinate d jointly over all LN cells, $z_{\ell cd} = (h_{\ell cd} - \bar{h}_{\cdot\cdot d})/s_d$, where $\bar{h}_{\cdot\cdot d}$ and s_d are the coordinate's mean and standard deviation over all cells (a coordinate with zero variance is centered and left unscaled), so that a few high-variance coordinates do not dominate the sums below (Timkey & van Schijndel, 2021; Kovaleva et al., 2021; Sun et al., 2024). Standardization is joint, never per language, because per-language standardization would remove the quantity being measured. Figure 7 (Appendix T) summarizes the computation.

Decomposition. We write each standardized vector as the overall mean plus an average language effect, an average sentence effect, and a remainder. Let $\bar{\boldsymbol{z}}$ be the overall (grand) mean over the grid, $\bar{\boldsymbol{z}}_\ell$ the mean over the N sentences of language ℓ, and $\bar{\boldsymbol{z}}_c$ the mean over the L languages of sentence c. The language effect $\boldsymbol{a}_\ell = \bar{\boldsymbol{z}}_\ell - \bar{\boldsymbol{z}}$ and the sentence effect $\boldsymbol{b}_c = \bar{\boldsymbol{z}}_c - \bar{\boldsymbol{z}}$ are differences of means, and $\boldsymbol{r}_{\ell c}$ is what remains:

$$\boldsymbol{z}_{\ell c} = \bar{\boldsymbol{z}} + \boldsymbol{a}_\ell + \boldsymbol{b}_c + \boldsymbol{r}_{\ell c}. \tag{1}$$

The language effect is the part of a language's difference that is the same for every sentence; language differences that depend on the sentence stay in the remainder ("effect" is used in its statistical, not causal, sense). We measure the size of each effect by a sum of squares (SS), the sum of the squared lengths of its vectors; because the language effect is repeated across the N sentences of its language and the sentence effect across the L languages of its sentence, those counts appear as multipliers,

$$\mathrm{SS}_{\mathrm{lang}} = N \sum_{\ell=1}^{L} \|\boldsymbol{a}_\ell\|^2, \qquad \mathrm{SS}_{\mathrm{sent}} = L \sum_{c=1}^{N} \|\boldsymbol{b}_c\|^2, \qquad \mathrm{SS}_{\mathrm{res}} = \sum_{\ell,c} \|\boldsymbol{r}_{\ell c}\|^2. \tag{2}$$

On the complete grid the cross terms cancel when summed over all cells, so $\sum_{\ell,c} \|\boldsymbol{z}_{\ell c} - \bar{\boldsymbol{z}}\|^2 = \mathrm{SS}_{\mathrm{lang}} + \mathrm{SS}_{\mathrm{sent}} + \mathrm{SS}_{\mathrm{res}}$ exactly (Appendix A); this holds for sums over the grid and does not mean

that individual language and sentence effects are perpendicular. It is the vector form of the classical balanced two-way analysis of variance with one observation per cell (Fisher, 1925). Language Factor Share is defined as

$$\mathrm{LFS} = \frac{\mathrm{SS}_{\mathrm{lang}}}{\mathrm{SS}_{\mathrm{lang}} + \mathrm{SS}_{\mathrm{sent}}}. \tag{3}$$

LFS is defined when $\mathrm{SS}_{\mathrm{lang}} + \mathrm{SS}_{\mathrm{sent}} > 0$. The remainder is left out of the ratio so that more unexplained variation does not make a model look language-neutral. It is not only noise, since language differences that vary by sentence also land there; we report its share beside the ratio (per model and layer in the supplementary outputs) and check the raw (unstandardized) sizes of the three parts when interpreting a change (§7). The sentence effect stands in for what the translations of a sentence share; it does not capture all of their meaning.

Interpretation. LFS is the language effect's share of the two effects together. At Qwen3-8B's layer of minimum language share, the language effect ($\mathrm{SS}_{\mathrm{lang}}$), the sentence effect ($\mathrm{SS}_{\mathrm{sent}}$), and the remainder ($\mathrm{SS}_{\mathrm{res}}$) make up 35.1%, 22.6%, and 42.3% of the grid's total standardized variation, so $\mathrm{LFS} = 35.1/(35.1+22.6) \approx 0.608$: the share is taken of the first two parts only. A lower LFS means a smaller language effect relative to the sentence effect; either effect, or both, may have changed. It does not mean that language identity is absent or that translations coincide. In the limiting cases LFS would be 1 if all sentence means coincided and 0 if all language means coincided, and the remainder can be large in either case. A linear classifier recovers language identity at 0.875 to 0.886 accuracy (chance 0.042) at every layer of Qwen3-0.6B (24 languages), including the layer with the smallest share, so a language can be easy to identify even when its share is small.

Noise reference. Finite random samples produce nonzero language and sentence effects even when neither factor has any structure, so LFS is not zero for random vectors. Under a noise model with no language or sentence structure and equal spread in every direction (Appendix B), the ratio of the expected sums of squares gives the noise reference

$$\mathrm{LFS}_{\mathrm{null}} = \frac{L-1}{(L-1)+(N-1)}, \tag{4}$$

0.298 on our 128×300 grid and 0.078 at N=1500. It depends on the grid size and is not a universal chance level; synthetic grids and pseudo-language splits reproduce it to three decimals (Appendix B), and measured grids lie far above it (§5).

What LFS can and cannot see. Multiplying all raw representations by the same nonzero scalar leaves LFS unchanged after joint standardization, so LFS cannot detect a uniform change in representation magnitude; we report magnitude-sensitive quantities separately (§7). LFS also does not say how two languages differ, for example whether one is shifted, rescaled, or rotated relative to the other; the mapping experiments in §5 test this.

Summarizing the layers. We measure LFS at every layer and summarize the result by three numbers: the *minimum layer*, where the language share is lowest; LFS at that layer; and the *decrease*, how much LFS falls from layer 0 to the minimum layer. Layer 0 is the embedding output, so a model with B blocks has $B+1$ layers to measure.

Simple cosine measures for comparison. A cosine similarity to English is the simplest alternative measure of the same representations, so we define two constructions on the raw sentence vectors $\boldsymbol{h}_{\ell c}$, with English as the reference language and $\bar{\boldsymbol{h}}_\ell = \frac{1}{N}\sum_c \boldsymbol{h}_{\ell c}$ the mean vector of language ℓ,

$$\mathrm{MatchedCos}_\ell = \frac{1}{N}\sum_c \cos(\boldsymbol{h}_{\ell c}, \boldsymbol{h}_{\mathrm{en},c}), \qquad \mathrm{LanguageMeanCos}_\ell = \cos(\bar{\boldsymbol{h}}_\ell, \bar{\boldsymbol{h}}_{\mathrm{en}}), \tag{5}$$

and report similarity. Two further per-language measures, a language-offset cosine on the standardized grid and a pairwise LFS on the two-language grid of English and one other language (unlike the global LFS over all languages, it depends on English), are defined in Appendix N; Table 11 records the preprocessing of every measure: the two cosines use raw vectors, whereas LFS, MEXA, and the offset cosine use the jointly standardized grid.

4 EXPERIMENTAL SETUP

Parallel text. We use NTREX-128 (Federmann et al., 2022), the WMT19 news test set (1,997 sentences in 123 documents) professionally translated from English into 128 languages, so every sentence exists in every language. We computed LFS at every layer using the first 300 sentences in each of the 128 languages, and for 14 models we saved 1,500-sentence representations at the minimum layer and four other layers; Appendix S follows six languages through the results. NTREX is translated from English; our comparison of original and translated WMT19 text (Appendix I) shifts the language share by $+0.02$ to $+0.06$ and preserves model ordering at Spearman 0.99.

Models. We measured a primary set of 17 base models across eleven families with the standard estimator of Eq. 3 (Table 4): we used ten while developing the method and added seven later under predictions fixed before their runs. We measured a second expansion set of 19 models from 13 families (three overlap the primary set; Appendix E) on a smaller grid, English plus the 40 context-transfer languages with 300 sentences, using a noise-corrected variant, LFS-VC, estimated by restricted maximum likelihood (Appendix C); 33 unique models in total. We study base checkpoints because the structural question concerns what pretraining produces; six base and instruction-tuned pairs change the decrease by at most 0.01 (Table 9, Appendix I). Table 3 lists the models, languages, sentences, estimator, and outcome or output of each analysis; the 14 models with saved vectors and the 19 context-transfer models do not overlap.

Benchmark accuracy and use of context. We ask whether representation measurements relate to two outcomes: accuracy on multilingual question-answering benchmarks, and the use of relevant context supplied in another language. **Belebele** (Bandarkar et al., 2024) is four-choice reading comprehension; we score 300 questions per language in 122 language variants by zero-shot log-likelihood under one uniform prompt. **INCLUDE** (Romanou et al., 2025) has natively authored questions in 30 overlapping languages, scored the same way, with no English counterpart per item. The **context-transfer test** (§6) asks whether a relevant preceding sentence in another language improves the likelihood of an English continuation relative to an unrelated sentence in the same language; it scores likelihoods and tests neither generated answers nor the language a model generates in.

We describe the three analysis designs and the covariate adjustment at the start of §6.

Reference measure. We use MEXA (Kargaran et al., 2025) as a published retrieval-based reference measure in the comparisons below, with its preprocessing specified for each analysis; it scores whether English sentences and their translations retrieve each other. Alignment-at-Risk (AaR@10), the mean retrieval margin of the worst-aligned tenth of sentences, summarizes the hardest cases; Table 10 (Appendix M) lists every diagnostic computed on the same grid.

5 LAYERWISE STRUCTURE ACROSS 33 MODELS

Language share is lowest at an intermediate layer. In all 17 models measured with the standard estimator and all 19 measured with LFS-VC, the language share is lowest at an intermediate layer and higher at the first and last layers (Figure 1a; Table 4, Appendix F); a lower share can reflect smaller differences between languages, larger differences between sentences, or both. The minimum lies at relative depth 0.06 to 0.66, most between 0.35 and 0.6. In the primary set the lowest LFS is about 0.60 to 0.90, above 0.5 and far above the noise reference of 0.298, so even at the minimum layer, differences between languages outweigh differences between sentences. The expansion values run lower (Llama-3.1-8B 0.479 against 0.626 on the primary grid); they come from a different estimator on a smaller set of 41 languages, so levels are compared only within each set.

Language share decreases by different amounts in similarly sized models. The decrease from layer 0 to the minimum layer ranges from 0.039 to 0.336 among the primary models, about 8.6-fold (Figure 1b). A larger decrease means that, by the minimum layer, differences between languages have shrunk more relative to differences between sentences; it is not a measure of multilingual ability. Within the Qwen3 series the decrease grows with size (0.206, 0.213, 0.314, 0.336), but differences between families of the same size are much larger: Qwen3-1.7B (0.213) has a larger decrease than

every model of 7B or more from another family except Llama-3.1-8B, while BLOOM, trained on a deliberately multilingual 46-language corpus, has the smallest decrease, which barely changes from 1.7B to 7B (0.039 to 0.043). This is a description, not an explanation: size alone does not account for the decrease, and these comparisons do not separate architecture, tokenizer, training data, or training procedure.

The pattern holds on new models and a second corpus. Two model families not used in development, SmolLM2-1.7B and Falcon3-7B, met predictions fixed before their runs (decreases of 0.068, predicted ≤ 0.15, and 0.140, predicted intermediate). On FLORES-200 devtest (116 languages, 300 sentences), all 17 models again have their minimum at an intermediate layer, and the decreases on the two corpora agree at Spearman 0.92. Sentence-subsampling intervals on the decrease (14 models; Appendix I) have 95 percent half-widths of 0.004 to 0.020 for 12 models and exceed the preset 0.03 bound for the two with a very early minimum, the one failed prediction of twelve.

A shift and rescaling explain most measured mapping error. At each model's minimum layer we map each language into a shared average space fitted across languages with increasingly flexible maps (a shift, a uniform rescaling, a rotation, a general linear map, and a nonlinear kernel map), each credited with how much it reduces the error on held-out sentences, those not used to fit it, beyond what randomly paired sentences achieve (Figure 3, Appendix D); the percentages are shares of that mapping error, not of $\mathrm{SS}_{\mathrm{lang}}$. A shift plus a rescaling accounts for 63 to 97% of the mapping error across the 14-model study, a rotation for -0.2 to $+0.7$%, and a general linear map for 0.3 to 8% (Table 2). Shifting and rescaling each language's representations therefore removes most measured error, and rotation adds little. This is consistent with prior work on adding language means (Sterz et al., 2025) and on mapping contextual representations under isotropic and isometric constraints (Xu & Koehn, 2021).

A multilingual average predicts a language's representations better than English does. At each model's layer of highest mean retrieval accuracy (MEXA), we project the standardized grid onto its first 64 principal components and, for each target language, fit a ridge regression (penalty 1) that predicts the target's vectors from the English vectors of the same sentences, and separately from the mean of all other languages except the target and English; we compare their 3-fold cross-validated R^2. The 127 targets and the average include three regional English varieties, and the components are fitted on the whole grid before the folds. The multilingual average predicts better than English for 113 to 124 of the 127 targets depending on the model, by a median $+0.06$ to $+0.16$ R^2. An average of many languages may gain partly because averaging reduces noise, so this does not establish that English plays no special internal role.

6 Language share, accuracy, and context use

Correlations are reported for LFS itself, so a negative correlation means that a lower language share goes with higher accuracy or a larger context benefit. Except where we say a protocol was fixed in advance, these analyses were planned after the outcomes had been seen.

