Cross-Lingual Alignment Without Joint Training: Do Monolingual Language Models Converge on Universal Representations?

Anonymous ACL submission

Abstract

Cross-lingual alignment in multilingual language models is typically attributed to joint training: shared parameters, mixed-language batches, or explicit alignment objectives. We ask whether monolingual models trained on non-parallel data learn alignable representations without joint training. By testing on strictly monolingual language models, such as the Goldfish model families and independently developed models from different research labs, we find three results. *Correlation*: these models develop alignable representational geometry across layers, with alignment strengthening as data scale, model scale, or linguistic proximity increases. *Construction*: a single Procrustes rotation fit on parallel sentences maps hidden states between models. *Causation*: the same rotation transfers functional content; patching a rotated English residual into a German model on a factual cloze flips the prediction to the donor's capital in most cases. We confirm that cross-lingual alignment can emerge from the structure of language and the information it carries rather than from joint training, and this points to practical future directions including model stitching, merging, and modular multilingual systems built from monolingual components.

1 Introduction

Multilingual language models produce representations in which cross-lingual semantically parallel inputs occupy close regions of activation space (Chang et al., 2022; Brinkmann et al., 2025). The research community has proposed various explanations in mechanism: shared subword vocabularies (Pires et al., 2019), similar input distributions (Conneau et al., 2020b), or shared architectural parameters (Dufter and Schütze, 2020). These studies only investigate the cause within the features of joint training, and no existing account predicts cross-lingual alignment between models trained independently on disjoint corpora.

In this paper, we ask whether joint training is required. We pose this question in its strictest form: if two models are trained independently, on completely distinct corpora and in different languages, with no shared parameters or alignment signal, do their representations still end up alignable? A positive answer would attest the emergence of language-agnostic conceptual representations purely from the corpora. This is predicted by the Platonic Representation Hypothesis (Huh et al., 2024), which has been tested cross-modally but not cross-linguistically. Our setting is strictly monolingual language models: each is a Transformer trained on a single language. For controlled comparison we use the Goldfish family (Chang et al., 2026). Our central findings replicate on five independently developed $\sim$1B-parameter monolingual models from different research groups, showing the effect holds at larger scale.

Our analysis follows a ***C**orrelation* $\rightarrow$ ***C**onstruction* $\rightarrow$ ***C**ausation* pipeline, each stage placing a stronger demand on the alignment. The correlational stage (Section 3) finds that independently trained monolingual models develop alignable representational geometry on parallel sentences, with alignment strengthening as data scale, model scale, or linguistic proximity increases, and persisting across models from different research groups. The constructional stage (Section 4) compares orthogonal rotation against more flexible mappings (affine, MLP) and finds that a single rotation suffices to align one model's representations with another's; the orthogonality constraint preserves the angular geometry that retrieval depends on, whereas the relaxed alternatives lower reconstruction error but weaken retrieval. The causal stage (Section 5) shows that the same rotation transfers functional content: patching a rotated English residual into a German model on a factual cloze task flips the prediction to the donor's capital in most cases,

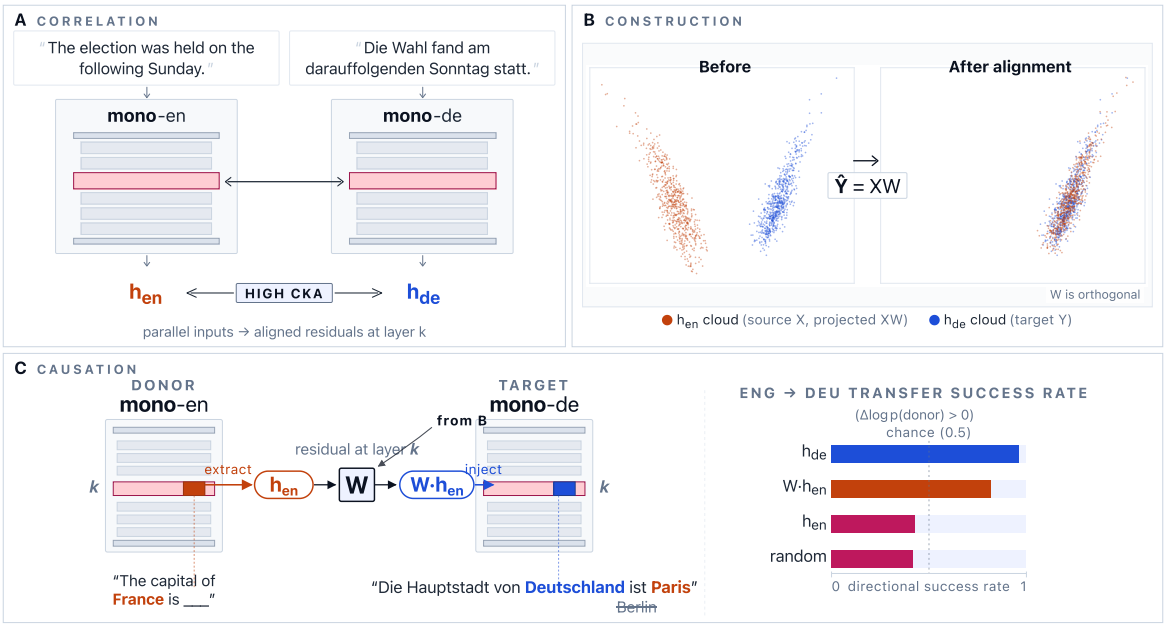

Figure 1: **Independent monolingual LMs share representational geometry; a rotation makes that geometry causally usable. (A)** Parallel English/German sentences fed to monolingual models yield residual streams h_{en}, h_{de} with high CKA. **(B)** A Procrustes rotation W fit on parallel sentence representations aligns the two clouds: $\widehat{Y}=XW$ overlaps the target. **(C)** The same W transfers factual content across models. Patching $W \cdot h_{\text{en}}$ at the residual of a MONO-DE prompt flips the German completion, recovering directional success rate ($\Delta \log p(\text{donor})>0$) most of the ceiling and well above baselines.

with the pattern holding across five unrelated target languages and approaching the within-model ceiling.

2 Related Work

2.1 Cross-Lingual Representations in Multilingual LMs

A growing literature finds that multilingual language models develop partly shared representational structure across languages. Early evidence in jointly trained encoders established that languages occupy overlapping linear subspaces (Chang et al., 2022). Subsequent work on decoder-only multilingual LMs has found a similarly partial language-agnostic geometry, with internal components shared across languages while still admitting language-specific divergences (Zhang et al., 2025; Wendler et al., 2024; Brinkmann et al., 2025; Dumas et al., 2025; Riemenschneider and Frank, 2025; Körner et al., 2026). This coneects to the Platonic Representation Hypothesis (Huh et al., 2024), which conjectures that representations converge toward a shared model of reality as models scale. Jha et al. (2026) support this at the embedding level, showing text embeddings can be translated across architectures without paired data.

2.2 Monolingual LMs

Most state-of-the-art large-scale LMs (Qwen et al., 2025; Team et al., 2024) and well-studied multilingual LMs (Devlin et al., 2019; Conneau et al., 2020a; Le Scao et al., 2022) are trained jointly across many languages. For our experiments, we require strictly monolingual models, in order to study whether language-agnostic structure can emerge independently of multilingual joint training. Recent efforts have produced such models, namely the Goldfish models (Chang et al., 2026), which include comparably trained monolingual models for 350 languages. These provide a controlled setting for our experiments. There are also several monolingual models developed by language communities: Orlando et al. (2024); Liu et al. (2025); Corrêa et al. (2025); Ociepa et al. (2025). We use this to test the scalability of our findings, albeit in a less controlled setting.

2.3 Aligning Independently-Trained LMs

Several studies have found that independently trained representations can be aligned *post hoc*. Early work in Bilingual Lexicon Induction (BLI) demonstrates that monolingual word embedding spaces are often related through approximately

orthogonal transformations, enabling bilingual lexicon induction without parallel supervision (Mikolov et al., 2013; Xing et al., 2015a; Smith et al., 2017; Artetxe et al., 2018; Conneau et al., 2018). Orthogonality constraints usually improve stability and performance relative to unconstrained linear mappings, suggesting that independently trained semantic spaces preserve similar global geometry. Geometric alignment measures, such as Centered Kernel Alignment (Kornblith et al., 2019), which builds on the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), have been used to probe multilingual encoders (Muller et al., 2021). However, correlational similarity is insufficient on its own for making claims about shared representation, as CKA can be dominated by a small number of high-variance components (Davari et al., 2022), inflated by model width and depth with only local neighbourhood agreement surviving null calibration (Gröger et al., 2026). Additionally, nearest-neighbour alignment (as was used in Huh et al., 2024) can be fragile to evaluation scale (Koepke et al., 2026). Mechanistic interpretability approaches provide a stronger test of representational convergence by testing functional interchangeability: whether aligned states are usable by the receiving model's computation. In this work, we use cross-model activation patching (Dumas et al., 2025; Körner et al., 2026) as a causal test. If a representation from one model can be projected into another model's residual stream and successfully drive downstream behaviour on a new task, then the alignment must preserve functionally relevant structure rather than correlational similarity alone.

3 Representational Alignment

We first ask whether independently trained monolingual models converge on similar internal representations of parallel semantic content. We address this correlationally by measuring CKA between model pairs under matched and shuffled controls.

3.1 Experimental Setup

Models. We start by using strictly monolingual Goldfish causal language models (Chang et al., 2026).[1] All models share an identical Transformer architecture and training data size, but have disjoint vocabularies, parameters, and training corpora. The family uses two architectures matched to data scale: the 5 MB and 10 MB models have 39M parameters, while the 100 MB and 1000 MB models have 125M. Unless otherwise stated, we use the 1000 MB tier.

Parallel data. We evaluate on four parallel-sentence datasets: FLORES-200 (Team et al., 2022), Tatoeba (Tiedemann, 2020), OPUS (Tiedemann, 2012), and BouQUET (Andrews et al., 2025). For each dataset and language pair, we extract hidden states from every Transformer layer, including the embedding layer.

Similarity metric. We measure cross-model similarity with linear CKA (Kornblith et al., 2019):

$$\mathrm{CKA}(X, Y) = \frac{\mathrm{HSIC}(X, Y)}{\sqrt{\mathrm{HSIC}(X, X)\,\mathrm{HSIC}(Y, Y)}},$$

which is invariant to orthogonal transformations and isotropic rescaling. To control for spurious similarity, we contrast a *matched* condition, which consists ofaligned parallel sentences, with a *shuffled* control. In this case, we randomly sample a sentence from the same dataset. We compare across all 36 language pairs.