Analysis designs. Each result states what one observation is: a language within one model (within-model; per-model correlations averaged over models, with a model-bootstrap interval), a model (across-model; measure and outcome aggregated over languages), or a model–language pair (adjusted analyses, which resample models, languages, or both together, "crossed", because models and languages repeat across rows). Languages with more available text or a different tokenization may differ in both a measure and a task score, so we report unadjusted and adjusted relationships; following Lin et al. (2019) and Xia et al. (2020), the adjustment regresses outcome and measure on $\log_{10}$ CulturaX token counts (Nguyen et al., 2024), a proxy for how much text exists in each language rather than for any model's training data, 0/1 indicators for the Glottolog macro-family (the top-level language family; Hammarström et al., 2024) and the script, and tokenizer fertility, the number of tokens per word (Ahia et al., 2023; Petrov et al., 2023); these proxies do not remove all confounding. The main per-language benchmark analysis also gives each model its own constant and uses a logit accuracy scale whose bottom is the chance level of 0.25 (Appendix G); the comparisons in Table 13 use raw accuracy, and the context-transfer analysis has no per-model constant because its predictor is one value per model. Spearman is written ρ and Pearson r; intervals are 2,000-replicate bootstrap

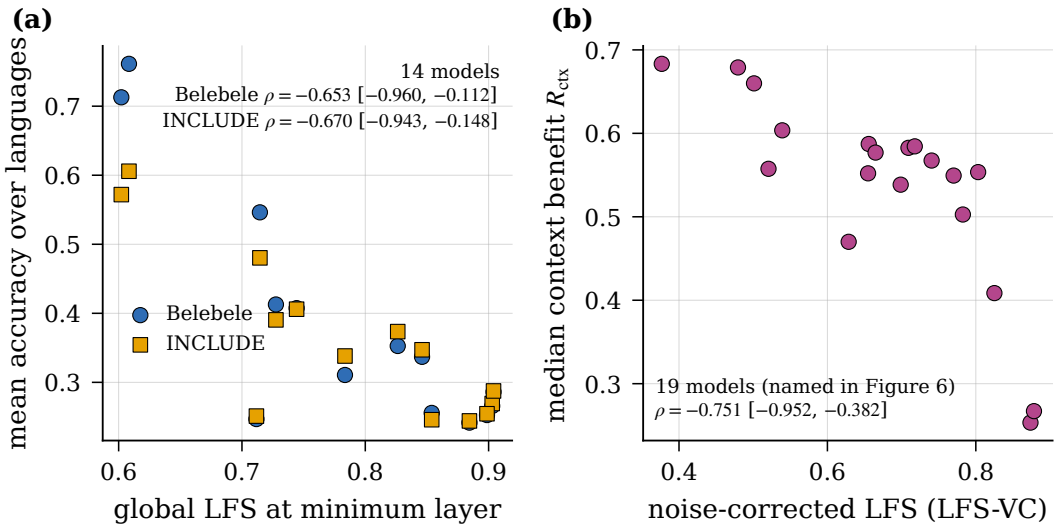

Figure 2: Language share, benchmark accuracy, and context use across models; each marker is one model on one outcome, with unadjusted Spearman correlations with 2,000-replicate model-bootstrap intervals. (a) Global LFS at each model's minimum layer against mean Belebele (circles) and INCLUDE (squares) accuracy, for the 14 models with saved vectors. (b) LFS-VC at the minimum layer against the median normalized context benefit R_{ctx} over 34 languages, for the 19 context-transfer models (named in Figure 6). Panel (a) concerns average accuracy; an association with the gap from English is not established (§6).

percentile intervals over the stated unit, except the exploratory cosine and continuation checks, which resample models 5,000 times.

Lower language share goes with higher average benchmark accuracy. Each observation is one model. Across the 14 models with saved vectors, a lower global LFS at the minimum layer goes with higher mean accuracy on Belebele ($\rho = -0.653$ $[-0.960, -0.112]$) and INCLUDE ($\rho = -0.670$ $[-0.943, -0.148]$; Figure 2a). We use these 14 models because they are the ones on which every measure below can be compared. An association with accuracy relative to English is not established: mean Belebele accuracy and English accuracy rank these models almost identically ($\rho = 0.96$), and in an exploratory check LFS shows no clear association with the mean gap to English, a language's accuracy minus the model's English accuracy ($\rho = 0.22$ $[-0.35, 0.62]$). Within a model, English accuracy is a constant, so correlating with the gap gives the same ranking as correlating with accuracy. Across all 33 models, LFS-VC also correlates with mean accuracy ($\rho = -0.567$ $[-0.744, -0.282]$ for Belebele over 40 common languages, -0.676 $[-0.830, -0.414]$ for INCLUDE over 30), but the two groups of models were measured on different grids with different rules for choosing the layer, so we treat this only as a secondary check (Appendix G.1). The covariate-adjusted per-language analyses give each model its own constant, which removes its overall accuracy level; because global LFS is one value per model, those analyses cannot test it, and we use them only for per-language measures (Appendix G). Pooled benchmarks also carry contamination and annotation errors (Deng et al., 2024; Gema et al., 2025).

Lower language share goes with more use of context in another language. For an English target sentence and its preceding same-document sentence from NTREX, we compare the loss of the target (its negative log-likelihood) after the preceding sentence in each of 41 context languages with its loss after an unrelated sentence of similar length from another document in the same language. The difference, $\text{benefit}_\ell = \text{NLL}_{\text{mismatched},\ell} - \text{NLL}_{\text{matched},\ell}$, the loss after the unrelated sentence minus the loss after the relevant one, is the benefit of relevant context; dividing it by the English-context benefit gives $R_{\text{ctx},\ell}$, so this outcome is already measured relative to English. After a four-model pilot that could not support a model-level claim, we fixed the protocol in advance and evaluated 19 new models from 13 families (Appendix H). Across models, a lower LFS-VC goes with a larger

Table 1: Use of context in another language, 19 models from 13 families. Upper block: one model–language row per observation (646 rows, 34 languages); covariate-adjusted Spearman with the target shown (the LFS-VC row with R_{ctx} is the preregistered result); 2,000-replicate bootstrap intervals resampling models (Model) or models and languages together (Crossed). Lower block: one observation per model (Figure 2b).

Measure	Target	ρ	Model 95% CI	Crossed 95% CI
Adjusted model–language rows (n=646)				
LFS-VC	R_{ctx}	$\mathbf{-0.46}$	$[-0.69, -0.11]$	$[-0.71, -0.11]$
MEXA (reference)	R_{ctx}	0.697	$[0.50, 0.85]$	$[0.45, 0.85]$
AaR@10 (worst decile)	R_{ctx}	0.581	$[0.33, 0.78]$	$[0.29, 0.78]$
LFS-VC	raw benefit	-0.623	$[-0.79, -0.30]$	$[-0.80, -0.32]$
Model level (n=19; one observation per model)				
LFS-VC	median R_{ctx}	-0.751	$[-0.952, -0.382]$	–
MEXA (model mean)	median R_{ctx}	0.653	$[0.273, 0.856]$	–

benefit: $\rho = -0.751$ $[-0.952, -0.382]$ with one observation per model (Figure 2b), against 0.653 $[0.273, 0.856]$ for MEXA averaged over languages. Table 1 gives the preregistered covariate-adjusted result, $\rho = -0.46$ on 646 model–language rows (after correcting one model's inputs; Appendix R), where MEXA reaches 0.70. Because LFS-VC has one value per model, the 646 rows add no evidence about LFS beyond the 19 models; they let the adjustment use language-level information, and the model bootstrap resamples whole models. The association holds with the standard estimator (-0.461 against -0.459; Appendix C), when any one model family is left out (-0.55 to -0.38), and across 80 adjustment specifications (-0.89 to -0.36, every interval excluding zero; Appendix Q).

Within a model, simpler measures rank languages alike. One observation per language: matched-translation and language-mean cosine give nearly the same ranking of languages (mean $\rho = 0.994$ $[0.989, 0.997]$), matched cosine and MEXA agree at 0.866 $[0.814, 0.914]$, and pairwise LFS (English plus one language) tracks MEXA almost exactly with the opposite sign (-0.963 $[-0.970, -0.956]$). Their unadjusted relations with per-language Belebele accuracy are similar in strength, $|\rho|$ from 0.48 to 0.55 (Figure 5, Appendix N). For ranking the languages of one model, a cosine or retrieval score therefore gives much the same answer as the LFS-derived measure.

Across-model comparisons depend on preprocessing. One observation per model: on the 14 models with saved vectors, raw matched cosine, MEXA, and global LFS agree only partly with one another, and the intervals of mean cosine and mean MEXA with mean Belebele accuracy include zero (Appendix N). These differences may partly reflect preprocessing: cosine was computed on raw representations, which at intermediate layers are dominated by a direction shared by all sentences (Appendix N), whereas LFS and MEXA used standardized ones. In an exploratory comparison of 19 models on a common 35-language grid, correlations with mean normalized benefit from relevant foreign-language context were -0.70 for standard LFS, $+0.81$ for standardized cosine, and -0.12 for raw cosine. The standardized-cosine correlation is larger in magnitude by 0.11 (95% model-bootstrap interval $[-0.13, 0.44]$), so neither measure has a demonstrated advantage (Appendix N).

A gain beyond retrieval is not established. As a secondary check, we predicted each model's median context benefit with its family held out of the fit: adding LFS-VC to mean MEXA lowered the mean absolute error from 0.0860 to 0.0676, but the reduction, 0.018 $[-0.003, 0.040]$, includes zero, as it does with AaR@10 in place of MEXA (Appendix P).

7 Interpreting changes in language share

Measures that ignore overall scale cannot see uniform shrinking. LFS and MEXA are unchanged when all representations are multiplied by the same nonzero constant, so they cannot register a uniform shrinking of representation size, which differs from dimensional collapse (Jing et al., 2022). When the non-English languages are shrunk toward the overall mean by 30, 60, or 90 percent (three models) only the content-variance preservation check (CVP, each language's total within-language variance

relative to the starting checkpoint) responds, returning $(1 - m)^2$ for shrinkage m as predicted, while LFS moves by at most 0.05 and MEXA by 0.01. A trained model shows the same pattern: the word-level alignment loss of Cao et al. (2020), a squared difference between representations of word-aligned pairs in a sentence and its translation, is smallest when every representation is zero ($\boldsymbol{h} = 0$); we use it without their regularizer. These representation measurements use 300 parallel sentences in English, German, Hindi, and Swahili at fixed layer 8, with each checkpoint standardized on this grid. Their absolute values are not directly comparable with the 128-language profiles. Fine-tuning Qwen3-0.6B with this loss for 400 steps (three seeds) shrank the mean representation norm from 451 to 42 and the raw variance of the sentence means from 13,791 to 699 (Figure 4, Appendix L), while LFS moved only from 0.445 to 0.493 and MEXA rose by $+0.09$ over a control trained with the language-modeling loss alone: the ratio hides a twenty-fold fall in the raw sentence variation. The shrunk model matched its control on a 40-language Belebele panel (0.395 versus 0.390), and conditions that lost about six points showed little change in the measured representation statistics, so CVP flags changed representations, not lost ability.

Lowering LFS did not improve the tested outcomes. Across 18 training runs (six conditions, three seeds each), LFS and worst-decile retrieval (AaR@10) were positively correlated (Pearson $+0.92$): runs with a lower share also retrieved the hardest sentences less well. In the earlier 128-language pilot, fine-tuning with LFS itself as a loss nearly doubled the decrease (0.206 to 0.389) within 400 steps, while average Belebele accuracy fell 2.9 points against the control's 1.7. A variance-preservation penalty did not stop the shrinking at any weight tested (1 to 1,000). LFS alone is therefore not supported as a training objective.

8 Limitations

LFS measures the balance of the two average effects: it excludes interactions from its ratio, is blind to a common scaling of all vectors, and its level depends on the grid, pooling, and standardization (Appendix J); a small share does not mean that language information is gone or that meaning is preserved. The across-model benchmark, cosine, and gain-beyond-retrieval comparisons were planned after the outcomes had been seen, and related model families limit the independence of observations. Validation is confined to translated text except for the INCLUDE and WMT19 controls, and the cross-model associations are observational. Likelihood scoring isolates context use from decoding settings but does not test generated output. An exploratory continuation study found no clear association at the preselected layer between LFS and automatically detected off-language output. The share of confidently classified outputs also varied with LFS, and an AI-assisted reading found false detections and missed language changes (Appendix O).

9 Conclusion

Language Factor Share makes the balance between average language differences and sentence differences measurable across a model's layers. Across 33 open base models, language share was lowest at an intermediate layer and higher at the input and output. This pattern repeated on a second collection of translations for 17 models, while the size of the decrease varied widely among similarly sized models. Lower language share was also associated with higher average benchmark accuracy and greater benefit from relevant context in another language when predicting English text. We found no clear association with the performance gap relative to English. The practical value of LFS is to help researchers locate where this balance changes and, from its two separately reported parts, see whether the language part, the sentence part, or both changed. In our training experiments, checking representation size separately revealed substantial shrinking that the share alone missed. These experiments did not establish that lowering LFS improves performance, so a smaller share is not proof of a better model. LFS therefore provides a way to investigate multilingual structure and interpret its changes, while direct tests remain necessary to assess model performance.

Reproducibility statement

The main representation grids derive from public NTREX-128 text and public base checkpoints: the 19 expansion models at pinned revisions, and the primary-set grids and the 14 saved-vector dumps

identified by Hugging Face id (Appendix E). The estimator is defined in Eqs. 2 and 3; Appendix A proves the decomposition identity and Appendix B derives the noise reference. A manifest links every main result to its saved evidence and, where available, its producing script, input hashes, and job identifier; it identifies the results available only through logged runs (the original-versus-translated shift, the whitening comparison, and the training-run Belebele deltas). Predictions for the new model families, the synthetic fault test suite, the intervention conditions, and the 19-model context-transfer expansion were frozen in dated documents before the corresponding runs, and every failed prediction is reported; Appendix R summarizes the corrections. The supplementary material contains the analysis code, saved outputs, a source-to-result manifest, and the preregistration documents. From them, the figures can be redrawn from saved summary values, and statistics computed from saved per-language and per-model observations can be recomputed on CPU; the manifest identifies results that are available only as stored summaries or logs, and inputs that were not included. Recomputing the representation measurements themselves, rerunning training, and rescoring benchmarks require the public checkpoints, GPUs, and the NTREX-128 and FLORES-200 data; the 1,500-sentence embedding dumps are not included because of their size. Model revisions are pinned for the 19 expansion models only.

Ethics statement

This work analyzes publicly released base language models on public benchmark text and involves no human subjects. Its aim is to improve the evaluation of multilingual models so that non-English users are served better.

AI use statement

We used generative AI assistants to implement and review analysis code, including the code that generates the synthetic test grids; to design and check analysis protocols; to draft and check mathematical derivations, including the decomposition identity and the noise reference value; to interpret results; to draft and edit text; to produce figures; to screen model outputs in the exploratory continuation study (Appendix O); and to organize the literature review. No reported number was generated as text by an AI assistant: every number was computed by our analysis scripts on the stated inputs and checked against the saved artifacts by the authors, and the derivations were checked numerically against synthetic grids (Appendix B). We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1541–1568, 2025. doi: 10.18653/v1/2025.ijcnlp-long.83. URL https://aclanthology.org/2025.ijcnlp-long.83/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Deniz Bayazit, Badr AlKhamissi, and Antoine Bosselut. Lingua franca or probing artifact? Rethinking latent language in multilingual LLMs. *arXiv preprint arXiv:2609.00155*, 2026. URL https://arxiv.org/abs/2609.00155.

J. Ben Tamo, Daniel Carlander-Reuterfelt, Jonathan Rubin, Oleg Poliannikov, Dezhi Hong, and Mingxian Wang. LinguaMap: Which layers of LLMs speak your language and how to tune them? In *International Conference on Learning Representations*, pp. 33–58, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/00295cede6e1600d344b5cd6d9fd4640-Abstract-Conference.html.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/2002.03518.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.9. URL https://aclanthology.org/2022.emnlp-main.9/.

Nadezhda Chirkova and Vassilina Nikoulina. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pp. 695–708, 2024. doi: 10.18653/v1/2024.inlg-main.53. URL https://aclanthology.org/2024.inlg-main.53/.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.04087.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482/.

Shuoyang Ding and Philipp Koehn. Evaluating saliency methods for neural language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5034–5052, 2021. doi: 10.18653/v1/2021.naacl-main.399. URL https://aclanthology.org/2021.naacl-main.399/.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, 2022. doi: 10.18653/v1/2022.sumeval-1.4. URL https://aclanthology.org/2022.sumeval-1.4/.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925. URL https://wellcomecollection.org/works/vb9nd67r.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, 2025. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2024. URL https://glottolog.org/.

Songbo Hu, Ivan Vulić, and Anna Korhonen. Quantifying language disparities in multilingual large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4003–4018, 2025. doi: 10.18653/v1/2025.emnlp-main.199. URL https://aclanthology.org/2025.emnlp-main.199/.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2110.09348.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. MEXA: Multilingual evaluation of English-centric LLMs via cross-lingual alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. doi: 10.18653/v1/2025.findings-acl.1385. URL https://aclanthology.org/2025.findings-acl.1385/.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pp. 414–435, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.wmt-1.23. URL https://aclanthology.org/2025.wmt-1.23/.

Philipp Koehn and Kevin Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pp. 9–16, 2002. doi: 10.3115/1118627.1118629. URL https://aclanthology.org/W02-0902/.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6919–6971, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.384. URL https://aclanthology.org/2024.naacl-long.384/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3392–3405, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.300. URL https://aclanthology.org/2021.findings-acl.300/.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1565–1575, 2019. doi: 10.18653/v1/D19-1167. URL https://aclanthology.org/D19-1167/.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. Locating language-specific information in contextualized embeddings. *arXiv preprint arXiv:2109.08040*, 2021. URL https://arxiv.org/abs/2109.08040.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1663–1674, 2020. doi: 10.18653/v1/2020.findings-emnlp.150. URL https://aclanthology.org/2020.findings-emnlp.150/.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3125–3135, 2019. doi: 10.18653/v1/P19-1301. URL https://aclanthology.org/P19-1301/.

Kelly Marchisio, Youngser Park, Ali Saad-Eldin, Anton Alyakin, Kevin Duh, Carey Priebe, and Philipp Koehn. An analysis of Euclidean vs. graph-based framing for bilingual lexicon induction from word embedding spaces. In *Findings of the Association for Computational Linguistics:*

EMNLP 2021, pp. 738–749, 2021. doi: 10.18653/v1/2021.findings-emnlp.64. URL https://aclanthology.org/2021.findings-emnlp.64/.

Kelly Marchisio, Neha Verma, Kevin Duh, and Philipp Koehn. IsoVec: Controlling the relative isomorphism of word embedding spaces. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6019–6033, 2022. doi: 10.18653/v1/2022.emnlp-main.404. URL https://aclanthology.org/2022.emnlp-main.404/.

Boris Marinov, Angira Sharma, Christian Schroeder de Witt, Philip Torr, Anisoara Calinescu, and Jialin Yu. When language representations interact: Separability and cross-lingual effects in LLMs. *arXiv preprint arXiv:2606.14347*, 2026. URL https://arxiv.org/abs/2606.14347.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021. doi: 10.18653/v1/2021.eacl-main.189. URL https://aclanthology.org/2021.eacl-main.189/.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4226–4237, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.377/.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Qiwei Peng and Anders Søgaard. Concept space alignment in multilingual LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5511–5526, 2024. doi: 10.18653/v1/2024.emnlp-main.315. URL https://aclanthology.org/2024.emnlp-main.315/.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1608. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/74bb24dca8334adce292883b4b651eda-Abstract-Conference.html.

Wessel Poelman and Miryam de Lhoneux. Form and meaning in intrinsic multilingual evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2503–2521, 2026. doi: 10.18653/v1/2026.eacl-long.113. URL https://aclanthology.org/2026.eacl-long.113/.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010. URL https://jmlr.org/papers/v11/radovanovic10a.html.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ced46a50befedcb884ccf0cbe8c3ad23-Abstract-Conference.html.

Cheril Shah, Yashashree Chandak, Atharv Mahesh Mane, Benjamin K. Bergen, and Tyler A. Chang. Correlations between multilingual language model geometry and crosslingual transfer performance. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4059–4066, 2024. URL https://aclanthology.org/2024.lrec-main.361/.

Nadav Shani and Ali Basirat. Language dominance in multilingual large language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 137–148, 2025. doi: 10.18653/v1/2025.blackboxnlp-1.7. URL https://aclanthology.org/2025.blackboxnlp-1.7/.

Hannah Sterz, Fabian David Schmidt, Goran Glavaš, and Ivan Vulić. ReCoVeR the target language: Language steering without sacrificing task performance. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19390–19405, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1056. URL https://aclanthology.org/2025.findings-emnlp.1056/.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2402.17762.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Hinata Tezuka and Naoya Inoue. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31742–31792, 2025. doi: 10.18653/v1/2025.emnlp-main.1618. URL https://aclanthology.org/2025.emnlp-main.1618/.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Alexander Tsvetkov and Alon Kipnis. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7971–7989, 2024. doi: 10.18653/v1/2024.findings-emnlp.468. URL https://aclanthology.org/2024.findings-emnlp.468/.

Manan Uppadhyay, Prashant Kodali, Pranjal Chitale, Reshma Ramaprasad, Himanshu Beniwal, and Sunayana Sitaram. DEPART: DEcomposing PARiTy across multilingual LLMs. *arXiv preprint arXiv:2605.28163*, 2026. URL https://arxiv.org/abs/2605.28163.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? On the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/84fb08ae30ab266953d1ff72eb0611b1-Abstract-Conference.html.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.764. URL https://aclanthology.org/2020.acl-main.764/.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.379. URL https://aclanthology.org/2022.emnlp-main.379/.

Haoran Xu and Philipp Koehn. Cross-lingual BERT contextual embedding space mapping with isotropic and isometric conditions. *arXiv preprint arXiv:2107.09186*, 2021. URL https://arxiv.org/abs/2107.09186.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. How does alignment enhance LLMs' multilingual capabilities? a language neurons perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34799–34807, 2026. doi: 10.1609/aaai.v40i41.40782. URL https://ojs.aaai.org/index.php/AAAI/article/view/40782.

Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. Lens: Rethinking multilingual enhancement for large language models. *arXiv preprint arXiv:2410.04407*, 2024. URL https://arxiv.org/abs/2410.04407.

Xiutian Zhao, Philipp Koehn, Björn Schuller, and Berrak Sisman. Multilingual emotion neurons in large audio-language models. *arXiv preprint arXiv:2608.08772*, 2026. URL https://arxiv.org/abs/2608.08772.

A Decomposition identity

Let $\boldsymbol{a}_\ell = \bar{\boldsymbol{z}}_\ell - \bar{\boldsymbol{z}}$, $\boldsymbol{b}_c = \bar{\boldsymbol{z}}_c - \bar{\boldsymbol{z}}$, and $\boldsymbol{r}_{\ell c} = \boldsymbol{z}_{\ell c} - \bar{\boldsymbol{z}}_\ell - \bar{\boldsymbol{z}}_c + \bar{\boldsymbol{z}}$, so that $\boldsymbol{z}_{\ell c} - \bar{\boldsymbol{z}} = \boldsymbol{a}_\ell + \boldsymbol{b}_c + \boldsymbol{r}_{\ell c}$. On a complete grid $\sum_\ell \boldsymbol{a}_\ell = 0$, $\sum_c \boldsymbol{b}_c = 0$, $\sum_c \boldsymbol{r}_{\ell c} = 0$ for every ℓ, and $\sum_\ell \boldsymbol{r}_{\ell c} = 0$ for every c. Expanding $\sum_{\ell c} \|\boldsymbol{a}_\ell + \boldsymbol{b}_c + \boldsymbol{r}_{\ell c}\|^2$, the cross term $2\sum_{\ell c} \boldsymbol{a}_\ell^\top \boldsymbol{b}_c = 2(\sum_\ell \boldsymbol{a}_\ell)^\top (\sum_c \boldsymbol{b}_c) = 0$, the cross term $2\sum_{\ell c} \boldsymbol{a}_\ell^\top \boldsymbol{r}_{\ell c} = 2\sum_\ell \boldsymbol{a}_\ell^\top (\sum_c \boldsymbol{r}_{\ell c}) = 0$, and likewise for $\boldsymbol{b}_c$. Hence $\sum_{\ell c} \|\boldsymbol{z}_{\ell c} - \bar{\boldsymbol{z}}\|^2 = N\sum_\ell \|\boldsymbol{a}_\ell\|^2 + L\sum_c \|\boldsymbol{b}_c\|^2 + \sum_{\ell c} \|\boldsymbol{r}_{\ell c}\|^2 = \mathrm{SS}_{\mathrm{lang}} + \mathrm{SS}_{\mathrm{sent}} + \mathrm{SS}_{\mathrm{res}}$. The multipliers N and L arise because each language mean is repeated across N sentences and each sentence mean across L languages.

B Noise reference

Suppose the cells were generated as $\boldsymbol{z}_{\ell c} = \boldsymbol{C}_c + \boldsymbol{G}_\ell + \boldsymbol{\varepsilon}_{\ell c}$, with fixed sentence vectors $\boldsymbol{C}_c$, fixed language vectors $\boldsymbol{G}_\ell$, and i.i.d. isotropic noise $\boldsymbol{\varepsilon}_{\ell c}$ of variance σ^2 per coordinate. Then the language means are $\bar{\boldsymbol{z}}_\ell = \boldsymbol{G}_\ell + \bar{\boldsymbol{C}} + \bar{\boldsymbol{\varepsilon}}_{\ell\cdot}$, and $\mathbb{E}\, N\sum_\ell \|\bar{\boldsymbol{\varepsilon}}_{\ell\cdot} - \bar{\boldsymbol{\varepsilon}}\|^2 = (L-1)D\sigma^2$ by the usual between-groups degrees of freedom. Thus $\mathbb{E}[\mathrm{SS}_{\mathrm{lang}}] = (L-1)D\sigma^2 + N\sum_\ell \|\boldsymbol{G}_\ell - \bar{\boldsymbol{G}}\|^2$ and symmetrically $\mathbb{E}[\mathrm{SS}_{\mathrm{sent}}] = (N-1)D\sigma^2 + L\sum_c \|\boldsymbol{C}_c - \bar{\boldsymbol{C}}\|^2$. With no structure the ratio of expectations is $(L-1)/(L+N-2)$, the noise reference of Eq. 4; this is a ratio of expected sums of squares, not the exact expectation of the ratio after empirical standardization; the noise reference is not a lower bound, since strong sentence structure with weak language structure drives LFS below it. Synthetic checks across $L \in [2, 128]$, $N \in [50, 3000]$, $D \in [64, 1024]$ and pseudo-language splits of real representations (0.0015 measured against 0.00143 predicted at $(2, 700)$; 0.0849 against 0.0846 at $(12, 120)$) confirm the formula. In the same saved runs (four models, 20 repetitions each, at each model's layer of minimum language share), real grids of English plus 11 other languages at $L = 12$, $N = 120$ were compared with artificial groups formed by splitting English sentences into 12 disjoint groups of 120: every model's lowest real-grid value (0.24 to 0.67) exceeds its highest artificial-group value (0.088 to 0.090). This separates real language structure from this control; it does not show that LFS captures all language information. LFS is nearly but not exactly size-independent: on fixed language sets it moves by -0.006 to -0.010 from N=100 to N=1000, against -0.065 to -0.125 for a retrieval-based score on the same draws.