Three representations. We compute CKA under three pooling strategies. Our sentence-level representation includes (i) *Mean pooling*: average token hidden states across the sequence, and (ii) *SGPT position-weighted pooling* (Muennighoff, 2022): weight each token by its position in the sequence before averaging, emphasising information accumulated toward the end of the context, which proves to work better for causal language models; and (iii) *Token-level word-aligned pooling*: isolate lexical correspondence by averaging subword hidden states within each aligned word span, using SimAlign (Jalili Sabet et al., 2020) word alignments over the parallel sentence pair. For a word w with subword token set $T(w)$ and per-token hidden states $\mathbf{h}_{t,\ell}$ at layer ℓ, $\mathbf{z}_{w,\ell} = \frac{1}{|T(w)|}\sum_{t \in T(w)} \mathbf{h}_{t,\ell}$.

Per-pair CKA values for all three settings on all four datasets are in Appendix A.

3.2 Cross-Lingual Alignment Across 36 Pairs

We show the gap between the matched and shuffled conditions in Figure 2. We show the per-layer trajectory for a single representative pair (English–French) under all three pooling-strategies and the shuffled condition. We also report Δ between the two, where a larger delta indicates more improvement over the baseline. Table 1 averages the gap across all pairs in four datasets. Figure 3 unpacks the averages to a 9×9 matrix.

[1] https://huggingface.co/goldfish-models

Dataset	Mean Pooling			Token-Aligned			SGPT		
	Matched	Shuffled	Δ	Matched	Shuffled	Δ	Matched	Shuffled	Δ
FLORES	0.72	0.15	+0.57	0.38	0.03	+0.35	**0.78**	0.17	+0.61
Tatoeba	0.69	0.16	+0.53	0.36	0.03	+0.33	**0.75**	0.18	+0.57
OPUS	0.68	0.15	+0.53	0.35	0.03	+0.32	**0.73**	0.17	+0.56
BouQUET	0.70	0.16	+0.54	0.37	0.03	+0.34	**0.76**	0.18	+0.58
Avg.	*0.70*	*0.15*	*+0.54*	*0.36*	*0.03*	*+0.34*	*0.76*	*0.18*	*+0.58*

Table 1: **Mean last-layer CKA across all 36 Goldfish (1000 MB) language pairs**, by representation and dataset. Δ is the matched−shuffled gap. Full per-dataset 9 × 9 matrices are in Appendix Tables 8–19.

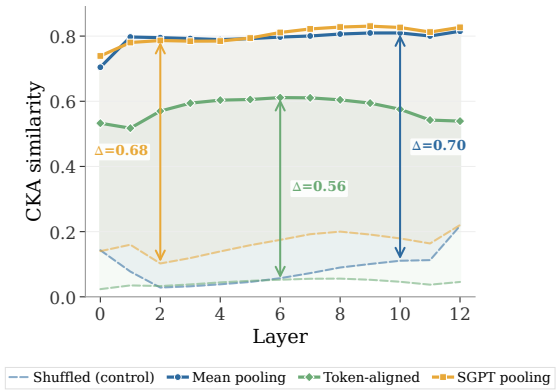

Figure 2: **Layerwise English–French CKA** (Goldfish 1000 MB) under three representations (mean pooling, token-aligned, SGPT). Solid: matched; dashed: shuffled.

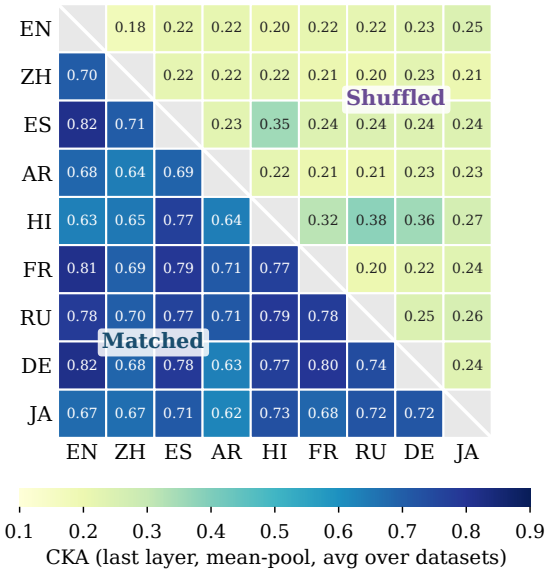

Figure 3: **Last-layer pairwise CKA (Goldfish 1000 MB, FLORES-200).** Lower triangle: matched; upper triangle: shuffled.

Semantic alignment exceeds architectural noise. Matched CKA is higher than the shuffled baseline at every layer and for every language pair (Figures 2 and 3). Under SGPT pooling, last-layer matched CKA averages 0.78 on FLORES, compared to 0.17 shuffled; the consistent gap is preserved across datasets (Table 1). Since the shuffled control uses the same sentences and only randomises the pairing, this gap cannot be explained by shared Transformer architecture or surface token statistics. Therefore, these results indicate stronger alignment than would be expected by chance.

Alignment is consistent across depth. Matched CKA is stable across layers under all three representations (Figure 2): in every case the layer-to-layer variation is small relative to the matched–shuffled gap.

Linguistic similarity modulates alignment. Figure 3 suggests a typological gradient: linguistically related pairs such as English–French (0.81), English–German (0.82) exhibit stronger alignment than distant pairs such as Chinese–Arabic (0.64) or Chinese–Hindi (0.65). We further test this by correlating last-layer matched CKA against URIEL typological distances (Littell et al., 2017) and a measure of model performance across the 36 Goldfish pairs. Following Chang et al. (2026), we use mean sentence negative log-likelihood (NLL) over parallel sentences to measure model quality. URIEL syntactic distance is the strongest single predictor (r=−0.60, ρ=−0.64, p<.001). Mean NLL is also a significant predictor of CKA alignment (r=−0.44, ρ=−0.43, p=.01). Full correlations and visualisations are in Appendix D.

3.3 Alignment Scales with Training Data

We use all four training-data scales available for Goldfish (5, 10, 100, 1000 MB) for each language, enabling a controlled study of how alignment grows with training data. Figure 4 plots mean off-diagonal matched CKA against training size

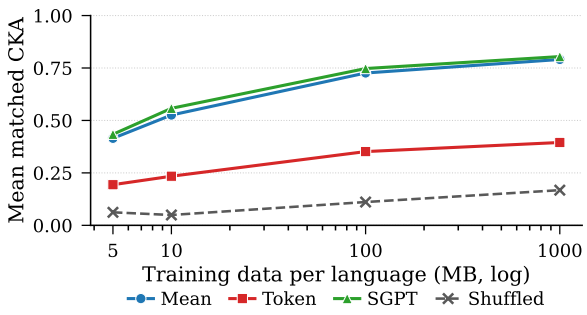

Figure 4: **Alignment scales with training data.** Mean matched last-layer CKA grows monotonically from 5 MB to 1000 MB under all three representations. Full 9×9 heatmaps at every scale are provided in Appendix Figures 14–16.

	Pyth. EN	Pyth. ZH	Tuc. PT	Biel. PL	Min. IT
Pyth. EN	—	.16	.19	.13	.22
Pyth. ZH	.70	—	.18	.12	.21
Tuc. PT	.84	.69	—	.14	.25
Biel. PL	.67	.61	.66	—	.17
Min. IT	.82	.70	.80	.63	—

Table 2: **Last-layer CKA across five ∼1B monolingual models** (SGPT). Lower triangle (blue): matched. Upper triangle (red): shuffled. Detailed model specifications are in Appendix Table 5.

for each of the three representations, with shuffled baselines for reference. Across all three representations, matched CKA increases monotonically with training-data scale.

A cross-size comparison pairs Goldfish models trained on different amounts of data (1000 MB against 100 MB, and 10 MB against 5 MB) within each model size. We see the diagonal (same language) dominates every row, and off-diagonal matched CKA still sits above the shuffled baseline at both pairings. This suggests that as model quality improves, cross-lingual representational alignment gets stronger. This is consistent with the Platonic Representation Hypothesis (Huh et al., 2024), which predicts that representational convergence increases with model capability. Further details are provided in Appendix B.2.

3.4 Alignment Persists Between Larger Models

A natural concern is that alignment within Goldfish models is only an artifact from shared design choices across the family. We therefore test whether alignment survives when models differ in architecture, vocabulary, training data, and research group. As a first check, comparing Goldfish English against Pythia-160M (Biderman et al., 2023) and Goldfish Chinese against Zh-Pythia-160M (Liu et al., 2025) also yields matched SGPT CKA above the shuffled baseline at every layer, albeit in a U-shaped profile (more discussion in Appendix C.1). We then compare five independently developed ∼1B monolingual models: Pythia-1.4B EN, Zh-Pythia-1.4B, Tucano-1B PT (Corrêa et al., 2025), Bielik-1.5B PL (Ociepa et al., 2025), Minerva-1B IT (Orlando et al., 2024), full specifications in Appendix Table 5. We conduct the same experiment as above. Table 2 reports last-layer CKA (SGPT) for all $\binom{5}{2} = 10$ model pairs on FLORES-200. Matched CKA exceeds the shuffled baseline in all 10 pairs (0.71 vs. 0.18 on average), confirming that the alignment pattern extends to models sharing nothing but the objective of modelling text.

CKA establishes similarity up to an unspecified transformation: it shows the two spaces are alignable but does not identify the map that aligns them. In the following section, we will establish causal evidence for meaningful alignment across models.

4 Cross-Lingual Representation Reconstruction

Section 3 established that independently trained monolingual models produce representations whose pairwise similarity exceeds shuffled baselines. We now move from correlation to *reconstruction*: given hidden states from a source-language model, can we recover the corresponding target-language representations using a simple learned mapping, and what hypothesis class is required?

4.1 Setup

We study all pairs of the Goldfish 1000 MB monolingual models from Section 3, and fit projections in both directions, yielding 72 directional pairs in total. As SGPT resulted in the strongest alignment, we use this method at every layer including the embedding layer. We use the same four parallel datasets as Section 3, yielding ∼5,000 parallel sentences per language pair (80/20 train/test split, seed 42).[2] On top of the resulting per-layer matrices $X \in \mathbb{R}^{n \times d}$ (source) and $Y \in \mathbb{R}^{n \times d}$ (target), we fit a hierarchy of three projection methods that each

[2] ≤50-character sentences filtered out.

relax a structural constraint. All three minimise squared reconstruction error.

(i) Procrustes. The most constrained mapping: a single rotation/reflection $W \in \mathbb{R}^{d\times d}$ that preserves all pairwise inner products in the source space (Schönemann, 1966).

$$W^\star = \underset{W^\top W = I}{\arg\min} \; \|XW - Y\|_F^2, \qquad (1)$$

solved in closed form by $W^\star = UV^\top$, where $U\Sigma V^\top = X^\top Y$ is the SVD.