C The REML variant (LFS-VC)

Treating Eq. 2 as a balanced two-way crossed random-effects layout with one observation per cell, the expected mean squares are $\mathbb{E}[\mathrm{MS}_L] = \sigma_E^2 + N\sigma_L^2$, $\mathbb{E}[\mathrm{MS}_C] = \sigma_E^2 + L\sigma_C^2$, $\mathbb{E}[\mathrm{MS}_E] = \sigma_E^2$, giving noise-corrected variance components per coordinate; the non-negative REML solution sets a negative component to zero. Summing over coordinates, $\text{LFS-VC} = \sigma_L^2/(\sigma_L^2 + \sigma_C^2)$. The two estimators are

mathematically distinct, and their levels are not comparable: LFS-VC subtracts an estimated noise component, and the expansion profiles also use a smaller grid (Llama-3.1-8B minimum 0.626 under standard LFS on the primary grid against 0.479 under LFS-VC on the 41-language expansion grid). Standard LFS is primary for the profile results; the expansion set and the context-transfer analysis use LFS-VC and are labeled as such throughout. Recalculation of standard LFS from the saved standardized sufficient statistics (September 2026) selected the same minimum layer as LFS-VC in all 19 expansion models; the adjusted context-transfer associations were -0.461 and -0.459, respectively, and the model-level Spearman correlations -0.763 and -0.751. These comparisons support the robustness of that association to estimator choice in this set, without making the two estimators' levels interchangeable. A completeness check over the 19 models and their 637 layers found no coordinate at any reported layer in the branch where both initial component estimates are negative. The two estimators are still reported separately and never pooled, and the historical LFS-VC analyses keep their label.

D MAPPING ANALYSIS

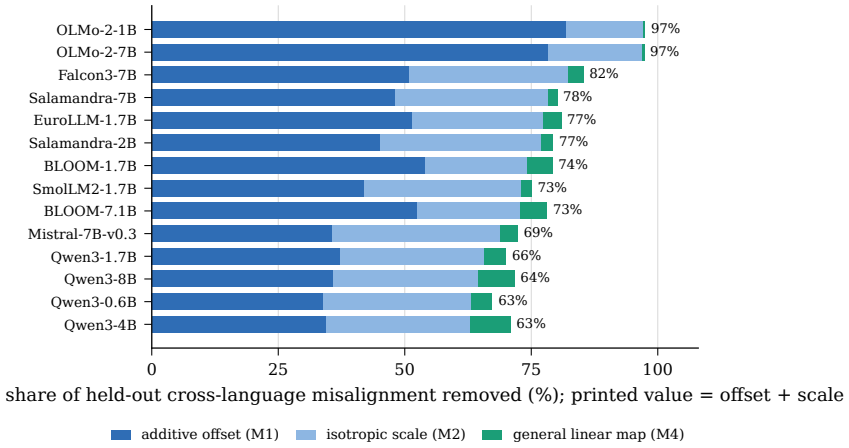

Figure 3: Mapping analysis at the layer of minimum language share for the 14-model study (N=1500, common rank 64). Additive offset and isotropic scale dominate every model; the rotation share, given per model in Table 2, never exceeds 0.7%.

Figure 3 shows the decomposition and Table 2 gives the per-model held-out shares at the common rank. Each language is mapped into a generalized-Procrustes consensus (an average space fitted across languages); a map is credited only with its held-out error reduction beyond per-language permutation nulls, on 20 cross-fitted splits. The procedure recovers injected rotations of 0.1 to 0.8 rad, and additional orthogonal alignment provides little held-out improvement under this protocol. The document-purged re-audit on Qwen3-1.7B leaves offset plus scale at about 64% and rotation at about 0.1% (same-document leakage in the original splits). Per (model, language) pair, a linear or simpler map is selected in 74% of 1,792 cases; the 26% nonlinear selections concentrate in three models (Falcon3-7B, OLMo-2-1B, OLMo-2-7B).

E MODELS AND REVISIONS

The 19 expansion models were run at the pinned Hugging Face revisions below. The 17 primary-set models are listed in Table 4; their saved grid and dump metadata record the Hugging Face id only, so their snapshot revisions are not pinned in the released artifacts.

Table 2: Mapping analysis at each model's layer of minimum language share for the 14-model study: mean held-out error-reduction share across 128 languages at the common rank 64, for the nested maps offset (M1), isotropic scale (M2), rotation (M3), general linear (M4), and nonlinear kernel (M5). A negative nonlinear share is held-out overfitting of the kernel map, reported as measured.

Model	Min. layer	Offset	Scale	Rotation	Linear	Nonlinear	Offset+scale
BLOOM-1.7B	14	54.1%	20.1%	0.6%	5.0%	−6.3%	74.2%
BLOOM-7.1B	17	52.4%	20.5%	0.7%	5.2%	−7.0%	72.9%
EuroLLM-1.7B	14	51.5%	25.8%	0.4%	3.6%	−1.7%	77.4%
Falcon3-7B	13	50.9%	31.4%	0.1%	3.0%	1.1%	82.3%
Mistral-7B-v0.3	2	35.7%	33.2%	0.0%	3.4%	−10.5%	68.9%
OLMo-2-1B	9	81.8%	15.4%	−0.2%	0.3%	0.5%	97.2%
OLMo-2-7B	12	78.4%	18.5%	−0.2%	0.4%	0.7%	96.9%
Qwen3-0.6B	8	33.8%	29.4%	0.0%	4.1%	−12.4%	63.2%
Qwen3-1.7B	13	37.2%	28.6%	0.1%	4.2%	−10.8%	65.8%
Qwen3-4B	17	34.5%	28.4%	0.1%	8.0%	−12.1%	62.9%
Qwen3-8B	19	35.9%	28.5%	0.1%	7.3%	−9.5%	64.4%
Salamandra-2B	6	45.2%	31.8%	−0.1%	2.3%	−1.2%	77.0%
Salamandra-7B	18	48.2%	30.1%	0.1%	1.8%	1.2%	78.3%
SmolLM2-1.7B	15	42.0%	31.0%	0.1%	2.1%	−5.3%	73.0%

Tag	Hugging Face id	Revision (first 12)	Family
Qwen2.5-0.5B	Qwen/Qwen2.5-0.5B	`060db6499f32`	Qwen2.5
Qwen2.5-1.5B	Qwen/Qwen2.5-1.5B	`8faed761d45a`	Qwen2.5
Qwen2.5-3B	Qwen/Qwen2.5-3B	`3aab1f1954e9`	Qwen2.5
Qwen2.5-7B	Qwen/Qwen2.5-7B	`d14972939875`	Qwen2.5
xglm-1.7B	facebook/xglm-1.7B	`d23a5e8e2164`	XGLM
xglm-2.9B	facebook/xglm-2.9B	`33c659ae27de`	XGLM
xglm-7.5B	facebook/xglm-7.5B	`732d59308a84`	XGLM
mGPT	ai-forever/mGPT	`40897bd7c8b4`	mGPT
OLMo-2-1124-13B	allenai/OLMo-2-1124-13B	`3fefddc1bf18`	OLMo-2
granite-3.1-8b-base	ibm-granite/granite-3.1-8b-base	`39975ba90995`	Granite
Yi-1.5-9B	01-ai/Yi-1.5-9B	`80d5471b1eae`	Yi-1.5
TowerBase-7B	Unbabel/TowerBase-7B-v0.1	`bd59c10d7600`	Tower
occiglot-7b-eu5	occiglot/occiglot-7b-eu5	`72d8b932ce09`	Occiglot
Sailor-1.8B	sail/Sailor-1.8B	`c687a96d5c3e`	Sailor
Sailor-7B	sail/Sailor-7B	`b8b49a0f0207`	Sailor
Llama-3.1-8B	meta-llama/Llama-3.1-8B	`d04e592bb4f6`	Llama-3.1
SmolLM2-360M	HuggingFaceTB/SmolLM2-360M	`f8027fd0eaee`	SmolLM2
Falcon3-3B-Base	tiiuae/Falcon3-3B-Base	`092c29e3114f`	Falcon3
Mistral-Nemo-Base-2407	mistralai/Mistral-Nemo-Base-2407	`a4477a2f9779`	Mistral

F PER-MODEL LAYER PROFILES

Table 4 lists every audited profile. The decrease is the layer-0 value (column First) minus the minimum.

G CONFOUNDER-ADJUSTMENT REGRESSION DETAILS

The adjustment regresses each outcome, transformed as $\mathrm{logit}((a - 0.25)/0.75)$ for accuracy a so that the four-choice chance level 0.25 maps to the bottom of the scale, on model fixed effects, $\log_{10}$ CulturaX tokens, Glottolog macro-family, script, and tokenizer fertility, with cluster-robust standard errors by language. On the 17-model grid (1,241 model-language rows) the fit has $R^2 = 0.824$; the token coefficient is $+0.286$ (SE 0.059, $z = 4.8$) and the fertility coefficient -0.103 (SE 0.022); no macro-family or script dummy is individually significant. An earlier development-grid run (10 models, 1,173 rows) gave a token coefficient of $+0.267$ (SE 0.051) and $R^2 = 0.86$; we report the 17-model values. Rows lacking a token count or a fertility value are dropped before fitting, which is why the

Table 3: Analysis populations. Each row lists what a result in the main text was computed on.

Analysis	Models	Languages	Sentences	Estimator	Outcome or output
Layer profiles, primary (Figure 1)	17	128	300	standard LFS	minimum layer, minimum LFS, decrease
Layer profiles, expansion	19	41 (English + 40)	300	LFS-VC	same
FLORES-200 replication	17	116	300	standard LFS	same
Sentence-subsampling intervals	14	128	300 of 1,500	standard LFS	decrease at fixed layers
Mapping analysis	14	128	1,500	nested maps M1–M5	held-out error reduction
English vs. multilingual-average predictor	17	127 targets	300	ridge on 64 components	3-fold cross-validated R^2
Context transfer	19	41 (34 scored for all)	100 pairs	LFS-VC	R_{ctx}
Model-level benchmark sensitivity (retrospective)	33 (14 + 19)	40 or 30 common	1,500 or 300	LFS-VC, two grids	model-mean accuracy
Pooled benchmark panel	33	69 of 73	1,726 of 2,001 rows	MEXA, AaR@10	pooled Belebele and INCLUDE residual
Word-alignment, six-condition, and weight-sweep evaluations	Qwen3-0.6B	4 (eng, deu, hin, swa), fixed layer 8	300	standard LFS	components, MEXA, AaR, CVP
Earlier direct-LFS pilot	Qwen3-0.6B	128, full profiles	300	standard LFS	layer profiles, decrease
Cosine and pairwise measures (follow-up)	14 with saved vectors	127	300	standard LFS; raw and standardized measures	69 model–layer records; 8,763 language rows
Comparison with benchmarks (follow-up)	14 with saved vectors	70 (Belebele), 30 (INCLUDE) per model, median	300	per-language measures	within- and across-model ρ
Incremental prediction (follow-up)	19	34	100 pairs	LFS-VC, MEXA	leave-one-family-out MAE, held-out ρ
Matched-preprocessing cosine (exploratory)	19	35 (English + 34)	300	standard LFS; raw, centered, standardized cosine	mean R_{ctx}, model ρ
Short-continuation persistence (exploratory)	19	11 + English control	20 prompts per language	standard LFS, 41-language grid	detected off-language share

analyzed row counts in this appendix are below the panel sizes. The residual α from this fit is the adjusted outcome; the same design without model fixed effects is used for the context-transfer analysis. In the follow-up analyses, macro-family and script enter as one indicator column per category with the first category in alphabetical order as reference, rows lacking a token count or a fertility value are dropped, and the reproduced clipped-logit transform is $\mathrm{logit}((a - 0.25)/0.75)$ with accuracy clipped to $[0.26, 0.99]$, which clipped 181 of 980 Belebele rows; the follow-up's primary rows use raw accuracy. Removing one covariate at a time from the within-model benchmark adjustment, or replacing the raw outcome by the original clipped logit, flips the sign of 7 of 60 adjusted coefficients of the per-language measures, all of them measures whose fully adjusted $|\rho|$ is at most 0.13 (the

Table 4: All 36 audited layer profiles: 17 standard-LFS profiles on the primary grid (128 languages, 300 sentences; noise reference 0.298) and 19 REML LFS-VC profiles on the expansion grid (English plus 40 languages, 300 sentences; the standard-LFS reference for this grid would be $40/339 = 0.118$, which is not a baseline for the noise-corrected estimator). Three models appear in both; levels are not comparable across the two blocks.

Model	Layers	Min. layer	Rel. depth	First	Minimum	Last	Decrease
Primary set, standard LFS							
Qwen3-8B-Base	37	19	0.53	0.945	0.608	0.958	0.336
Llama-3.1-8B	33	14	0.44	0.961	0.626	0.936	0.335
Qwen3-4B-Base	37	17	0.47	0.916	0.602	0.956	0.314
salamandra-2b	25	6	0.25	0.966	0.712	0.935	0.254
Qwen3-1.7B-Base	29	13	0.46	0.927	0.715	0.960	0.213
Mistral-7B-v0.3	33	2	0.06	0.956	0.744	0.963	0.212
Qwen3-0.6B-Base	29	8	0.29	0.933	0.728	0.958	0.206
granite-3.1-8b-base	41	14	0.35	0.898	0.709	0.969	0.190
salamandra-7b	33	18	0.56	0.963	0.784	0.938	0.180
Falcon3-7B-Base	29	13	0.46	0.967	0.826	0.962	0.140
OLMo-2-1124-7B	33	12	0.38	0.981	0.846	0.956	0.135
Yi-1.5-9B	49	20	0.42	0.973	0.860	0.963	0.113
EuroLLM-1.7B	25	14	0.58	0.951	0.854	0.937	0.097
OLMo-2-0425-1B	17	9	0.56	0.980	0.903	0.958	0.077
SmolLM2-1.7B	25	15	0.62	0.972	0.904	0.973	0.068
bloom-7b1	31	17	0.57	0.927	0.884	0.953	0.043
bloom-1b7	25	14	0.58	0.938	0.899	0.948	0.039
Expansion set, REML LFS-VC							
Mistral-Nemo-Base-2407	41	13	0.33	0.907	0.377	0.822	0.530
Llama-3.1-8B	33	14	0.44	0.966	0.479	0.883	0.487
Qwen2.5-7B	29	14	0.50	0.953	0.501	0.911	0.452
Qwen2.5-1.5B	29	15	0.54	0.923	0.521	0.939	0.403
Qwen2.5-3B	37	19	0.53	0.924	0.539	0.932	0.385
TowerBase-7B	33	2	0.06	0.974	0.665	0.943	0.309
Qwen2.5-0.5B	25	11	0.46	0.909	0.629	0.935	0.280
Sailor-7B	33	14	0.44	0.934	0.656	0.954	0.279
occiglot-7b-eu5	33	2	0.06	0.969	0.699	0.968	0.270
granite-3.1-8b-base	41	15	0.38	0.911	0.655	0.958	0.256
OLMo-2-1124-13B	41	19	0.47	0.985	0.740	0.916	0.244
mGPT	25	13	0.54	0.951	0.709	0.861	0.242
Yi-1.5-9B	49	20	0.42	0.984	0.803	0.950	0.181
xglm-2.9B	49	21	0.44	0.936	0.770	0.981	0.166
xglm-1.7B	25	13	0.54	0.944	0.783	0.987	0.161
Sailor-1.8B	25	11	0.46	0.951	0.825	0.956	0.126
SmolLM2-360M	33	20	0.62	0.979	0.874	0.983	0.106
Falcon3-3B-Base	23	11	0.50	0.979	0.878	0.972	0.101
xglm-7.5B	33	21	0.66	0.790	0.718	0.850	0.072

pairwise share on INCLUDE, the language contribution, and the offset norm); the measures above 0.3 keep their sign.

Adjusted benchmark estimates by support. Raw per-language correlations of 0.50 to 0.54 (MEXA and AaR@10, 17-model grid) fall to 0.19 to 0.31 after adjustment. For MEXA the adjusted correlation is 0.167 on 14 models (522 model–language rows; $[0.10, 0.23]$ resampling languages, $[-0.07, 0.41]$ resampling models) and 0.220 $[0.16, 0.26]$ on the 33-model panel (1,726 rows; resampling languages). Because the adjustment removes each model's mean, a predictor with one value per model, such as global LFS or LFS-VC, is uninformative by construction here, and we do not report it.

G.1 Retrospective model-level benchmark sensitivity

As part of the submission validation we also correlated each model's noise-corrected share (LFS-VC) with its mean benchmark accuracy, one observation per model. The analysis is a retrospective audit of saved outputs, not a preregistered test. The panel combines two cohorts: the 14 primary models with 1,500-sentence dumps, with LFS-VC from 128-language, 1,500-sentence component grids, and the 19 expansion models, with LFS-VC from 41-language, 300-sentence grids, each at the layer recorded in its own measurement artifact (mGPT with its corrected value; Appendix R). Models with both a primary and an expansion measurement use their frozen expansion measurement, a rule that does not depend on the outcomes; the alternative measurements are kept separately. For

each cohort and outcome, accuracy is averaged equally over the languages scored for every model in that cohort, and the Spearman correlation (average ranks) is taken across models, with 2,000 model-bootstrap replicates (seed 13). There is no covariate adjustment, and model resampling is not family resampling. Because the two cohorts differ in grid size and layer selection, cohort membership can contribute to the pooled coefficient, so the 33-model rows are mixed-protocol sensitivities, not a harmonized validation; they neither establish causality nor contradict the weak adjusted associations above, which use a different unit. The pooled accuracy row averages, for each model and language, whichever of Belebele and INCLUDE is available before averaging over languages; it is secondary. Item-level scoring logs are incomplete for part of the benchmark panel. Table 5 gives all nine rows. The primary-cohort Belebele coefficient equals, by coincidence, the standard-LFS value of Section 6 (-0.653); the estimators, grids, and intervals differ.

Table 5: Retrospective associations between LFS-VC and model-mean benchmark accuracy; one observation per model; Spearman ρ with 95 percent model-bootstrap intervals (2,000 replicates). The stored results are correlations with the sentence share $1 - \text{LFS-VC}$; here each coefficient is negated and each interval reversed. The 33-model rows combine two representation grids and layer-selection protocols.

Cohort	Measurement grid	Outcome (common languages)	ρ	95% CI
All 33 models	mixed	Belebele (40)	-0.567	$[-0.744, -0.282]$
All 33 models	mixed	INCLUDE (30)	-0.676	$[-0.830, -0.414]$
Primary 14	$128 \times 1{,}500$	Belebele (70)	-0.653	$[-0.956, -0.065]$
Primary 14	$128 \times 1{,}500$	INCLUDE (30)	-0.679	$[-0.960, -0.120]$
Expansion 19	41×300	Belebele (40)	-0.583	$[-0.854, -0.189]$
Expansion 19	41×300	INCLUDE (30)	-0.735	$[-0.940, -0.368]$
Secondary: available-target pooled accuracy				
All 33 models	mixed	pooled (40)	-0.583	$[-0.754, -0.299]$
Primary 14	$128 \times 1{,}500$	pooled (70)	-0.648	$[-0.960, -0.063]$
Expansion 19	41×300	pooled (40)	-0.634	$[-0.871, -0.239]$

H CONTEXT-TRANSFER PROTOCOL

The candidate measure is one number per model fixed in advance, LFS-VC at the layer of minimum language share, taken from the expansion-set estimator (the protocol states the association for the sentence share $1 - \text{LFS-VC}$; we report it for LFS-VC, which flips the sign of every correlation and interval and changes nothing else). The standard-LFS recalculation from saved sufficient statistics selects the same minimum layer in all 19 models and gives an adjusted association of -0.461 against -0.459 (Appendix C); this result is reported as an LFS-VC result because that is the estimator the frozen protocol named, and the recalculation is a sensitivity check, not a relabeling. Because it is constant within a model, the 646-row correlation reflects between-model variation with languages as repeated measures; the model-bootstrap interval and the model-level Spearman of -0.751 over 19 models are the statements that account for this. Pairs: for each NTREX document, the first adjacent pair whose context sentence has at least five English words and whose target has 8 to 45 words; the first 100 distinct documents; a deterministic minimum-cost derangement assigns each pair a wrong-document context of matched English length (mean absolute mismatch 0.34 words). Scoring: teacher-forced fp32 NLL over target tokens only, context and target separated by a newline; 41 context languages; 8,300 sequences per model. Ten models initially exceeded a tokenizer-length ceiling and were rerun with a 1,024-token and then a 2,048-token limit; no example, revision, or scoring rule changed. Table 6 gives the per-model values.

Attention (secondary result). On the same pairs (Qwen3-0.6B, 41 languages, frozen predictions), matched context draws more attention than mismatched in 41 of 41 languages, peaking just below the LFS minimum, and content attention tracks alignment after confounder adjustment (0.53 $[0.27, 0.73]$); the attention-to-behavior link does not survive confounder adjustment and is scored as a failure.

Table 6: The 19 models: LFS-VC at the layer of minimum language share and median context-transfer values over the 34 languages scored for every model.

Model	Family	LFS-VC	Median R_{ctx}	Median raw benefit
Mistral-Nemo-Base-2407	Mistral	0.377	0.683	0.658
Llama-3.1-8B	Llama-3.1	0.479	0.679	0.689
Qwen2.5-7B	Qwen2.5	0.501	0.660	0.705
Qwen2.5-1.5B	Qwen2.5	0.521	0.557	0.600
Qwen2.5-3B	Qwen2.5	0.539	0.604	0.665
Qwen2.5-0.5B	Qwen2.5	0.629	0.470	0.493
granite-3.1-8b-base	Granite	0.655	0.552	0.388
Sailor-7B	Sailor	0.656	0.587	0.525
TowerBase-7B	Tower	0.665	0.577	0.413
occiglot-7b-eu5	Occiglot	0.699	0.538	0.466
mGPT	mGPT	0.709	0.583	0.478
xglm-7.5B	XGLM	0.718	0.584	0.474
OLMo-2-1124-13B	OLMo-2	0.740	0.567	0.537
xglm-2.9B	XGLM	0.770	0.549	0.442
xglm-1.7B	XGLM	0.783	0.503	0.417
Yi-1.5-9B	Yi-1.5	0.803	0.554	0.458
Sailor-1.8B	Sailor	0.825	0.409	0.378
SmolLM2-360M	SmolLM2	0.874	0.253	0.234
Falcon3-3B-Base	Falcon3	0.878	0.267	0.235

I Replication, uncertainty, and instruction tuning

Protocol and frozen predictions: a dated addendum written before the runs (in the supplementary material). FLORES-200 devtest, first 300 sentences, 116 of the 128 NTREX languages (English plus 115 of the 127 non-English languages) mapped by an explicit table (regional English and French variants and eight languages absent from FLORES-200 are excluded).

Original versus translated text. On seven WMT19 languages measured on native text and on translations of English (300 pairs per direction, 14 models), translated text carries a slightly higher language share ($+0.02$ to $+0.06$) and slightly lower retrieval alignment, but model ordering is preserved at Spearman 0.99: comparative conclusions survive.

Instruction tuning. For six base and instruct pairs (Qwen3 0.6B to 8B, Qwen2.5-7B, Llama-3.1-8B) the profile is unchanged within 0.01 in the decrease (Table 9), the minimum layer agrees within two layers for five of six, and the final-layer share is higher for the instruct model in five of six pairs by at most 0.003 (Table 9): the profile is largely preserved across these six base and instruction-tuned comparisons, consistent with the zero-shot cross-lingual transfer of English-only instruction tuning reported by Chirkova & Nikoulina (2024).

The failed prediction. Of the twelve predictions frozen for the second-corpus, subsampling, and instruction-tuning runs, eleven passed. The failure is the precision bound: the frozen criterion required a 95 percent half-width below 0.03 for every model, and two of fourteen exceeded it (Mistral-7B 0.045, Salamandra-2B 0.059; Table 8). Both have a sharp minimum at an early layer (2 and 6), where the decrease depends more on which sentences are drawn. The failure is reported as scored; neither model is excluded from any analysis. The ranking check in the same run passed: across all 2,000 sentence subsamples, the ordering of the 14 models by the decrease in language share remained correlated with the reference ordering at Spearman 0.90 or higher (median 0.996). These subsamples use the fixed stored input and minimum layers and shared subsets of 300 of the 1,500 sentences, not minima reselected from fresh complete profiles. The largest cross-corpus change is granite-3.1-8B, whose decrease shrinks from 0.190 on NTREX to 0.074 on FLORES-200 while its per-layer profile still correlates at 0.946 (Table 7); its NTREX decrease should be read with that sensitivity in mind.

Table 7: FLORES-200 replication of the 17-model profile. The decrease is layer-0 LFS minus the minimum. The last two columns are post hoc descriptives: Pearson correlation of the two per-layer profiles and the largest per-layer difference.

Model	Decrease NTREX	Decrease FLORES	Min. layer NTREX	Min. layer FLORES	Intermediate minimum	Profile r	Max. layer diff.
Qwen3-0.6B-Base	0.206	0.187	8	13	yes	0.974	0.074
Qwen3-1.7B-Base	0.213	0.225	13	13	yes	0.974	0.066
Qwen3-4B-Base	0.314	0.332	17	17	yes	0.999	0.037
Qwen3-8B-Base	0.336	0.372	19	19	yes	0.999	0.041
OLMo-2-0425-1B	0.077	0.071	9	9	yes	1.000	0.007
OLMo-2-1124-7B	0.135	0.138	12	12	yes	0.998	0.007
Mistral-7B-v0.3	0.212	0.180	2	15	yes	0.936	0.099
bloom-1b7	0.039	0.039	14	14	yes	0.992	0.008
bloom-7b1	0.043	0.047	17	17	yes	0.995	0.006
EuroLLM-1.7B	0.097	0.100	14	14	yes	0.998	0.007
SmolLM2-1.7B	0.068	0.059	15	15	yes	0.993	0.015
Falcon3-7B-Base	0.140	0.134	13	13	yes	0.980	0.046
salamandra-2b	0.254	0.176	6	14	yes	0.944	0.094
salamandra-7b	0.180	0.190	18	20	yes	0.989	0.030
granite-3.1-8b-base	0.190	0.074	14	15	yes	0.946	0.116
Yi-1.5-9B	0.113	0.120	20	20	yes	0.989	0.023
Llama-3.1-8B	0.335	0.372	14	14	yes	0.990	0.062

Table 8: Sentence-subsampling intervals on the decrease: 2,000 subsets of 300 sentences drawn without replacement from the 1,500-sentence dumps at the fixed stored layers, identical index sets for every model. The grid column is the main first-300-sentence value.