(ii) Affine. We relax orthogonality: W is unconstrained and a bias $b \in \mathbb{R}^d$ is allowed.[3]

$$(W^\star, b^\star) = \underset{W,b}{\arg\min} \; \|XW + \mathbf{1}b^\top - Y\|_F^2. \quad (2)$$

(iii) MLP. We further relax linearity using a one-hidden-layer ReLU network f_θ of width d.[4]

$$\theta^\star = \underset{\theta}{\arg\min} \; \|f_\theta(X) - Y\|_F^2, \qquad (3)$$

$$f_\theta(x) = W_2\,\mathrm{ReLU}(W_1 x + b_1) + b_2. \qquad (4)$$

We evaluate on three complementary metrics with distinct geometric meaning: **P@1** (cosine retrieval, capturing local neighbourhood preservation), **MSE** (mean squared error; pointwise reconstruction quality), and post-projection **linear CKA** (relational structure). P@1 is reported in two variants: P@1_{std} retrieves against the test pool, while P@1_{hard} retrieves against the combined train+test pool.

4.2 Rotation Is Nearly Sufficient

Table 3 summarises reconstruction across all 72 directions. Three observations follow. **(i) All three projections are highly effective.** The identity baseline retrieves nothing (P@1_{std}=0.2%); a single Procrustes rotation leads to 88.7% retrieval accuracy, with MSE down from 2.258 to 0.689. A learned mapping recovers the matching target representation in the vast majority of cases. **(ii) Affine and MLP win MSE and CKA, as expected.** Both methods minimise reconstruction error directly and operate in a strictly larger hypothesis class than Procrustes, so unsurprisingly

[3]Solved in closed form via centring with ridge regularisation $\lambda\|W\|_F^2$; λ is tuned on a held-out validation slice over $\{10^{-2}, \ldots, 10^{4}\}$.

[4]Trained with Adam (lr=10^{-4}), weight decay 10^{-2}, batch size 64, up to 200 epochs with early stopping (patience 20) on a 10% internal validation split.

Method	P@1_{std}	P@1_{hard}	MSE	CKA
Procrustes	**.887**$_{\pm.021}$	**.775**$_{\pm.028}$	.689$_{\pm.026}$	.713$_{\pm.023}$
Affine	.814$_{\pm.027}$	.663$_{\pm.036}$	**.439**$_{\pm.017}$	**.762**$_{\pm.020}$
MLP	.809$_{\pm.029}$	.643$_{\pm.038}$	.456$_{\pm.017}$	.760$_{\pm.020}$
Identity	.002	.000	2.258	.713

Table 3: **Method comparison across 36 Goldfish language pairs** (final layer). Each cell shows the mean across 36 pairs with the bootstrapped 95% CI in brackets (10,000 resamples); **bold** marks the best method per metric. P@1_{std} retrieves against the test-only pool; P@1_{hard} against the combined train+test pool. MSE and CKA are pool-independent. Per-pair scatter is in Appendix Figures 12 and 17.

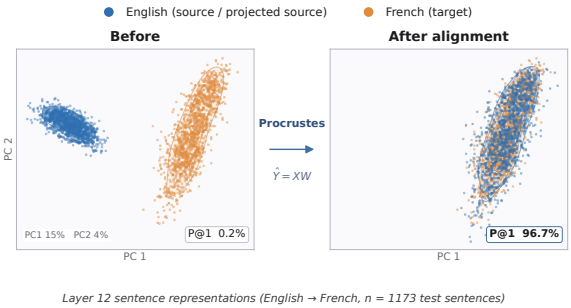

Figure 5: **Procrustes mapping, eng–fra, layer 12** (2D PCA over the joint test set). English in blue (source, Procrustes-mapped); French in yellow (target). The per-layer full version is in Figure 18.

they attain lower MSE on all 36 pairs (Affine 0.439, MLP 0.456 vs. Procrustes 0.689) and higher post-projection CKA (rotation does not change CKA). **(iii) Procrustes nonetheless wins retrieval.** Procrustes beats the more flexible methods on P@1 on both the standard and the hard pool, while its MSE stays within range of Affine's. We hypothesize that this as an objective-metric mismatch: every method minimises Euclidean reconstruction error, but P@1 is rank-based and direction-sensitive. Because Procrustes is pure rotation, it preserves every angle in the source space, whereas Affine and MLP have no such constraint and would rescale axes anisotropically, lowering MSE at the cost of the preserving angular geometry. The orthogonality constraint preserves better neighbourhood structure.

Figure 5 visualises this directly for `eng–fra` at the final layer: projecting both languages into a shared 2D PCA basis and overlaying the Procrustes-mapped English representations on the matched French targets, the two clouds coincide while the mapped source retains the geometry of the original English cloud. This is related to a pre-Transformer

line of work on word-embedding alignment, where orthogonal mappings are shown to outperform unconstrained linear ones on retrieval (Mikolov et al., 2013; Xing et al., 2015b; Smith et al., 2017; Artetxe et al., 2018; Conneau et al., 2018). We observe the same dissociation in transformer hidden states.

4.3 Layers and residual structure analysis

Next, we verify that the mapping we find is reliable across layers, not only at the final layer reported in Table 3. Sweeping Procrustes across all 13 layers (Appendix E.1), retrieval stays well above the identity baseline at every layer and peaks in mid layers (6–8): `eng–fra` P@1=0.981 at layer 8; `eng–rus` 0.987 at layer 6. We confirm that rotation is usable across the depth of the network. A complementary dimensional analysis (Appendix E.2) examines where the post-Procrustes residual lives. The residual dimension is high-rank and concentrated outside the principal directions of the target; this is why neither Affine relaxation nor a one-layer MLP improves retrieval.

Our reconstruction experiment so far works geometrically; we now turn to Section 5 and test whether this rotation would work under a causal intervention setting.

5 Cross-Model Activation Patching

Reconstruction shows that a single rotation maps one model's representations onto another's geometrically. We now ask whether the same rotation, applied with no further training, could transfer to a different task functionally. We test this on a factual-recall task using cross-model activation patching, which substitutes a model's forward pass and measures whether the prediction follows the substitution. We test this using a factual recall task, which also carries functional content when used as a causal intervention.

5.1 Task and Intervention

Factual probe. We construct country→capital facts in six languages (English, French, German, Spanish, Japanese, Chinese), with hand-curated cloze prompts and answers, e.g. "`The capital of France is`" → `Paris`; "`Die Hauptstadt von Frankreich ist`" → `Paris`. The country word's last subword position is identified per language by a verb-marker heuristic (details in Appendix F).

Span patching at the concept position. For a (donor, target) cross-fact pair (with donor $\neq$ target country), we run the source-language model on the donor's source-language prompt and cache the residual stream $\mathbf{h}_S^{(k)}[c_S]$ at the donor country's last subword position c_S, for every layer $k \in \{0, \ldots, L\}$ (L=12). The target-language model is then run on its own target-language target prompt, but at the target country position c_T, we replace the residual $\mathbf{h}_T^{(k)}[c_T]$ with the projected donor residual $f^{(k)}\left(\mathbf{h}_S^{(k)}[c_S]\right)$ for all $k \in \{j, \ldots, L\}$. The intervention follows two design choices from Dumas et al. (2025): (i) the patch spans layers $j \to L$ rather than firing at a single layer, since a single-layer last-token patch is suppressed by downstream computation; and (ii) the patch lands on the concept word rather than the last token, so that the injected information must propagate through attention to influence the answer logits.

Conditions. We compare four sources for the patched residual: *within-lang* ($\mathbf{h}_T$, the target model patches its own donor representation, no projection; the ceiling on what any cross-lingual map could match), *Procrustes* ($W^{(k)}\mathbf{h}_S$, the rotation described in Section 4, applied per layer with no further training), *unprojected* ($\mathbf{h}_S$, raw source-model residual, basis-mismatch control), and *shuffle* ($\sim \mathcal{N}(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$, a Gaussian sample per layer matching the source model's activation mean and full covariance, distributional control). Affine (Section 4) was also evaluated but is excluded since it underperforms Procrustes on this probe; see Appendix F for the explanation.

Metric. We score with the *directional success rate*: the fraction of cross-fact pairs where $\Delta \log p(\text{donor capital}) > 0$ in the target language under patching, with Wilson 95% CIs over 870 cross-fact pairs per cell. The directional rate is approximately invariant to source-model strength asymmetries; raw $\Delta \log p$ is contaminated by the fact that some models hold country→capital knowledge more strongly than others.

5.2 Cross-Model Transfer of Factual Content

We report Procrustes as the cross-lingual projection of record (see Appendix F for discussions on Affine). Figure 6 reports cross-model transfer across five `eng`→{`fra`, `deu`, `spa`, `jpn`, `zho`} target languages at the patch-start layer that best separates Procrustes from controls per target. Three observations follow.

(i) Establishing the ceiling by patching with

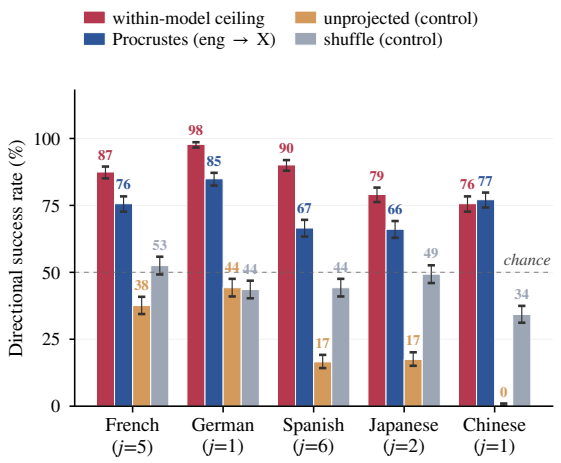

Figure 6: **Directional success rate (%) for eng→{fra, deu, spa, jpn, zho}**, n=870 cross-fact pairs per cell, Wilson 95% CIs. Bars: within-model ceiling, Procrustes, unprojected residual, and shuffled Gaussian sample. Per target, the patch start layer $j \in \{1, \ldots, 8\}$ is selected to maximise `within` and `Procrustes`. Dashed line at 50% chance.

the same-language representation. Before evaluating any cross-model projection we verify that the intervention itself can carry factual content when no cross-lingual transfer is required. Within-model patching yields directional success rates of 76–98% across the five targets, reproducing the Dumas et al. (2025) concept-position span-patch design in our setting and giving an upper bound on what any cross-lingual map could match.

(ii) The fitted rotation from Section 4 successfully transfers factual concepts between monolingual models. The Procrustes map from Section 4 was fit on sentence-end pooled activations. We evaluate it here without retraining on a different task (factual cloze rather than news), at a different token position (the concept word mid-prompt rather than the sequence end), and scored against a different readout (next-token logits rather than cosine retrieval). Remarkably, the rotation reaches directional success rates of 66–85% across the five targets, against a 76–98% within-model ceiling, recovering most of what an oracle same-model intervention would achieve. We also see that the pattern holds across different languages and scripts.

(iii) Control settings (shuffle and unprojected) are cleanly near or below the chance line. The shuffled control sits approximately at the 50% chance line, as expected for a random perturbation. However, the unprojected control falls clearly below chance. We hypothesise that unprojected carries the donor concept in a basis that is mismatched with the target model, so when read by the target model's LM head it acts as a coherent signal pointing in the wrong direction, and sometimes produces the opposite effect.

Taking eng→deu as a concrete illustration: patching the German model's own donor representation flips the prediction to the donor's capital 98% of the time; the rotated English residual recovers this in 85% of cross-fact pairs; a random Gaussian sample sits near 50%. A single orthogonal map, fit once on parallel sentence pools, recovers most of what an oracle same-model intervention would do, while uninformed signals stay at or below chance.

6 Conclusion and Discussion

We set out asking whether strictly monolingual language models could be aligned at all. At the **correlational** stage, alignment between models exceeds shuffled controls on all pairs. At the **constructional** stage, an orthogonal rotation recovers one model's hidden states from another's. At the **causal** stage, the same rotation could patch a English residual into a German model and flip the factual prediction in 85% of cases. This evidence confirms that cross-lingual alignment can emerge from the structure of language and the world it describes, with no joint training required.

More broadly speaking, we could think of multilinguality as a special case of the more general phenomenon of representational convergence (Moschella et al., 2023). The cross-lingual structure observed in multilingual language models is partly inherited from this convergence and partly produced by joint training (K et al., 2020; Artetxe et al., 2020); and our results suggest that the inherited component is substantial, and that alignment already present in monolingual models can be recovered post hoc. This opens a modular alternative to current multilingual pipelines: a small set of anchor pairs is in principle enough to stitch independently trained monolingual specialists into a shared space (Lenc and Vedaldi, 2015; Bansal et al., 2021), and the same machinery extends to model merging across languages (Ilharco et al., 2023; Yadav et al., 2023). Whether stitched or merged monolingual models match jointly trained ones on downstream tasks, and how to scale anchor-based projection to typologically distant pairs, are valuable future directions.

Limitations

Language and model coverage. The nine-language Goldfish analysis and the five-model ∼1B replication are weighted towards higher-resource, predominantly Indo-European languages, reflecting which truly monolingual checkpoints are publicly available. We cap at ∼1B parameters because most larger releases marketed as monolingual retain measurable multilingual content in their pretraining data. Since alignment increases monotonically with training data (Section 3.3), we expect the same pattern to extend to 7B+ monolingual checkpoints, but this remains untested.

Parallel-data assumption. Rotations are fitted on parallel sentence pairs. Although this is small in absolute terms, sentence-aligned data is still not free for low-resource languages, and we do not evaluate unsupervised mapping alternatives.

Scope of the causal evaluation. Cross-model activation patching is run on a single relation (country→capital) at a single concept-bearing span. Multi-token concepts, additional relational families, and open-ended generation fall outside this design, so the observed transferability is only shown on one well-controlled probe.

Use of AI assistants.

We use AI coding assistants for implementation support. All experimental design, analyses, and claims are the authors' own.

References

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. 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)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2026. Goldfish: Monolingual language models for 350 languages. *arXiv preprint arXiv:2408.10441*.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *Preprint*, arXiv:1710.04087.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 6022–6034, Online. Association for Computational Linguistics.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. Tucano: Advancing Neural Text Generation for Portuguese. *Patterns*, 6(11).

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2022. Reliability of cka as a similarity measure in deep learning. *Preprint*, arXiv:2210.16156.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Fabian Gröger, Shuo Wen, and Maria Brbić. 2026. Revisiting the platonic representation hypothesis: An aristotelian view. *Preprint*, arXiv:2602.14486.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *Preprint*, arXiv:2405.07987.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John Morris. 2026. Harnessing the universal geometry of embeddings. *Advances in Neural Information Processing Systems*, 38:45963–45987.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. *Preprint*, arXiv:1912.07840.

A Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A Efros. 2026. Back into plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR.

Felicia Körner, Max Müller-Eberstein, Anna Korhonen, and Barbara Plank. 2026. When meanings meet: Investigating the emergence and quality of shared concept spaces during multilingual language model training. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3149–3169, Rabat, Morocco. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Karel Lenc and Andrea Vedaldi. 2015. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–999.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2025. A systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *Preprint*, arXiv:1309.4168.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable

zero-shot latent space communication. *Preprint*, arXiv:2209.15430.

Niklas Muennighoff. 2022. SGPT: GPT Sentence Embeddings for Semantic Search. *Preprint*, arXiv:2202.08904.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. 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: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025. Bielik v3 small: Technical report. *Preprint*, arXiv:2505.02550.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Frederick Riemenschneider and Anette Frank. 2025. Cross-lingual generalization and compression: From language-specific to shared neurons. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13470–13491, Vienna, Austria. Association for Computational Linguistics.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *Preprint*, arXiv:1702.03859.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of ACL*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015a. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL-HLT*, pages 1006–1011.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015b. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Proceedings of NeurIPS*.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling. In *International Conference on Learning Representations*, volume 2025, pages 95372–95392.

A Full Pairwise CKA Results

Tables 8–19 (Appendix G) report the complete pairwise last-layer linear CKA values for all $\binom{9}{2} = 36$ language pairs across nine monolingual 1000 MB Goldfish models. Each cell shows *matched / shuffled* CKA, where *matched* denotes CKA computed on parallel (translation-equivalent) sentences and *shuffled* denotes a control in which the sentence order of one language is randomly permuted. Tables are grouped by pooling method: sentence-level mean pooling (Tables 8–11), token-level word-aligned (Tables 12–15), and SGPT position-weighted pooling (Tables 16–19), each evaluated on FLORES, Tatoeba, OPUS, and BouQUET. Missing entries (—) indicate language pairs for which parallel data was unavailable in that dataset.

B CKA Comparison with Training Data

This appendix supports Section 3.3 with the full 9×9 pairwise CKA matrices at each of the four Goldfish training-data scales (5 MB, 10 MB, 100 MB, 1000 MB), and with a cross-size comparison that pairs models trained on *different* amounts of data. All analyses use last-layer pairwise linear CKA across the same nine monolingual Goldfish languages from Appendix A, with a shuffled control throughout.

B.1 Does Alignment Grow with Training Data?

Figures 14–16 (Appendix G) show the full 9×9 pairwise CKA matrices at each of the four data scales, for all three CKA settings: sentence-level mean pooling, token-level word-aligned, and SGPT position-weighted pooling.

Findings. Across all three CKA settings, matched off-diagonal CKA increases monotonically from 5 MB to 1000 MB. Shuffled baselines remain near zero at every scale, so the matched−shuffled gap widens with more training data. This confirms and extends the size-scaling result from Section 3.3: increased data exposure yields more semantically structured internal representations, consistently across all pooling strategies.

B.2 Cross-Size Comparison: Do Models of Different Scales Align?

Figures 7–9 show cross-size CKA matrices, where Model A and Model B are trained on *different* amounts of data. We consider two cross-size pairings: 1000 MB vs. 100 MB, and 10 MB vs. 5 MB.

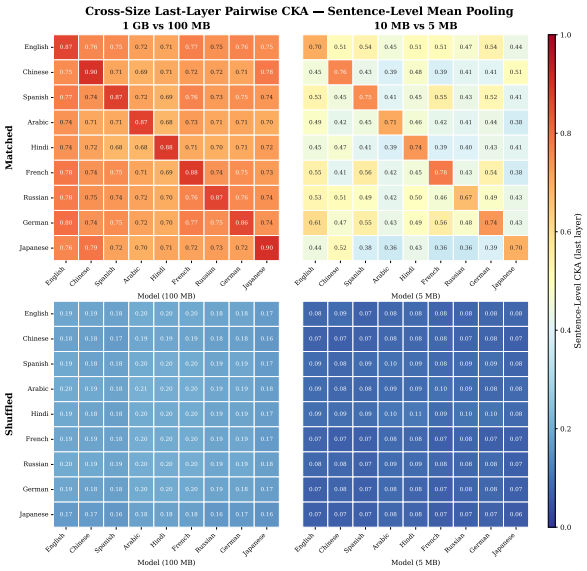

Figure 7: Sentence-Level Mean Pooling: cross-size CKA at the last layer. Left pair: 1000 MB models (rows) vs. 100 MB models (columns). Right pair: 10 MB vs. 5 MB. Within each pair, matched and shuffled are shown separately. Diagonal entries correspond to the *same language* compared across sizes; off-diagonal entries correspond to *different languages* at different scales.

Findings. Two consistent patterns emerge across all three settings.

(i) Same-language diagonal dominates. When comparing, say, the English 1000 MB model against the English 100 MB model, the diagonal CKA is substantially higher than any off-diagonal cell in the same row.

(ii) Matched signal survives across sizes. In the 1000 MB vs. 100 MB comparison, off-diagonal matched CKA values remain clearly above the shuffled baseline, indicating that independently trained monolingual models at different scales still encode partially compatible concept representations. The 10 MB vs. 5 MB comparison shows lower absolute values (consistent with the within-size scaling result), but the matched-vs-shuffled gap persists.

C Cross-Architecture Comparison

This appendix supports Section 3.4 with the full model specifications for both cross-architecture experiments, and a first-layer analysis that complements the last-layer matrix in the main text.

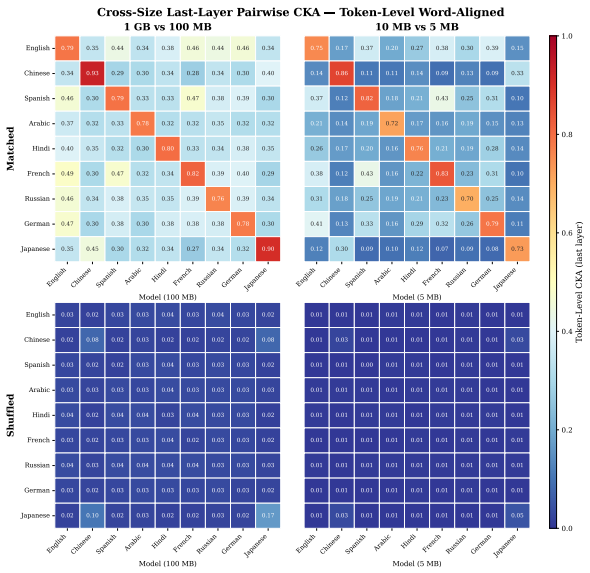

Figure 8: Token-Level Word-Aligned: cross-size CKA. Same layout as Figure 7.

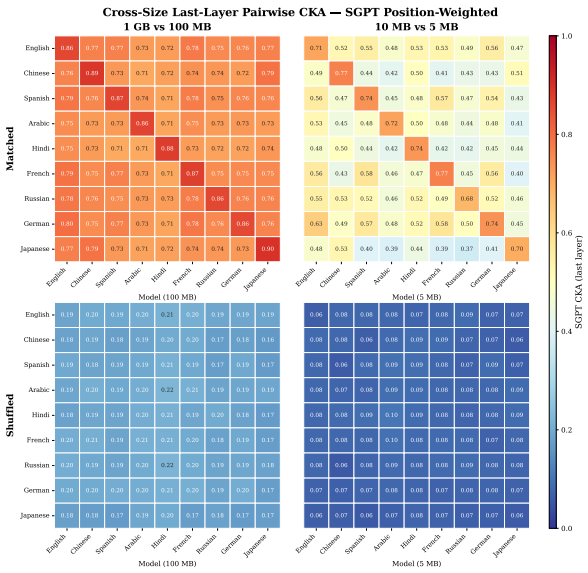

Figure 9: SGPT Position-Weighted Pooling: cross-size CKA. Same layout as Figure 7.