Model	Min. layer	Decrease, first 300 (grid)	Median, 300 of 1,500	95% interval
Qwen3-8B-Base	19	0.336	0.352	[0.341, 0.363]
Qwen3-4B-Base	17	0.314	0.338	[0.324, 0.351]
salamandra-2b	6	0.254	0.313	[0.254, 0.372]
Mistral-7B-v0.3	2	0.212	0.246	[0.199, 0.289]
Qwen3-1.7B-Base	13	0.213	0.229	[0.219, 0.239]
Qwen3-0.6B-Base	8	0.206	0.225	[0.205, 0.246]
salamandra-7b	18	0.180	0.187	[0.181, 0.195]
Falcon3-7B-Base	13	0.140	0.153	[0.141, 0.167]
OLMo-2-1124-7B	12	0.135	0.140	[0.132, 0.149]
EuroLLM-1.7B	14	0.097	0.101	[0.096, 0.107]
OLMo-2-0425-1B	9	0.077	0.081	[0.073, 0.090]
SmolLM2-1.7B	15	0.068	0.075	[0.067, 0.083]
bloom-7b1	17	0.043	0.048	[0.044, 0.053]
bloom-1b7	14	0.039	0.043	[0.039, 0.049]

J MEASUREMENT DEPENDENCE

On five models, the absolute decrease moves by 0.13 to 0.20 between mean pooling and final-token, unit-norm, or length-matched pooling; relative to the mean-pooling ordering, the cross-model rank correlation is 0.90 for final-token and for unit-norm pooling but 0.60 for length-matched truncation, so ranking robustness is scoped to this tested set. Tokenizer fertility correlates with each language's contribution to $\mathrm{SS}_{\mathrm{lang}}$ at $+0.42$ to $+0.64$, so part of the language component is tokenization. Full ZCA whitening, a different preprocessing transformation evaluated at each model's fixed evaluated layer rather than on new full profiles, reshuffles the middle of the 12-model ranking (Spearman 0.45 against a frozen prediction of > 0.8); this is sensitivity to preprocessing, not evidence about where language identity resides (Appendix K).

K ROTATION AND NATIVE COORDINATES

Three transformations should be kept apart. Common scalar rescaling and orthogonal rotation of the already standardized grid leave LFS unchanged. Coordinate standardization of raw representations is part of the estimator, and rotating the raw hidden states before it can change LFS. Full whitening is a different preprocessing transformation; the whitening comparison in Appendix J, at fixed evaluated layers, shows sensitivity to that choice rather than locating language identity in particular directions. Separately, the concentration of language energy in particular native coordinates is not rotation-invariant (median Gini coefficient 0.855 native versus 0.542 after random rotations), so no "language neuron" claim follows.

Table 9: Base versus instruction-tuned checkpoints on the NTREX grid.

Base → instruct	Decrease base	Decrease instruct	Min. layer base	Min. layer instruct	Final LFS base	Final LFS instruct
Qwen3-0.6B-Base → Qwen3-0.6B	0.206	0.211	8	8	0.958	0.956
Qwen3-1.7B-Base → Qwen3-1.7B	0.213	0.216	13	8	0.960	0.963
Qwen3-4B-Base → Qwen3-4B	0.314	0.320	17	17	0.956	0.957
Qwen3-8B-Base → Qwen3-8B	0.336	0.343	19	18	0.958	0.960
Qwen2.5-7B → Qwen2.5-7B-Instruct	0.287	0.295	4	4	0.959	0.961
Llama-3.1-8B → Llama-3.1-8B-Instruct	0.335	0.332	14	14	0.936	0.937

L THE SHRINKING CONDITION AS A RATIO AND AS COMPONENTS

Figure 4 shows the pretrained checkpoint and four of the six training conditions of §7 read two ways: as the LFS ratio, and as the raw variance of the sentence means, which the ratio does not show.

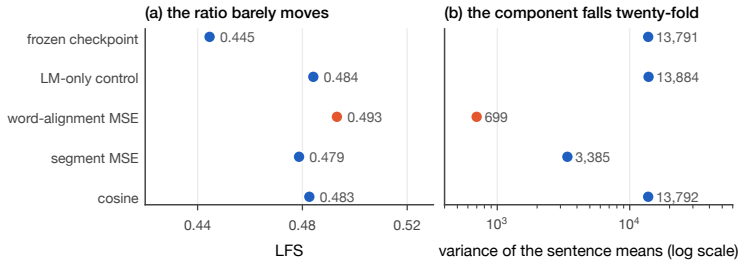

Figure 4: The pretrained checkpoint and four of the six training conditions on Qwen3-0.6B. (a) LFS separates the shrunk word-alignment condition from the cosine condition by 0.01. (b) The raw variance of the sentence means separates them by a factor of twenty. Measurements use the four-language, 300-sentence evaluation grid at fixed layer 8; trained-condition points average three seeds, and the frozen point is the pretrained checkpoint.

M COMPLEMENTARY DIAGNOSTICS

Table 10 lists the diagnostics computed on the same grid as LFS, the question each answers, and what it exposes. Two transitions are visible through depth: input-script mass read out by logit-lens decoding falls early (relative depth 0.05 to 0.19), while the LFS minimum lies at relative depth 0.3 to 0.5 in all four models tested, so the logit-lens readout of the input script falls before the language share reaches its minimum. Middle-layer states are far from every output token embedding (maximum cosine about 0.10); this geometric comparison alone does not identify an internal language.

Table 10: The intrinsic measures used in this paper. MEXA (Kargaran et al., 2025) is an external retrieval-based measure reported alongside for reference.

Question	Measure	What it exposes
Language vs. sentence share, per layer	LFS across layers; decrease; family differences	
Did magnitude or sentence-mean variance shrink?	raw SS_{lang}, SS_{sent}, SS_{res}, norm vs. frozen reference	shrinking a ratio hides
Is the cross-language map simple?	nested maps M1–M5, held-out error reduction	additive vs. rotational vs. nonlinear
Does English or a multilingual average predict a language better?	held-out R^2, English vs. average of others	English-centricity
Do the hardest sentences align?	AaR@10, worst-decile margin	tail failures that means hide

N COSINE MEASURES, AGREEMENT, AND LAYER CONDITIONS

Prespecified follow-up on previously examined outcomes; the analysis plan was frozen before any cosine value was read. For the 14 primary models with saved vectors, at every saved layer, for the 127 languages other than the English reference (including three regional English varieties) and sentences 0 to 299 of the grid, we computed the matched and language-mean cosines of Eq. 5 on the raw sentence vectors, and three measures on the jointly standardized grid: the *language-offset cosine*, the cosine between the centered language means $\boldsymbol{a}_\ell$ and $\boldsymbol{a}_{\mathrm{en}}$ of the full 128-language grid (because the standardized grid has zero grand mean, this is also the language-mean cosine on standardized vectors); the *pairwise LFS*, Eq. 3 recomputed on the two-language grid of English and language ℓ after re-standardizing that grid jointly (a reference-dependent extension, not the global LFS; on a two-language grid the two centered offsets are opposites, which is why offsets are taken on the full grid); and the *language contribution*, $N\|\boldsymbol{a}_\ell\|^2/\mathrm{SS}_{\mathrm{lang}}$, one language's share of the full-grid language sum of squares, which depends on the selected language set but does not use English as a reference. The saved 1,500-sentence representations of the 14 models are stored in fp16 after fp32 pooling, and all sums are computed in fp32; a parity check recomputed the whole-grid LFS from those dumps and matched the stored grid values within 1.1×10^{-4} over the 69 model–layer records. Table 11 records the preprocessing of every measure and Table 13 the values plotted in Figure 5.

Table 14 gives the within-model agreement between the measures at the layer of minimum language share; Table 15 gives per-model means at layer 0 and at the minimum layer. Across the 14 models the relations between the model-level aggregates depend on the layer condition: at the minimum layer global LFS correlates with mean matched cosine at $\rho = -0.57$ $[-0.86, -0.06]$ and with mean MEXA at -0.51 $[-0.92, +0.09]$, while for mean matched cosine and mean MEXA no clear association is detected ($+0.04$ $[-0.53, +0.70]$); at layer 0 the three coefficients are -0.14, -0.01, and $+0.52$, and at the deepest saved layer $+0.10$, -0.75, and -0.31 (one observation per model; 2,000-replicate model bootstrap). An earlier audit reported $+0.07$ for global LFS against mean MEXA across the same 14 models at the minimum layer; that value used the raw-vector MEXA values of an offset-removal study, not the standardized-grid MEXA used here, and the two are not interchangeable. The raw cosines at intermediate layers are dominated by a direction shared by all sentences: the earlier all-pairs cosine between unrelated sentences (13 languages) at the same layers is 0.67 to 0.998, close to the matched cosine, so the matched cosine there reflects anisotropy more than translation alignment.

Table 11: Measures compared in §6, their level, and their preprocessing.

Measure	Level	Input and preprocessing
Global LFS (Eq. 3)	one value per model and layer	jointly standardized full grid
Matched cosine (Eq. 5)	per language	raw sentence vectors
Language-mean cosine (Eq. 5)	per language	raw sentence vectors
Language-offset cosine	per language	language offsets on the standardized full grid
Matched cosine, 19-model check	per model	raw, globally centered, or jointly standardized 35-language grid
MEXA (Kargaran et al., 2025)	per language	stored implementation on the standardized grid
Pairwise LFS	per language	English plus one language, jointly re-standardized
Language contribution	per language	one language's share of the full-grid $\mathrm{SS}_{\mathrm{lang}}$
AaR@10	per language	worst-decile retrieval margin on the standardized grid
LFS-VC	one value per model and layer	noise-corrected REML variance components (Appendix C) on the jointly standardized grid (41 languages for the 19 context-transfer models)

Matched preprocessing on the context-transfer panel (exploratory). After the comparisons above, we extracted vectors for the 19 context-transfer models on one grid, English plus the 34 context languages that every model scored, with the first 300 sentences, at the layer selected from each model's stored LFS-VC profile (for mGPT, layer 12 with position-corrected vectors; its corrected minimum is layer 13, Appendix R). On the same vectors we computed the standard LFS and the matched cosine of Eq. 5 under three preprocessings: raw, after subtracting one grand-mean vector, and after the joint coordinate standardization used by LFS. The outcome is each model's mean normalized context benefit over the 34 languages, one observation per model; intervals resample the 19 models 5,000 times, and paired differences use the same draws. This analysis was specified after the outcomes had been seen and is not adjusted for covariates, so it is not a replacement for the preregistered result in Table 1, which uses LFS-VC on 646 covariate-adjusted model–language rows, nor for the separate model-level analysis of median normalized context benefit.

Neither measure has an established advantage: the paired difference includes zero at both layer rules. Against raw cosine alone the difference favors LFS (0.819 $[0.262, 1.259]$ after reversing the sign of

Table 12: Exploratory matched-preprocessing comparison on the 19 context-transfer models: Spearman correlation with mean normalized context benefit (35-language grid, 300 sentences, one observation per model, 5,000 model resamples). LFS keeps its own sign; the last rows compare strength as the standardized-cosine correlation minus the magnitude of the LFS correlation.

Measure	Layer selected from the LFS-VC profile	Half-depth layer
Standard LFS (language share)	-0.696 $[-0.920, -0.248]$	-0.563 $[-0.862, -0.073]$
Matched cosine, raw	-0.123 $[-0.564, +0.354]$	-0.070 $[-0.484, +0.402]$
Matched cosine, globally centered	$+0.319$ $[-0.129, +0.685]$	$+0.346$ $[-0.103, +0.722]$
Matched cosine, jointly standardized	$+0.809$ $[+0.500, +0.941]$	$+0.777$ $[+0.383, +0.957]$
Magnitude difference (standardized cosine)	$+0.112$ $[-0.130, +0.440]$	$+0.214$ $[-0.027, +0.542]$
Without mGPT: LFS; standardized cosine	-0.725; $+0.787$	-0.550; $+0.746$

LFS), but that contrast omits the preprocessing control. Within individual models, raw cosine has strong language-level associations with context benefit; the weak raw-cosine result concerns model averages. Model resampling does not fully represent dependence within model families or uncertainty in the saved context-benefit estimates, and these comparisons are not adjusted for multiplicity.

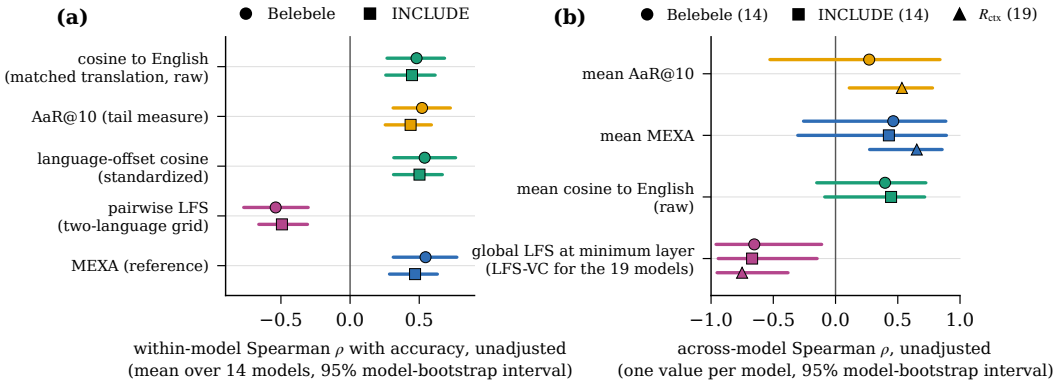

Figure 5: Table 13 plotted, with the language-offset cosine added in panel (a); unadjusted Spearman correlations with 2,000-replicate model-bootstrap intervals. (a) Within a model: mean over the 14 models of the per-model correlation between a per-language measure and per-language accuracy (circles Belebele, squares INCLUDE). Pairwise LFS keeps its own sign, so its negative values mean that a lower two-language share goes with higher accuracy. (b) Across models, one value per model: correlation with mean Belebele and INCLUDE accuracy (14 models) and with the median context benefit (19 models; LFS-VC for the LFS row).