C.1 Goldfish vs. Pythia

We compare Goldfish against Pythia-160M (Biderman et al., 2023) (for English) and Zh-Pythia-160M (Liu et al., 2025) (for Chinese), which share Goldfish's hidden dimension and depth but differ in tokenizer, training data, and training pipeline. Figure 10 reports per-layer SGPT CKA. Two observations stand out: matched CKA exceeds the shuffled baseline at *every* layer for both English and Chinese, and the layerwise profile is U-shaped, with alignment high at embedding and final layers but depressed in the middle. This suggests that independently trained architectures converge on similar input- and output-facing semantic representations, but can diverge in their intermediate processing strategies. Table 4 lists the four models used.

Model	Lang	Vocab
Goldfish 125M	EN	51 200
Pythia-160M	EN	50 304
Goldfish 125M	ZH	51 200
Zh-Pythia-160M	ZH	50 304

Table 4: Models used in the Goldfish vs. Pythia comparison of Section 3.4. All four models share hidden dimension 768, 12 layers, and 12 attention heads; they differ in vocabulary size (shown), tokenizer, training data, and training procedure.

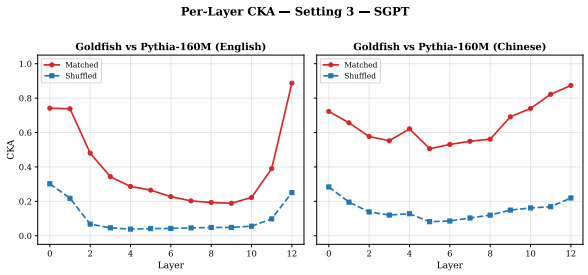

Figure 10: **Goldfish vs. Pythia per-layer CKA (SGPT).** Left: English (Goldfish-en vs. Pythia-160M). Right: Chinese (Goldfish-zh vs. Zh-Pythia-160M). Red: matched; blue: shuffled.

C.2 Diverse Independently-Developed Monolingual Models

Section 3.4 (Table 2) reports last-layer SGPT CKA across five independently developed ∼1B monolingual models that share no architecture, tokenizer, training data, or research group. Here we provide the full model specifications and a first-layer comparison that complements the last-layer matrix in the main text.

Models. Table 5 lists the five models. They span three architecture families (GPT-NeoX, LLaMA, Qwen2.5/Mistral), five languages, and five research groups, and differ in hidden dimension, depth, attention head count, and vocabulary.

Setup. We compute SGPT-pooled linear CKA for all $\binom{5}{2} = 10$ model pairs, with matched and shuffled controls. Because these models differ in depth, per-layer trajectories are not directly com-

Model	Lang	Arch	Hidden	Layers	Heads
Pythia-1.4B	EN	GPT-NeoX	2048	24	16
Zh-Pythia-1.4B	ZH	GPT-NeoX	2048	24	16
Tucano-1B	PT	LLaMA	2048	22	32
Bielik-1.5B	PL	Qwen2.5	1536	28	12
Minerva-1B	IT	Mistral	2048	16	16

Table 5: Independently developed monolingual models used in the shared-nothing comparison of Section 3.4.

parable; we therefore focus on the final pre-logit (last) layer, which represent the input and output boundaries of each model. Last-layer values are reported in Table 2 of the main text.

D Predictors of Alignment: Additional Details

For each of the 36 Goldfish language pairs, we correlate last-layer SGPT matched CKA against five pairwise predictors: four URIEL typological distances (Littell et al., 2017) (syntactic, genetic, geographic, inventory) and the mean per-sentence negative log-likelihood (NLL) of the two models on FLORES, which proxies how well each model has learned its language. URIEL distances are taken directly from the URIEL typological database: syntactic distance over morphosyntactic feature vectors, genetic distance over the language family tree, geographic distance as great-circle distance between language centroids, and inventory distance over phonological inventories. For each predictor we compute Pearson and Spearman correlations against last-layer SGPT CKA (matched condition).

Predictor	r	p_r	ρ	p_ρ
URIEL Syntactic	$\mathbf{-0.603}$	$<.001$	$\mathbf{-0.635}$	$<.001$
URIEL Genetic	-0.454	.005	-0.371	.026
Mean NLL	-0.440	.007	-0.425	.010
URIEL Geographic	-0.387	.020	-0.440	.007
URIEL Inventory	-0.209	.222	-0.123	.477

Table 6: **Predictors of last-layer matched CKA** across 36 Goldfish language pairs. r: Pearson; ρ: Spearman. p_r, p_ρ are two-sided.

Syntactic distance is the strongest single predictor of alignment ($r = -0.60$, $\rho = -0.64$, $p < .001$), followed by genetic distance, mean NLL, and geographic distance; phonological inventory distance is not significantly correlated. Figure 11 visualises the syntactic relationship directly: same-subfamily pairs cluster in the low-distance, high-CKA quadrant, while cross-family pairs drift toward the opposite corner. The four URIEL distances are themselves intercorrelated, and we do not attempt to dissociate their contributions; the ranking is reported as a descriptive summary of where alignment is strongest. The negative coefficient on mean NLL is consistent with a model-fit effect: pairs whose models have learned their language less well also show weaker pairwise alignment.

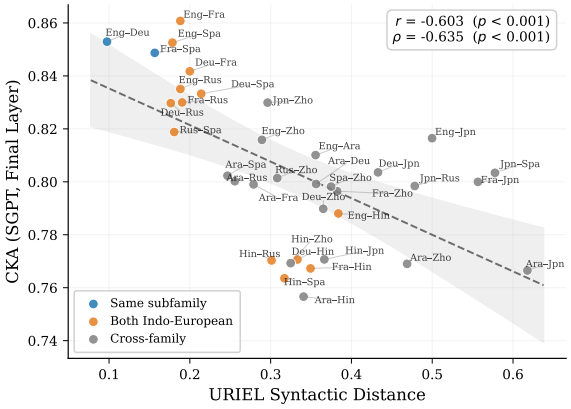

Figure 11: **Last-layer matched CKA vs. URIEL syntactic distance** (SGPT) across 36 Goldfish pairs. Points are coloured by genealogical relationship: same subfamily (blue), both Indo-European (orange), cross-family (grey). Dashed line: linear fit with 95% confidence band.

E Cross-Lingual Reconstruction: Method Comparison

This appendix supports Section 4 (*Cross-Lingual Representation Reconstruction*) with extended results comparing the three projection families (Procrustes, Affine with ridge, MLP) on all 36 language pairs at the final layer.

Per-pair comparison. Figure 12 plots Procrustes against Affine on retrieval P@1 (left) and test MSE (right) for each of the 36 language pairs at layer 12. Procrustes wins on P@1 for every pair (paired mean $\Delta = +0.112$, 95% CI $[+0.080, +0.143]$), while Affine wins on MSE for every pair. Together with Table 20, this dissociation shows that the metric on which a method is "best" depends on the metric being optimised: Affine attains lower MSE because its hypothesis class is strictly larger, but the orthogonality constraint of Procrustes preserves angular structure that is better for cosine retrieval.

Aggregate summary. Table 20 (Appendix G) reports mean P@1, MSE, and CKA across the 36 pairs with bootstrap 95% CIs, and the number of

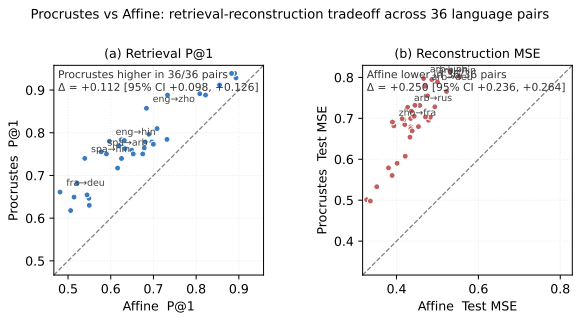

Figure 12: **Procrustes vs. Affine, all 36 pairs (layer 12).** Each point is one language pair. Left: retrieval P@1 against the train+test target pool; right: test MSE. Dashed line is $y = x$. Annotations report the paired mean difference Δ with bootstrap 95% CI (10,000 resamples).

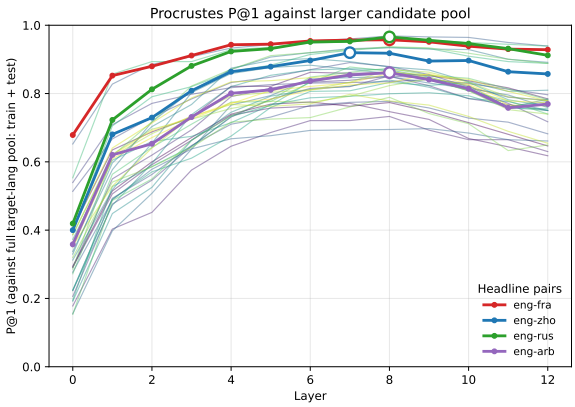

Figure 13: **Procrustes P@1 by layer** (train+test pool). Headline pairs (eng→{fra, rus, zho, arb}) bolded; remaining 32 pairs faint in background. Dots mark the per-pair best layer.

pairs on which each method achieves the best score for each metric.

Methods × metrics. Figure 17 (Appendix G) shows the full method×metric grid: each panel is one (method, metric) combination across all 36 pairs and 13 layers. This confirms that the ranking observed at layer 12 is stable across most layers, and that MLP and Affine track each other closely, but Procrustes consistently dominates on retrieval.

Visualising the rotation. Figure 18 (Appendix G) provides a qualitative view of what Procrustes is doing. We project the English source representations and the matched French target representations into the same 2D PCA basis and overlay the Procrustes-mapped English points. This illustrates that a single rotation of the English space lands each sentence near its French counterpart.

E.1 Layer-wise Analysis

We restrict to Procrustes and sweep layers 0–12 for all 36 pairs. Figure 13 reports retrieval P@1 by layer, with four representative pairs highlighted: eng→fra (typologically close), eng→rus (mid), eng→zho and eng→arb (distant). We see that retrieval is non-monotonic in depth and peaks at layers 6–8 across pairs (eng→fra: layer 8, P@1 =0.981; eng→rus: layer 6, 0.987; eng→zho: layer 8, 0.977; eng→arb: layer 8, 0.922), with the location of the peak weakly tracking typological distance.

E.2 Dimensional Analysis

Procrustes leaves a residual. A natural follow-up question is what that residual looks like, because the answer determines whether a more flexible map could have closed the gap. Two possibilities could happen: Either the residual error is concentrated along a small number of directions, in which case the two spaces are nearly isometric except along those axes and a low-rank correction would suffice; or the error is spread across many directions, in which case no simple relaxation will help. We distinguish the two by taking an SVD of the residual and comparing its spectrum to that of the target.

Setup. Let $X, Y \in \mathbb{R}^{n \times d}$ be the per-layer source and target representations and $W^\star$ the Procrustes solution from Section 4.1. The residual is $R = XW^\star - Y$. We take its SVD, $R = U_R \Sigma_R V_R^\top$, and the target's, $Y = U_Y \Sigma_Y V_Y^\top$, and compute two kinds of summary.

The first asks how spread-out a spectrum is. The entropy-based *effective rank*

$$\mathrm{eff_rank}(M) = \exp\Big(-\sum_i \tilde{\sigma}_i \log \tilde{\sigma}_i\Big), \quad \tilde{\sigma}_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2},$$

counts the directions a matrix effectively occupies (a perfectly rank-1 matrix has eff_rank 1; a uniform spectrum across d directions has eff_rank d). We complement it with k_α, the smallest k whose top singular values carry an α fraction of the squared norm.

The second asks where the residual sits relative to the target. We project R onto the top-k right-

singular vectors of Y, denoted $V_Y^{(k)}$, and report

$$\text{frac_out}_k = 1 - \frac{\|RV_Y^{(k)}\|_F^2}{\|R\|_F^2},$$

the fraction of residual energy lying outside the top-k directions of Y. A small value means the residual lives in the same subspace the target uses; a large value means it lives off to the side, in directions the target itself barely populates.

	Eff. rank		**Resid.** k_α		frac_out_k	
Pair	Y	R	50%	90%	k=10	k=50
eng–fra	67	246	90	357	.93	.73
eng–rus	45	202	74	273	.93	.73
eng–zho	43	180	69	265	.91	.69
eng–arb	53	147	71	306	.88	.68

Table 7: **Spectra of Y and the post-Procrustes residual R at layer 8** (d=768). Columns: target and residual effective rank; smallest k for which the top-k singular values of R capture 50%/90% of its squared norm; fraction of residual energy outside the top-k right-singular subspace of Y.

Findings. Table 7 reports these statistics for English paired with French, Russian, Chinese, and Arabic at the peak retrieval layer (layer 8). Three patterns hold across pairs and, with mild variation, across all 13 layers.

(i) The target is low-rank; the residual is not. The target representations use only 43–67 directions out of 768, and 90% of their variance fits in roughly 160–200 directions. The residual is far more diffuse: effective rank 147–246, with half its energy already requiring 70–90 singular values and 90% requiring 265–360. The error is therefore not stored in a small handful of bad axes that a targeted fix could correct; it is spread thinly across many directions.

(ii) The residual lives outside the part of space the target uses. Of the residual's total energy, 88–93% lies outside the top-10 directions of Y, and 68–74% outside the top-50. The Procrustes rotation lands cleanly in the subspace where the target stores its variance and leaves error in the directions the target barely populates. This is true even though the residual is large in absolute terms (Frobenius norm 0.79–0.93 of the target's at layer 8): a residual can be large in raw magnitude while being functionally irrelevant if it sits where the model does not read.

(iii) The pattern is stable across layers. The same picture holds throughout the network. Residual effective rank exceeds target effective rank by a factor of at least two (typically three to six). frac_out_{10} stays above 0.79, and frac_out_{50} above 0.57, at every layer for every headline pair. No layer admits a low-dimensional collapse of the residual.

These observations reconcile a tension in the main results. Procrustes leaves a residual that is large in Frobenius norm yet small in functional consequence, because the residual is high-rank and largely orthogonal to the subspace where the target stores its variance. There is no concentrated low-dimensional structure for an Affine relaxation or a one-layer MLP to fit, which explains why neither improves retrieval over Procrustes. What the data support is approximate isometry along the dominant directions of the target, with diffuse mismatch in the remaining ones.

F Span Patching: Setup, Conditions, and Auxiliary Numbers

This appendix supports Section 5 with extended specification of the country→capital probe, the conditions, and the per-pair numbers underlying Figure 6.

Facts and prompts. The probe consists of 30 country→capital facts, each available as a complete (prompt, answer) pair in six languages: English, French, German, Spanish, Japanese, and Chinese. Prompts share a fixed cloze form per language (e.g. "`The capital of` X `is`", "`La capitale de` X `est`", "`Die Hauptstadt von` X `ist`", "`La capital de` X `es`", "Xの首都は", "X的首都是"). Answers are the target language's name for the capital, scored using each language's first subword id, computed in context so that BPE splits are handled correctly per tokenizer.

Span vs. single-layer patching. A preliminary experiment patched the residual at a single layer at the prompt's last-token position (eng→fra at the position immediately before the LM head). Within-model $\Delta \log p(\text{donor})$ was $\sim$+0.07 nat, which doesn't seem to work well. This is becuase the intervention is too narrow: a single-layer last-token patch is immediately consumed by the LM head and downstream layers, with no opportunity to propagate. The span variant fixes this by (i) patching from j to L, so each layer sees the patched

residual at its input; and (ii) patching at the country (concept) position rather than the last token, so the patched information must move through attention to reach the answer logits.

Conditions. Four conditions per cross-fact cell are reported in the main results:

- `within-lang`: target model, target-language donor prompt, identity projection. Within-model ceiling.
- `Procrustes`: source model on source-language donor prompt; per-layer orthogonal map $W^{(k)}$ from Section 4, fit on FLORES/OPUS/BouQUET sentence-end activations.
- `unprojected`: source-model residual at c_S, identity projection. Controls for basis mismatch.
- `shuffle`: Gaussian sample $\sim \mathcal{N}(\boldsymbol{\mu}_S^{(k)}, \boldsymbol{\Sigma}_S^{(k)})$ per layer with full-covariance Cholesky factorisation (eigendecomposition fallback for rank-deficient covariances), where $(\boldsymbol{\mu}_S^{(k)}, \boldsymbol{\Sigma}_S^{(k)})$ are the moments of the source-model activation pool. Distributional control.

A fifth condition, `Affine` (the per-layer linear+bias projection from Section 4, no orthogonality constraint), was evaluated and is discussed below; we exclude it from the main results for reasons later in this section.

Settings. Goldfish 1000 MB models per language; 12 transformer blocks; hidden dim 768; per-language vocabulary; max sequence length 512; seed 42. Span starts $j \in \{1, \ldots, 8\}$; span end fixed at $L = 12$ (final block). Cross-fact pairs per (condition, j) cell: $30^2 - 30 = 870$ (excluding the diagonal where donor = target). The country position is identified per language by a verb marker preceding the noun phrase; the country's last subword is the token immediately before this marker. All cells run on a single GPU.

Why Affine is excluded. We also evaluated the Section 4 Affine projection on every eng→X pair. Affine produces a magnitude artefact rather than directional transfer. Inspection of $\Delta_d - \Delta_t = \Delta \log p(\text{donor capital}) - \Delta \log p(\text{target capital})$, a stricter directional metric that requires the donor capital to rise *more* than the target capital, is negative for Affine: it inflates the magnitude of the patched residual enough that the LM head boosts *both* capitals, with the target capital often boosted more. The sign of $\Delta \log p(\text{donor capital})$ on Affine therefore reflects a uniform bias shift rather than functional transfer of the donor concept, so we report Procrustes as the cross-lingual projection of record and omit Affine from Figure 6.

G Supplementary Tables and Figures

This appendix collects the wide tables and figures referenced from Appendices A, B, and E.

Table 8: Last-layer pairwise CKA (*matched / shuffled*) — Sentence-Level Mean Pooling, FLORES dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.80 / 0.22	0.84 / 0.23	0.80 / 0.24	0.78 / 0.23	0.85 / 0.24	0.82 / 0.24	0.85 / 0.24	0.81 / 0.21
zho	0.80 / 0.22	—	0.78 / 0.21	0.75 / 0.23	0.76 / 0.22	0.78 / 0.22	0.79 / 0.23	0.77 / 0.23	0.82 / 0.20
spa	0.84 / 0.23	0.78 / 0.21	—	0.78 / 0.24	0.74 / 0.23	0.83 / 0.24	0.80 / 0.24	0.82 / 0.24	0.78 / 0.21
arb	0.80 / 0.24	0.75 / 0.23	0.78 / 0.24	—	0.73 / 0.23	0.78 / 0.25	0.78 / 0.25	0.78 / 0.25	0.75 / 0.22
hin	0.78 / 0.23	0.76 / 0.22	0.74 / 0.23	0.73 / 0.23	—	0.75 / 0.24	0.75 / 0.24	0.76 / 0.24	0.77 / 0.21
fra	0.85 / 0.24	0.78 / 0.22	0.83 / 0.24	0.78 / 0.25	0.75 / 0.24	—	0.82 / 0.24	0.83 / 0.24	0.78 / 0.21
rus	0.82 / 0.24	0.79 / 0.23	0.80 / 0.24	0.78 / 0.25	0.75 / 0.24	0.82 / 0.24	—	0.82 / 0.25	0.79 / 0.22
deu	0.85 / 0.24	0.77 / 0.23	0.82 / 0.24	0.78 / 0.25	0.76 / 0.24	0.83 / 0.24	0.82 / 0.25	—	0.79 / 0.21
jpn	0.81 / 0.21	0.82 / 0.20	0.78 / 0.21	0.75 / 0.22	0.77 / 0.21	0.78 / 0.21	0.79 / 0.22	0.79 / 0.21	—

Table 9: Last-layer pairwise CKA (*matched / shuffled*) — Sentence-Level Mean Pooling, Tatoeba dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.45 / 0.17	0.78 / 0.21	0.64 / 0.22	0.65 / 0.26	0.75 / 0.23	0.70 / 0.22	0.81 / 0.23	0.59 / 0.25
zho	0.45 / 0.17	—	0.60 / 0.23	0.51 / 0.31	0.50 / 0.20	0.57 / 0.21	0.53 / 0.21	0.63 / 0.26	0.43 / 0.16
spa	0.78 / 0.21	0.60 / 0.23	—	0.68 / 0.24	0.85 / 0.61	0.72 / 0.23	0.70 / 0.23	0.74 / 0.24	0.62 / 0.27
arb	0.64 / 0.22	0.51 / 0.31	0.68 / 0.24	—	—	0.65 / 0.26	0.66 / 0.25	0.65 / 0.25	0.58 / 0.26
hin	0.65 / 0.26	0.50 / 0.20	0.85 / 0.61	—	—	0.83 / 0.46	0.86 / 0.66	0.84 / 0.58	0.74 / 0.33
fra	0.75 / 0.23	0.57 / 0.21	0.72 / 0.23	0.65 / 0.26	0.83 / 0.46	—	0.67 / 0.21	0.75 / 0.25	0.53 / 0.25
rus	0.70 / 0.22	0.53 / 0.21	0.70 / 0.23	0.66 / 0.25	0.86 / 0.66	0.67 / 0.21	—	0.73 / 0.26	0.65 / 0.27
deu	0.81 / 0.23	0.63 / 0.26	0.74 / 0.24	0.65 / 0.25	0.84 / 0.58	0.75 / 0.25	0.73 / 0.26	—	0.61 / 0.26
jpn	0.59 / 0.25	0.43 / 0.16	0.62 / 0.27	0.58 / 0.26	0.74 / 0.33	0.53 / 0.25	0.65 / 0.27	0.61 / 0.26	—