O SHORT-CONTINUATION LANGUAGE PERSISTENCE (EXPLORATORY)

This study asks whether a base model, given an ordinary news sentence in a language, continues in that language without instruction. For each of 12 languages (English, German, French, Spanish, Russian, Arabic, Persian, Hindi, Chinese, Japanese, Indonesian, Swahili) we took 20 matched NTREX sentences, one from each of 20 documents not among the 300 sentences used for the representation grid, selected in file order with at least 40 letters in every language. The prompt is the sentence followed by a space, with no instruction, language label, or chat template; each of the 19 context-transfer models produced one greedy continuation of up to 64 new tokens (4,560 outputs; generation derives token positions from the attention mask, so left padding does not affect them). An unmodified langid 1.1.6 classifier over its 97 languages labels the generated suffix only: same language, off language, possibly mixed, uncertain (score below 0.90), too short (fewer than 20 letters), or empty. The primary outcome is the share of the 220 non-English outputs per model labeled off-language, averaged over the 11 languages; unresolved outputs are not counted as same-language. The predictor

Table 13: Comparison of the measures. Block A, within a model: each entry is the mean over the 14 models with saved vectors of the per-model Spearman correlation between a per-language measure and per-language accuracy at the layer of minimum language share, with the 2,000-replicate model-bootstrap interval of that mean; pairwise LFS keeps its own sign (negative: a lower two-language share goes with higher accuracy); the difference column compares strength, with the sign of pairwise LFS flipped before differencing; the adjusted column is the Spearman correlation of residuals after the resource adjustment on the raw accuracy scale. Block B, across models: one value per model (measures averaged over the languages with a score), Spearman with the mean accuracy of the 14 models with saved vectors and with the median context benefit of the 19-model panel; the 19-model entry of the LFS row uses LFS-VC. Average ranks for ties. Prespecified follow-up on previously examined outcomes.

A. Within a model (14 models with saved vectors, layer of minimum language share)

Measure	Belebele ρ [95% CI]	INCLUDE ρ [95% CI]	Belebele ρ, covariate-adjusted	Unadjusted difference from MEXA [95% CI]
Cosine to English (matched translation; raw)	$+0.481$ $[+0.267, +0.683]$	$+0.448$ $[+0.258, +0.615]$	-0.116	-0.065 $[-0.112, -0.025]$
MEXA (reference; standardized)	$+0.546$ $[+0.312, +0.772]$	$+0.470$ $[+0.285, +0.631]$	$+0.316$	–
AaR@10 (standardized)	$+0.520$ $[+0.311, +0.725]$	$+0.438$ $[+0.254, +0.588]$	$+0.166$	-0.025 $[-0.080, +0.018]$
Pairwise LFS (derived)	-0.538 $[-0.769, -0.303]$	-0.493 $[-0.660, -0.308]$	-0.164	-0.008 $[-0.018, +0.003]$

B. Across models (one value per model)

Measure	Mean Belebele accuracy, 14 models, ρ [CI]	Mean INCLUDE accuracy, 14 models, ρ [CI]	Median R_{ctx}, 19 models, ρ [CI]
Global LFS at the minimum layer	-0.653 $[-0.960, -0.112]$	-0.670 $[-0.943, -0.148]$	-0.751 $[-0.952, -0.382]$ (LFS-VC)
Mean cosine to English (raw)	$+0.398$ $[-0.151, +0.726]$	$+0.446$ $[-0.087, +0.716]$	separate comparison (Table 12)
Mean MEXA	$+0.464$ $[-0.257, +0.886]$	$+0.429$ $[-0.303, +0.891]$	$+0.653$ $[+0.273, +0.856]$
Mean AaR@10	$+0.270$ $[-0.527, +0.839]$	–	$+0.533$ $[+0.110, +0.779]$

is the standard LFS on the 41-language grid at the layer selected from the stored LFS-VC profile (for mGPT, layer 12 with position-corrected vectors).

The primary association is inconclusive ($+0.467$ $[-0.039, +0.791]$; Table 16); at the secondary half-depth layer it is $+0.626$ $[+0.206, +0.863]$. Classifier coverage, the share of outputs labeled either same- or off-language, varies with LFS, so unresolved output is related to the predictor and the detected rate is not a clean measure of switching: a model that produces empty, repetitive, or unclassifiable text can have a low detected rate without continuing well in the prompt language. An AI-assisted reading of 1,582 selected outputs, not a random sample and not verified by people, found genuine switches (English after Swahili prompts for Yi-1.5-9B), foreign segments inside otherwise same-language text, false off-language labels on repetitive text, prompt copying, and whitespace-only output; no label, threshold, or correlation was changed after that reading. We therefore report this as a detector-based exploratory analysis, not as evidence about how often models actually switch language.

Table 14: Within-model agreement between per-language measures at the layer of minimum language share: Spearman correlation over the 127 non-reference languages inside each of the 14 models, mean over models with the 2,000-replicate model-bootstrap interval of the mean. Signs are original (the pairwise share is a language share, so its sign against similarity measures is negative). Prespecified follow-up on previously examined data.

Measure	Measure	Mean within-model ρ [95% CI]
Matched cosine	Language-mean cosine	$+0.994$ $[+0.989, +0.997]$
Matched cosine	Language-offset cosine	$+0.829$ $[+0.788, +0.869]$
Matched cosine	MEXA	$+0.866$ $[+0.814, +0.914]$
Matched cosine	Pairwise LFS	-0.908 $[-0.947, -0.863]$
Pairwise LFS	MEXA	-0.963 $[-0.970, -0.956]$
Language-offset cosine	MEXA	$+0.909$ $[+0.895, +0.923]$
MEXA	AaR@10	$+0.829$ $[+0.771, +0.888]$

Table 15: Per-model values at layer 0 and at the layer of minimum language share: global LFS on the full standardized grid (standard estimator), and matched cosine to English (raw vectors) and MEXA (standardized grid), each averaged over the 127 non-reference languages. Sorted by LFS at the minimum layer.

Model	Min. layer	LFS, layer 0	LFS, min. layer	Cosine, layer 0	Cosine, min. layer	MEXA, layer 0	MEXA, min. layer
Qwen3-4B-Base	17	0.916	0.602	0.252	0.974	0.101	0.541
Qwen3-8B-Base	19	0.945	0.608	0.445	0.957	0.153	0.601
salamandra-2b	6	0.966	0.712	0.585	0.979	0.197	0.201
Qwen3-1.7B-Base	13	0.927	0.715	0.264	0.957	0.081	0.494
Qwen3-0.6B-Base	8	0.933	0.728	0.247	0.992	0.097	0.337
Mistral-7B-v0.3	2	0.956	0.744	0.380	0.999	0.179	0.090
salamandra-7b	18	0.963	0.784	0.252	0.910	0.246	0.437
Falcon3-7B-Base	13	0.967	0.826	0.128	0.781	0.173	0.256
OLMo-2-1124-7B	12	0.981	0.846	0.307	0.552	0.127	0.290
EuroLLM-1.7B	14	0.951	0.854	0.271	0.858	0.222	0.392
bloom-7b1	17	0.927	0.884	0.555	0.890	0.197	0.334
bloom-1b7	14	0.938	0.899	0.522	0.845	0.287	0.285
OLMo-2-0425-1B	9	0.980	0.903	0.357	0.599	0.124	0.161
SmolLM2-1.7B	15	0.972	0.904	0.060	0.959	0.071	0.207

P INCREMENTAL PREDICTION DETAILS

Design: one row per model (19 models, 13 families); predictors are model-mean MEXA (mean of the 34 rows for those 34 languages), LFS-VC at the layer of minimum language share from the expansion protocol (never relabeled as standard LFS; a linear fit is unchanged by using $1 -$ LFS-VC instead), and model-mean AaR@10; outcome is the median R_{ctx} over the same 34 languages. For each family in turn, an ordinary-least-squares fit on the other families predicts the excluded family, with predictor means and standard deviations taken from the training rows only. Families are unweighted. Intervals are 2,000-replicate model bootstraps of the out-of-fold predictions (models resampled with replacement; the fits are not repeated inside the bootstrap), so they condition on the fitted predictions; paired gains resample the same models for both predictor sets. A rank-based variant fits ranks within the training fold and places test ranks relative to it. Table 17 lists every predictor set; the AaR-only predictor has a negative held-out correlation, so the AaR rows are unstable. The gain is not established: the MAE reduction of MEXA with LFS-VC over MEXA alone is 0.018 $[-0.003, 0.040]$ and its held-out ρ gain $+0.246$ $[-0.091, +0.608]$; under the rank-based variant the held-out ρ rises from 0.563 to 0.767 (no interval is stored for this variant); and with AaR@10 in place of MEXA the MAE reduction is 0.015 $[-0.017, 0.049]$.

Descriptive partial correlations on the same 19 rows (rank residualization; model bootstrap): LFS-VC with median R_{ctx} given MEXA, -0.745 $[-0.950, -0.393]$; MEXA given LFS-VC, $+0.644$ $[+0.183, +0.855]$; the two predictors correlate at $\rho = -0.34$. A partial rank correlation is a conditional association in this sample, not evidence of out-of-sample usefulness. Cosine does not enter this comparison; the separate exploratory cosine check on the same 19 models is in Appendix N.

Table 16: Exploratory continuation study: Spearman correlations across 19 models with 5,000 model resamples. A positive LFS coefficient means a larger language share goes with more detected off-language output.

Predictor and layer	Outcome	ρ [95% CI]
LFS, selected layer (primary)	detected off-language share	$+0.467$ $[-0.039, +0.791]$
Raw matched cosine, selected layer	detected off-language share	-0.430 $[-0.779, +0.083]$
LFS, selected layer	classifier coverage	-0.597 $[-0.853, -0.177]$
LFS, half-depth layer (secondary)	detected off-language share	$+0.626$ $[+0.206, +0.863]$
Raw matched cosine, half-depth layer	detected off-language share	-0.592 $[-0.815, -0.206]$

Table 17: All predictor sets of the leave-one-family-out prediction (19 models). MAE in R_{ctx} units.

Predictors	Held-out ρ [95% CI]	Held-out r	MAE [95% CI]	Rank-variant held-out ρ
MEXA	$+0.553$ $[+0.119, +0.786]$	$+0.382$	0.0860 $[0.059, 0.117]$	$+0.563$
LFS-VC	$+0.735$ $[+0.350, +0.938]$	$+0.674$	0.0717 $[0.051, 0.095]$	$+0.713$
MEXA + LFS-VC	$+0.798$ $[+0.442, +0.975]$	$+0.725$	0.0676 $[0.048, 0.090]$	$+0.767$
AaR@10	-0.435 $[-0.790, +0.118]$	-0.183	0.1000 $[0.059, 0.145]$	$+0.409$
AaR@10 + LFS-VC	$+0.661$ $[+0.233, +0.915]$	$+0.541$	0.0852 $[0.057, 0.118]$	$+0.795$
Paired gains: MEXA + share over MEXA, ρ $+0.246$ $[-0.091, +0.608]$, MAE reduction 0.018 $[-0.003, 0.040]$; AaR + share over AaR, MAE reduction 0.015 $[-0.017, +0.049]$.				

Q Result ledger for the context-transfer association

Table 18 lists the reported versions of the association between LFS-VC and cross-language context use, with its unit and provenance. The July 2026 informal figure of 0.59 on a resource-restricted subset has no stored artifact and is not used.

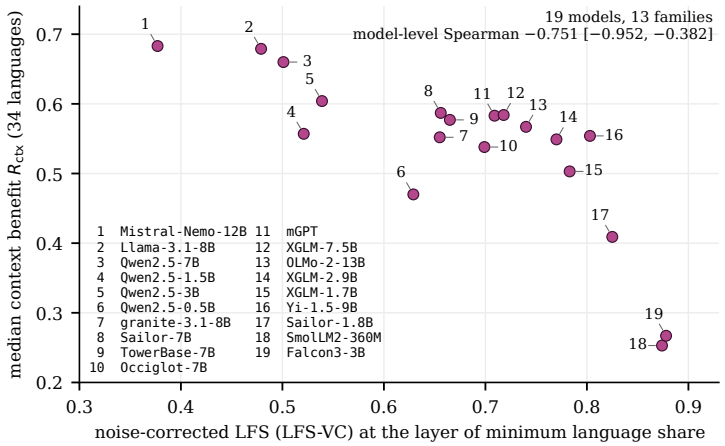

Figure 6: Model-level view of the context-transfer association: each point is one of the 19 frozen models, placed by LFS-VC at its layer of minimum language share (expansion protocol) and its median normalized context benefit over 34 languages. Numbers identify the models (key at lower left, numbered from left to right). The model-level Spearman correlation and its 2,000-replicate model-bootstrap interval come from the corrected rerun (Table 18); the adjusted row-level value of -0.46 (Table 1) is a separate analysis.

R Corrections ledger

Every discrepancy found during the project was recorded with its date, cause, and fix. In summary: a torn line in the run ledger (file locking added); a stale prototype result left in a results directory

(renamed so that globs fail loudly); a size-invariance claim stated too strongly (replaced by the reference value, a ratio of expected sums of squares, and its verifications); a permutation-null criterion written with an inverted inequality (corrected before scoring); function words contaminating a word-level concept grid (filtered at the concept definition); a collapse criterion that the control condition also failed (criterion rewritten); a criterion scored on sign alone (replaced by an effect-size criterion); a preregistration that mis-described its own held-out set (left as written, with the correction attached); an inert variance-preservation term in the word-level trainer (fixed and rerun); a frozen reference skipped by name in an evaluation script (identified by set membership instead); profile counts that combined the two estimators (reported separately since); an English-versus-average result stated as universal when it is model-specific (reported as a range); and, during the final revision, a regression statistic cited from the development-grid run rather than the 17-model artifact (corrected in Appendix G), and an offset-plus-scale range quoted from a three-model prototype at selected ranks (65 to 98 percent) where the 14-model common-rank artifacts give 63 to 97 percent (corrected throughout); the English-versus-average cross-validation described as two-fold where the producing script, which wrote the saved results, uses three folds (description corrected, no rerun); and global-LFS coefficients reported from the adjusted benchmark analysis, whose outcome has model fixed effects removed, which absorbs all between-model variation of a predictor with one value per model (those coefficients removed; the applicable per-language analyses retained; global-LFS associations reported separately at the model level). With that change the pooled-panel support in Appendix G is the MEXA and AaR@10 support of 1,726 rows over 69 languages rather than the 1,740 rows over 70 languages of the removed LFS-VC row; both come from the same saved output. Finally, a position-handling error affected mGPT, one of the 19 context-transfer models: its tokenizer pads batches on the left and our extractor passed no position indices, so the absolute positions of mGPT's real tokens, and therefore its representations, depended on the other sentences in a batch. Positions are now derived from the attention mask; the corrected extraction agrees across batch sizes 8 and 12 and with unpadded inputs, and reproduces the separately validated layer-12 values. The corrected full profile moves mGPT's minimum from layer 12 (LFS-VC 0.444) to layer 13 (0.709), still an intermediate layer. Every affected analysis of the 19 context-transfer models was rerun with unchanged code and only mGPT's inputs replaced, after a rerun on the original inputs reproduced every printed value (the LFS-VC coefficient to within 0.0002); all current results in this paper use the corrected inputs, and historical values appear only in this appendix and in the explicitly labeled block of Table 18. The exploratory cosine and continuation analyses retain the historically selected mGPT layer 12, using position-corrected representations; the corrected full-profile minimum is layer 13. The main changes are the adjusted association (-0.51 to -0.46), the model-level association (-0.770 to -0.751), and the incremental-prediction gain, whose interval now includes zero; the MEXA reference values also moved (0.591 to 0.697 adjusted, 0.558 to 0.653 at the model level). The other 18 models use rotary position encodings or pad on the right, so their representations are not affected by this position-handling error; the context-transfer scoring pads on the right, and the continuation study derives positions from the attention mask.