Table 10: Last-layer pairwise CKA (*matched / shuffled*) — Sentence-Level Mean Pooling, OPUS dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.83 / 0.09	0.81 / 0.17	0.70 / 0.20	0.34 / 0.09	0.83 / 0.16	0.80 / 0.16	0.79 / 0.18	0.53 / 0.27
zho	0.83 / 0.09	—	—	0.74 / 0.13	—	0.70 / 0.15	0.73 / 0.15	0.61 / 0.18	—
spa	0.81 / 0.17	—	—	—	—	—	—	—	—
arb	0.70 / 0.20	0.74 / 0.13	—	—	—	0.84 / 0.12	0.85 / 0.11	0.53 / 0.22	—
hin	0.34 / 0.09	—	—	—	—	—	—	—	—
fra	0.83 / 0.16	0.70 / 0.15	—	0.84 / 0.12	—	—	0.84 / 0.12	0.81 / 0.15	—
rus	0.80 / 0.16	0.73 / 0.15	—	0.85 / 0.11	—	0.84 / 0.12	—	0.62 / 0.24	—
deu	0.79 / 0.18	0.61 / 0.18	—	0.53 / 0.22	—	0.81 / 0.15	0.62 / 0.24	—	—
jpn	0.53 / 0.27	—	—	—	—	—	—	—	—

Table 11: Last-layer pairwise CKA (*matched / shuffled*) — Sentence-Level Mean Pooling, BouQUET dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.73 / 0.24	0.83 / 0.26	0.60 / 0.21	0.76 / 0.24	0.83 / 0.25	0.81 / 0.27	0.81 / 0.27	0.76 / 0.26
zho	0.73 / 0.24	—	0.74 / 0.22	0.55 / 0.20	0.69 / 0.24	0.73 / 0.23	0.75 / 0.22	0.73 / 0.24	0.75 / 0.25
spa	0.83 / 0.26	0.74 / 0.22	—	0.60 / 0.20	0.72 / 0.22	0.82 / 0.25	0.79 / 0.26	0.79 / 0.24	0.73 / 0.25
arb	0.60 / 0.21	0.55 / 0.20	0.60 / 0.20	—	0.54 / 0.20	0.57 / 0.20	0.56 / 0.21	0.57 / 0.19	0.54 / 0.20
hin	0.76 / 0.24	0.69 / 0.24	0.72 / 0.22	0.54 / 0.20	—	0.73 / 0.25	0.76 / 0.25	0.71 / 0.25	0.69 / 0.26
fra	0.83 / 0.25	0.73 / 0.23	0.82 / 0.25	0.57 / 0.20	0.73 / 0.25	—	0.80 / 0.24	0.79 / 0.25	0.73 / 0.25
rus	0.81 / 0.27	0.75 / 0.22	0.79 / 0.26	0.56 / 0.21	0.76 / 0.25	0.80 / 0.24	—	0.79 / 0.24	0.73 / 0.28
deu	0.81 / 0.27	0.73 / 0.24	0.79 / 0.24	0.57 / 0.19	0.71 / 0.25	0.79 / 0.25	0.79 / 0.24	—	0.75 / 0.24
jpn	0.76 / 0.26	0.75 / 0.25	0.73 / 0.25	0.54 / 0.20	0.69 / 0.26	0.73 / 0.25	0.73 / 0.28	0.75 / 0.24	—

Table 12: Last-layer pairwise CKA (*matched / shuffled*) — Token-Level Word-Aligned, FLORES dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.37 / 0.03	0.51 / 0.04	0.40 / 0.04	0.43 / 0.04	0.54 / 0.04	0.51 / 0.04	0.52 / 0.03	0.37 / 0.03
zho	0.37 / 0.03	—	0.32 / 0.03	0.33 / 0.03	0.37 / 0.03	0.32 / 0.03	0.36 / 0.03	0.31 / 0.03	0.43 / 0.10
spa	0.51 / 0.04	0.32 / 0.03	—	0.38 / 0.04	0.36 / 0.04	0.53 / 0.03	0.43 / 0.04	0.43 / 0.04	0.32 / 0.03
arb	0.40 / 0.04	0.33 / 0.03	0.38 / 0.04	—	0.35 / 0.04	0.38 / 0.04	0.40 / 0.04	0.35 / 0.04	0.34 / 0.03
hin	0.43 / 0.04	0.37 / 0.03	0.36 / 0.04	0.35 / 0.04	—	0.38 / 0.04	0.38 / 0.04	0.42 / 0.04	0.37 / 0.02
fra	0.54 / 0.04	0.32 / 0.03	0.53 / 0.03	0.38 / 0.04	0.38 / 0.04	—	0.44 / 0.04	0.44 / 0.03	0.31 / 0.03
rus	0.51 / 0.04	0.36 / 0.03	0.43 / 0.04	0.40 / 0.04	0.38 / 0.04	0.44 / 0.04	—	0.43 / 0.04	0.36 / 0.03
deu	0.52 / 0.03	0.31 / 0.03	0.43 / 0.04	0.35 / 0.04	0.42 / 0.04	0.44 / 0.03	0.43 / 0.04	—	0.32 / 0.03
jpn	0.37 / 0.03	0.43 / 0.10	0.32 / 0.03	0.34 / 0.03	0.37 / 0.02	0.31 / 0.03	0.36 / 0.03	0.32 / 0.03	—

Table 13: Last-layer pairwise CKA (*matched / shuffled*) — Token-Level Word-Aligned, Tatoeba dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.17 / 0.04	0.51 / 0.07	0.38 / 0.07	0.39 / 0.09	0.51 / 0.06	0.51 / 0.06	0.50 / 0.05	0.30 / 0.08
zho	0.17 / 0.04	—	0.27 / 0.07	0.33 / 0.12	0.23 / 0.05	0.22 / 0.04	0.24 / 0.06	0.25 / 0.06	0.44 / 0.19
spa	0.51 / 0.07	0.27 / 0.07	—	0.47 / 0.11	0.61 / 0.28	0.54 / 0.06	0.48 / 0.08	0.43 / 0.06	0.30 / 0.09
arb	0.38 / 0.07	0.33 / 0.12	0.47 / 0.11	—	—	0.39 / 0.09	0.46 / 0.09	0.38 / 0.09	0.32 / 0.10
hin	0.39 / 0.09	0.23 / 0.05	0.61 / 0.28	—	—	0.56 / 0.19	0.58 / 0.27	0.55 / 0.20	0.42 / 0.10
fra	0.51 / 0.06	0.22 / 0.04	0.54 / 0.06	0.39 / 0.09	0.56 / 0.19	—	0.52 / 0.05	0.46 / 0.05	0.26 / 0.07
rus	0.51 / 0.06	0.24 / 0.06	0.48 / 0.08	0.46 / 0.09	0.58 / 0.27	0.52 / 0.05	—	0.48 / 0.07	0.33 / 0.08
deu	0.50 / 0.05	0.25 / 0.06	0.43 / 0.06	0.38 / 0.09	0.55 / 0.20	0.46 / 0.05	0.48 / 0.07	—	0.29 / 0.07
jpn	0.30 / 0.08	0.44 / 0.19	0.30 / 0.09	0.32 / 0.10	0.42 / 0.10	0.26 / 0.07	0.33 / 0.08	0.29 / 0.07	—

Table 14: Last-layer pairwise CKA (*matched / shuffled*) — Token-Level Word-Aligned, OPUS dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.38 / 0.02	0.59 / 0.05	0.53 / 0.07	0.41 / 0.04	0.57 / 0.04	0.56 / 0.05	0.58 / 0.06	0.27 / 0.09
zho	0.38 / 0.02	—	—	0.35 / 0.02	—	0.36 / 0.02	0.39 / 0.02	0.47 / 0.03	—
spa	0.59 / 0.05	—	—	—	—	—	—	—	—
arb	0.53 / 0.07	0.35 / 0.02	—	—	—	0.46 / 0.04	0.50 / 0.04	0.41 / 0.09	—
hin	0.41 / 0.04	—	—	—	—	—	—	—	—
fra	0.57 / 0.04	0.36 / 0.02	—	0.46 / 0.04	—	—	0.50 / 0.04	0.51 / 0.04	—
rus	0.56 / 0.05	0.39 / 0.02	—	0.50 / 0.04	—	0.50 / 0.04	—	0.45 / 0.07	—
deu	0.58 / 0.06	0.47 / 0.03	—	0.41 / 0.09	—	0.51 / 0.04	0.45 / 0.07	—	—
jpn	0.27 / 0.09	—	—	—	—	—	—	—	—

Table 15: Last-layer pairwise CKA (*matched / shuffled*) — Token-Level Word-Aligned, BouQUET dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.34 / 0.04	0.54 / 0.06	0.37 / 0.06	0.44 / 0.06	0.54 / 0.05	0.53 / 0.06	0.51 / 0.05	0.38 / 0.06
zho	0.34 / 0.04	—	0.35 / 0.04	0.31 / 0.05	0.36 / 0.05	0.32 / 0.04	0.38 / 0.05	0.34 / 0.05	0.67 / 0.16
spa	0.54 / 0.06	0.35 / 0.04	—	0.40 / 0.06	0.39 / 0.06	0.58 / 0.06	0.50 / 0.06	0.48 / 0.06	0.35 / 0.05
arb	0.37 / 0.06	0.31 / 0.05	0.40 / 0.06	—	0.33 / 0.06	0.35 / 0.06	0.37 / 0.06	0.33 / 0.06	0.34 / 0.06
hin	0.44 / 0.06	0.36 / 0.05	0.39 / 0.06	0.33 / 0.06	—	0.41 / 0.06	0.40 / 0.06	0.44 / 0.06	0.37 / 0.05
fra	0.54 / 0.05	0.32 / 0.04	0.58 / 0.06	0.35 / 0.06	0.41 / 0.06	—	0.50 / 0.06	0.46 / 0.05	0.35 / 0.06
rus	0.53 / 0.06	0.38 / 0.05	0.50 / 0.06	0.37 / 0.06	0.40 / 0.06	0.50 / 0.06	—	0.47 / 0.06	0.38 / 0.06
deu	0.51 / 0.05	0.34 / 0.05	0.48 / 0.06	0.33 / 0.06	0.44 / 0.06	0.46 / 0.05	0.47 / 0.06	—	0.35 / 0.05
jpn	0.38 / 0.06	0.67 / 0.16	0.35 / 0.05	0.34 / 0.06	0.37 / 0.05	0.35 / 0.06	0.38 / 0.06	0.35 / 0.05	—