S Six focal languages

Each analysis uses its full prespecified language support (128 on the grid, 116 on FLORES-200, 41 on the context-transfer test with 34 scored for every model, 24 in the language probe; benchmark language coverage varies by analysis, see Table 3); no language was selected or removed after outcomes. For readers who want a fixed set to follow through the paper, Tables 19 and 20 report six high- and mid-resource languages spanning three scripts (Latin, Cyrillic, and Arabic): German, French, Spanish, Russian, Arabic, and Persian.

T Overview of the computation

Figure 7 summarizes §3 in one diagram, from the input sentences to LFS at every layer.

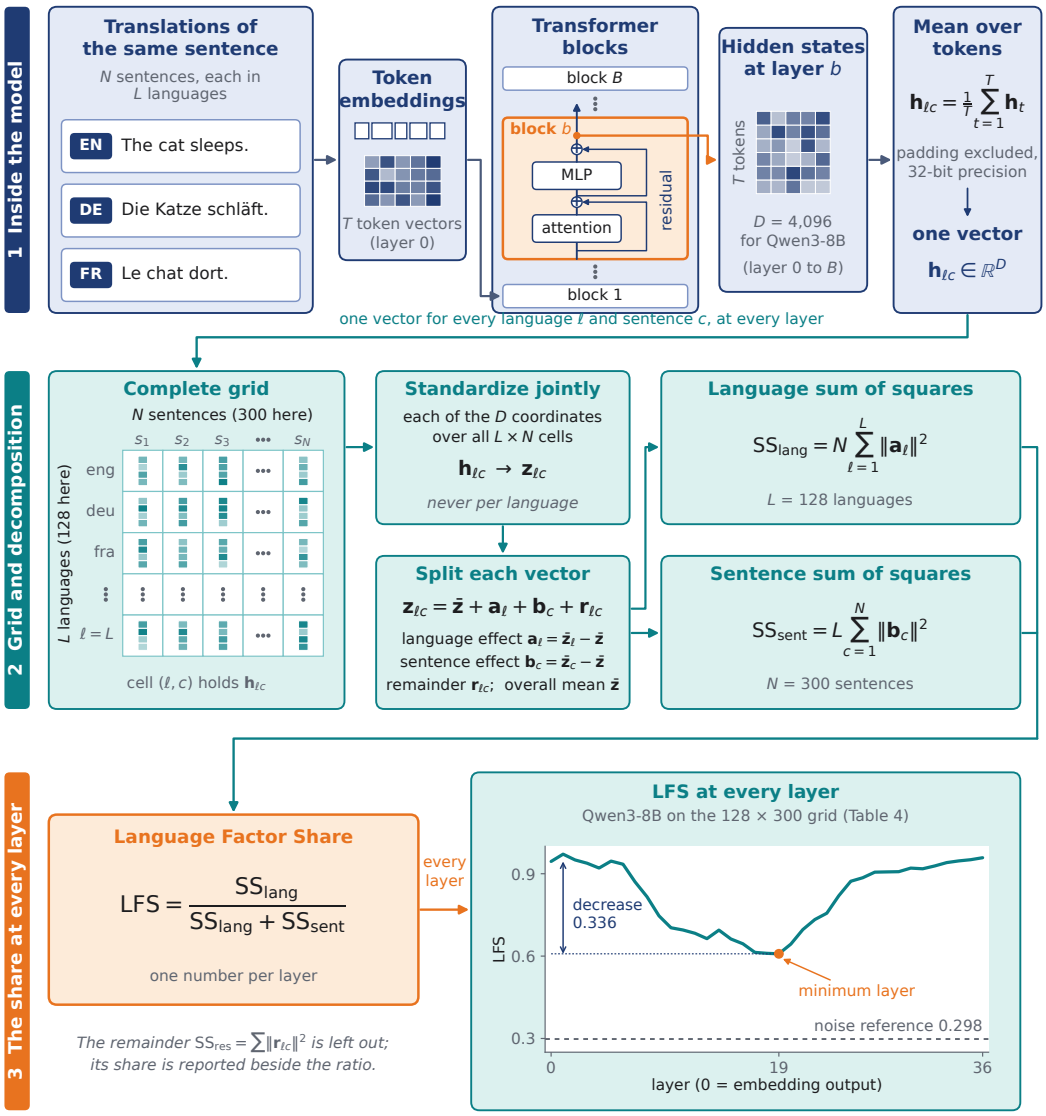

Figure 7: How LFS is computed. (1) Each of the $L \times N$ translated sentences is run through the model; at the chosen layer (0 is the embedding output, 1 to B the block outputs) its token vectors are averaged, excluding padding, into one vector $\boldsymbol{h}_{\ell c}$ of dimension D. (2) The vectors form a complete grid, are standardized jointly (never per language), and each is split into the overall mean, a language effect $\boldsymbol{a}_\ell$, a sentence effect $\boldsymbol{b}_c$, and a remainder $\boldsymbol{r}_{\ell c}$ (Eq. 1), whose sizes are measured by sums of squares (Eq. 2). (3) LFS is the language sum of squares as a share of the language and sentence sums (Eq. 3), computed at every layer; the profile shown is Qwen3-8B on the 128×300 grid (Table 4), and the dashed line is the noise reference (Eq. 4).

Table 18: The reported versions of the association between LFS-VC and cross-language context use, with its unit, support, and provenance, in three groups: an earlier four-model pilot (Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, OLMo-2-0425-1B), which did not include mGPT and is unaffected by the correction; historical 19-model values that used mGPT representations affected by the position error described in Appendix R, now superseded; and the corrected results.

Analysis	Predictor	Unit and support	Estimate	Provenance
Earlier four-model pilot; unaffected by the mGPT correction				
Four-model pilot	LFS-VC	model–language rows, 4 models	$\rho = -0.84$, model-cluster CI $[-0.91, 0.04]$	original protocol
Historical 19-model inputs (mGPT position error; superseded)				
Adjusted expansion result	LFS-VC	646 model–language rows, 19 models	$\rho = -0.51$ $[-0.71, -0.19]$ model, $[-0.72, -0.18]$ crossed; leave-one-family-out -0.60 to -0.44, family-bootstrap $[-0.76, -0.24]$	frozen protocol, 2026-08-24 run
Adjusted expansion result	MEXA	same rows	$\rho = 0.591$ $[0.33, 0.82]$	same run
Model level, original run	LFS-VC	19 model summaries, median over 34 languages	$\rho = -0.770$ $[-0.94, -0.44]$	2026-08-24 run
Model level, recomputed	LFS-VC	same 19 summaries (identical bytes)	$\rho = -0.770$ $[-0.942, -0.450]$; $r = -0.768$ $[-0.887, -0.586]$; leave-one-family-out -0.832 to -0.730 (ρ), -0.814 to -0.735 (r)	follow-up, 2026-09-23
Model level, recomputed	model-mean MEXA	same 19 summaries	$\rho = 0.558$ $[0.095, 0.832]$	follow-up, 2026-09-23
Estimator check	standard LFS	646 rows; 19 summaries	adjusted $\rho = -0.511$; model level -0.782	follow-up, 2026-09-23
Specification curve	LFS-VC	646 rows; 80 specifications	ρ from -0.821 to -0.378; 80 of 80 intervals exclude zero	follow-up, 2026-09-23
Corrected mGPT inputs (rerun with unchanged code, 2026-09-24)				
Adjusted expansion result	LFS-VC	646 model–language rows, 19 models	$\rho = -0.46$ $[-0.69, -0.11]$ model, $[-0.71, -0.11]$ crossed; leave-one-family-out -0.55 to -0.38, family-bootstrap $[-0.74, -0.17]$	corrected rerun
Adjusted expansion result	MEXA	same rows	$\rho = 0.697$ $[0.50, 0.85]$	corrected rerun
Model level	LFS-VC	19 model summaries, median over 34 languages	$\rho = -0.751$ $[-0.952, -0.382]$; $r = -0.777$ $[-0.892, -0.587]$; leave-one-family-out -0.811 to -0.707 (ρ), -0.815 to -0.747 (r)	corrected rerun
Model level	model-mean MEXA	same 19 summaries	$\rho = 0.653$ $[0.273, 0.856]$	corrected rerun
Estimator check	standard LFS	646 rows; 19 summaries	adjusted $\rho = -0.461$; model level -0.763	corrected rerun
Specification curve	LFS-VC	646 rows; 80 specifications	ρ from -0.885 to -0.359; 80 of 80 intervals exclude zero	corrected rerun
Matched preprocessing (exploratory)	standard LFS	19 models, 35-language grid, mean benefit	$\rho = -0.696$ $[-0.920, -0.248]$; 5,000 model resamples	follow-up, 2026-09-24

Table 19: Context transfer R_{ctx} (the reduction in prediction loss from relevant rather than unrelated context, normalized by the English reduction) for the six focal languages across the 19 frozen models of the expansion set.

Model	German	French	Spanish	Russian	Arabic	Persian
Falcon3-3B-Base	0.60	0.77	0.78	0.30	0.04	0.13
Llama-3.1-8B	0.79	0.76	0.74	0.69	0.62	0.66
Mistral-Nemo-Base-2407	0.77	0.79	0.78	0.72	0.68	0.65
OLMo-2-1124-13B	0.75	0.76	0.79	0.62	0.48	0.56
Qwen2.5-0.5B	0.59	0.68	0.69	0.54	0.45	0.32
Qwen2.5-1.5B	0.69	0.71	0.70	0.61	0.53	0.46
Qwen2.5-3B	0.69	0.73	0.73	0.65	0.60	0.53
Qwen2.5-7B	0.77	0.75	0.73	0.69	0.64	0.63
Sailor-1.8B	0.60	0.67	0.66	0.47	0.28	0.22
Sailor-7B	0.74	0.73	0.75	0.65	0.54	0.50
SmolLM2-360M	0.51	0.60	0.59	0.22	0.05	0.11
TowerBase-7B	0.73	0.78	0.77	0.68	0.43	0.44
Yi-1.5-9B	0.71	0.75	0.73	0.59	0.46	0.46
granite-3.1-8b-base	0.74	0.70	0.72	0.55	0.51	0.44
mGPT	0.79	0.77	0.77	0.67	0.52	0.56
occiglot-7b-eu5	0.82	0.78	0.78	0.61	0.36	0.34
xglm-1.7B	0.69	0.69	0.72	0.59	0.48	0.23
xglm-2.9B	0.72	0.72	0.73	0.62	0.55	0.25
xglm-7.5B	0.73	0.75	0.76	0.66	0.58	0.27

Table 20: Share of the language sum of squares $\mathrm{SS}_{\mathrm{lang}}$ carried by each focal language at the layer of minimum language share (1,500-sentence dumps, 14 models). A uniform share would be $1/128 = 0.0078$; a language above it contributes more than its share of the language component.

Model	German	French	Spanish	Russian	Arabic	Persian	Uniform share
EuroLLM-1.7B	0.0041	0.0040	0.0041	0.0042	0.0046	0.0094	0.0078
Falcon3-7B-Base	0.0065	0.0079	0.0082	0.0068	0.0062	0.0068	0.0078
Mistral-7B-v0.3	0.0058	0.0048	0.0050	0.0048	0.0096	0.0104	0.0078
OLMo-2-0425-1B	0.0065	0.0060	0.0053	0.0054	0.0057	0.0052	0.0078
OLMo-2-1124-7B	0.0058	0.0057	0.0052	0.0049	0.0042	0.0044	0.0078
Qwen3-0.6B-Base	0.0060	0.0059	0.0059	0.0058	0.0069	0.0051	0.0078
Qwen3-1.7B-Base	0.0062	0.0056	0.0055	0.0054	0.0068	0.0056	0.0078
Qwen3-4B-Base	0.0058	0.0055	0.0055	0.0052	0.0063	0.0049	0.0078
Qwen3-8B-Base	0.0052	0.0049	0.0051	0.0049	0.0061	0.0048	0.0078
SmolLM2-1.7B	0.0064	0.0055	0.0052	0.0061	0.0065	0.0062	0.0078
bloom-1b7	0.0060	0.0048	0.0047	0.0073	0.0055	0.0054	0.0078
bloom-7b1	0.0053	0.0042	0.0042	0.0061	0.0045	0.0057	0.0078
salamandra-2b	0.0060	0.0039	0.0059	0.0047	0.0121	0.0104	0.0078
salamandra-7b	0.0044	0.0043	0.0045	0.0045	0.0086	0.0110	0.0078