Table 16: Last-layer pairwise CKA (*matched / shuffled*) — SGPT Position-Weighted Pooling, FLORES dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.82 / 0.23	0.85 / 0.24	0.81 / 0.24	0.79 / 0.24	0.86 / 0.24	0.84 / 0.25	0.85 / 0.24	0.82 / 0.23
zho	0.82 / 0.23	—	0.80 / 0.22	0.77 / 0.23	0.77 / 0.23	0.80 / 0.24	0.80 / 0.23	0.79 / 0.24	0.83 / 0.21
spa	0.85 / 0.24	0.80 / 0.22	—	0.80 / 0.23	0.76 / 0.23	0.85 / 0.23	0.82 / 0.23	0.83 / 0.24	0.80 / 0.22
arb	0.81 / 0.24	0.77 / 0.23	0.80 / 0.23	—	0.76 / 0.25	0.80 / 0.25	0.80 / 0.25	0.80 / 0.24	0.77 / 0.24
hin	0.79 / 0.24	0.77 / 0.23	0.76 / 0.23	0.76 / 0.25	—	0.77 / 0.24	0.77 / 0.25	0.77 / 0.24	0.77 / 0.22
fra	0.86 / 0.24	0.80 / 0.24	0.85 / 0.23	0.80 / 0.25	0.77 / 0.24	—	0.83 / 0.24	0.84 / 0.24	0.80 / 0.22
rus	0.84 / 0.25	0.80 / 0.23	0.82 / 0.23	0.80 / 0.25	0.77 / 0.25	0.83 / 0.24	—	0.83 / 0.25	0.80 / 0.23
deu	0.85 / 0.24	0.79 / 0.24	0.83 / 0.24	0.80 / 0.24	0.77 / 0.24	0.84 / 0.24	0.83 / 0.25	—	0.80 / 0.22
jpn	0.82 / 0.23	0.83 / 0.21	0.80 / 0.22	0.77 / 0.24	0.77 / 0.22	0.80 / 0.22	0.80 / 0.23	0.80 / 0.22	—

Table 17: Last-layer pairwise CKA (*matched / shuffled*) — SGPT Position-Weighted Pooling, Tatoeba dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.43 / 0.17	0.81 / 0.21	0.65 / 0.22	0.66 / 0.25	0.76 / 0.24	0.72 / 0.23	0.81 / 0.23	0.63 / 0.24
zho	0.43 / 0.17	—	0.62 / 0.22	0.52 / 0.33	0.48 / 0.19	0.59 / 0.22	0.52 / 0.21	0.64 / 0.24	0.40 / 0.15
spa	0.81 / 0.21	0.62 / 0.22	—	0.69 / 0.24	0.87 / 0.65	0.76 / 0.23	0.74 / 0.24	0.76 / 0.24	0.65 / 0.26
arb	0.65 / 0.22	0.52 / 0.33	0.69 / 0.24	—	—	0.67 / 0.26	0.68 / 0.24	0.66 / 0.26	0.60 / 0.23
hin	0.66 / 0.25	0.48 / 0.19	0.87 / 0.65	—	—	0.83 / 0.51	0.85 / 0.63	0.85 / 0.57	0.75 / 0.32
fra	0.76 / 0.24	0.59 / 0.22	0.76 / 0.23	0.67 / 0.26	0.83 / 0.51	—	0.70 / 0.23	0.76 / 0.24	0.60 / 0.26
rus	0.72 / 0.23	0.52 / 0.21	0.74 / 0.24	0.68 / 0.24	0.85 / 0.63	0.70 / 0.23	—	0.74 / 0.26	0.68 / 0.26
deu	0.81 / 0.23	0.64 / 0.24	0.76 / 0.24	0.66 / 0.26	0.85 / 0.57	0.76 / 0.24	0.74 / 0.26	—	0.64 / 0.26
jpn	0.63 / 0.24	0.40 / 0.15	0.65 / 0.26	0.60 / 0.23	0.75 / 0.32	0.60 / 0.26	0.68 / 0.26	0.64 / 0.26	—

Table 18: Last-layer pairwise CKA (*matched / shuffled*) — SGPT Position-Weighted Pooling, OPUS dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.85 / 0.14	0.83 / 0.16	0.71 / 0.20	0.34 / 0.10	0.85 / 0.15	0.82 / 0.15	0.81 / 0.19	0.54 / 0.27
zho	0.85 / 0.14	—	—	0.74 / 0.14	—	0.72 / 0.16	0.74 / 0.16	0.63 / 0.18	—
spa	0.83 / 0.16	—	—	—	—	—	—	—	—
arb	0.71 / 0.20	0.74 / 0.14	—	—	—	0.86 / 0.13	0.86 / 0.13	0.53 / 0.23	—
hin	0.34 / 0.10	—	—	—	—	—	—	—	—
fra	0.85 / 0.15	0.72 / 0.16	—	0.86 / 0.13	—	—	0.85 / 0.11	0.82 / 0.13	—
rus	0.82 / 0.15	0.74 / 0.16	—	0.86 / 0.13	—	0.85 / 0.11	—	0.63 / 0.23	—
deu	0.81 / 0.19	0.63 / 0.18	—	0.53 / 0.23	—	0.82 / 0.13	0.63 / 0.23	—	—
jpn	0.54 / 0.27	—	—	—	—	—	—	—	—

Table 19: Last-layer pairwise CKA (*matched / shuffled*) — SGPT Position-Weighted Pooling, BouQUET dataset. All values computed on 1000 MB Goldfish models.

	eng	zho	spa	arb	hin	fra	rus	deu	jpn
eng	—	0.75 / 0.25	0.83 / 0.27	0.60 / 0.20	0.76 / 0.25	0.83 / 0.24	0.81 / 0.27	0.81 / 0.26	0.77 / 0.27
zho	0.75 / 0.25	—	0.75 / 0.23	0.55 / 0.20	0.70 / 0.25	0.74 / 0.24	0.76 / 0.23	0.74 / 0.24	0.76 / 0.24
spa	0.83 / 0.27	0.75 / 0.23	—	0.61 / 0.20	0.72 / 0.24	0.83 / 0.26	0.80 / 0.24	0.80 / 0.25	0.75 / 0.25
arb	0.60 / 0.20	0.55 / 0.20	0.61 / 0.20	—	0.54 / 0.20	0.57 / 0.21	0.56 / 0.20	0.56 / 0.20	0.55 / 0.21
hin	0.76 / 0.25	0.70 / 0.25	0.72 / 0.24	0.54 / 0.20	—	0.74 / 0.24	0.76 / 0.25	0.71 / 0.25	0.70 / 0.26
fra	0.83 / 0.24	0.74 / 0.24	0.83 / 0.26	0.57 / 0.21	0.74 / 0.24	—	0.81 / 0.24	0.79 / 0.26	0.75 / 0.25
rus	0.81 / 0.27	0.76 / 0.23	0.80 / 0.24	0.56 / 0.20	0.76 / 0.25	0.81 / 0.24	—	0.79 / 0.24	0.75 / 0.26
deu	0.81 / 0.26	0.74 / 0.24	0.80 / 0.25	0.56 / 0.20	0.71 / 0.25	0.79 / 0.26	0.79 / 0.24	—	0.76 / 0.25
jpn	0.77 / 0.27	0.76 / 0.24	0.75 / 0.25	0.55 / 0.21	0.70 / 0.26	0.75 / 0.25	0.75 / 0.26	0.76 / 0.25	—

Method	Mean P@1	Mean MSE	Mean CKA	Best P@1	Best MSE	Best CKA
Procrustes	0.775 [0.747, 0.803]	0.689 [0.662, 0.715]	0.713 [0.689, 0.735]	36/36	0/36	0/36
Affine	0.663 [0.628, 0.699]	0.439 [0.422, 0.455]	0.762 [0.741, 0.781]	0/36	36/36	32/36
MLP	0.643 [0.605, 0.682]	0.456 [0.440, 0.473]	0.760 [0.739, 0.778]	0/36	0/36	4/36

Table 20: **Method comparison across all 36 pairs at layer 12.** Cells show mean across 36 pairs with bootstrap 95% CI (10,000 resamples). "Best on metric" counts the number of pairs on which each method achieves the top score.

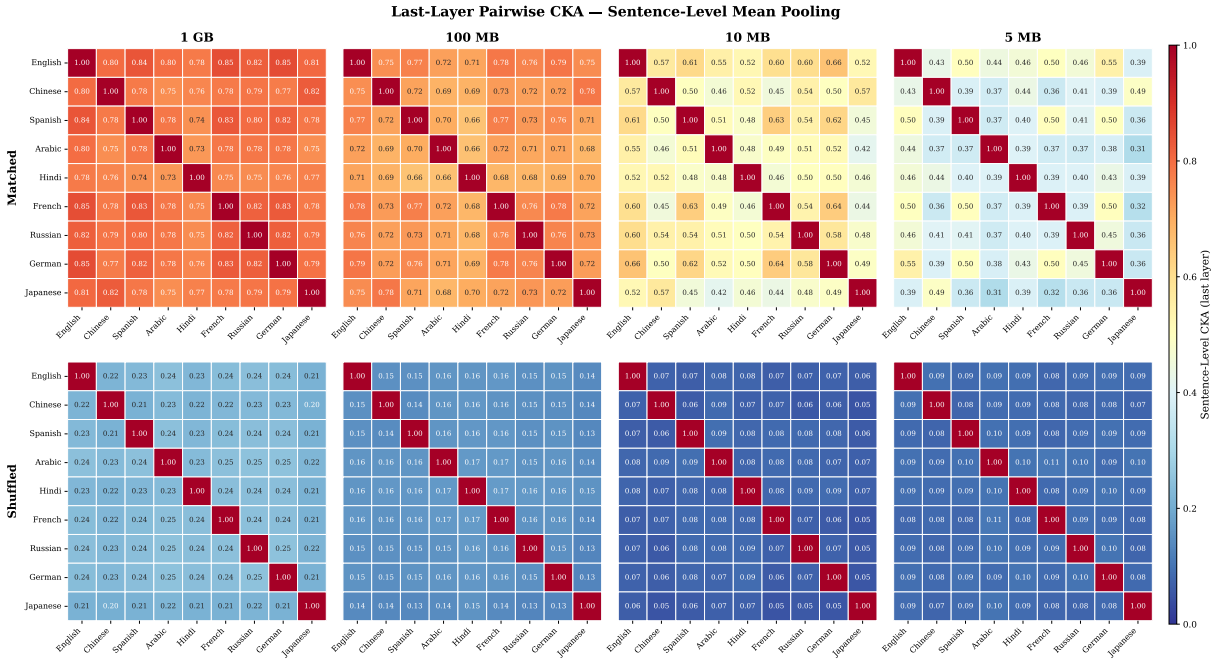

Figure 14: Setting 1 — Sentence-Level Mean Pooling: last-layer pairwise CKA across all nine Goldfish languages at four training-data scales (5 MB, 10 MB, 100 MB, 1000 MB). Top row: matched CKA. Bottom row: shuffled baseline. Off-diagonal matched values rise monotonically with scale, while shuffled values remain near zero throughout.

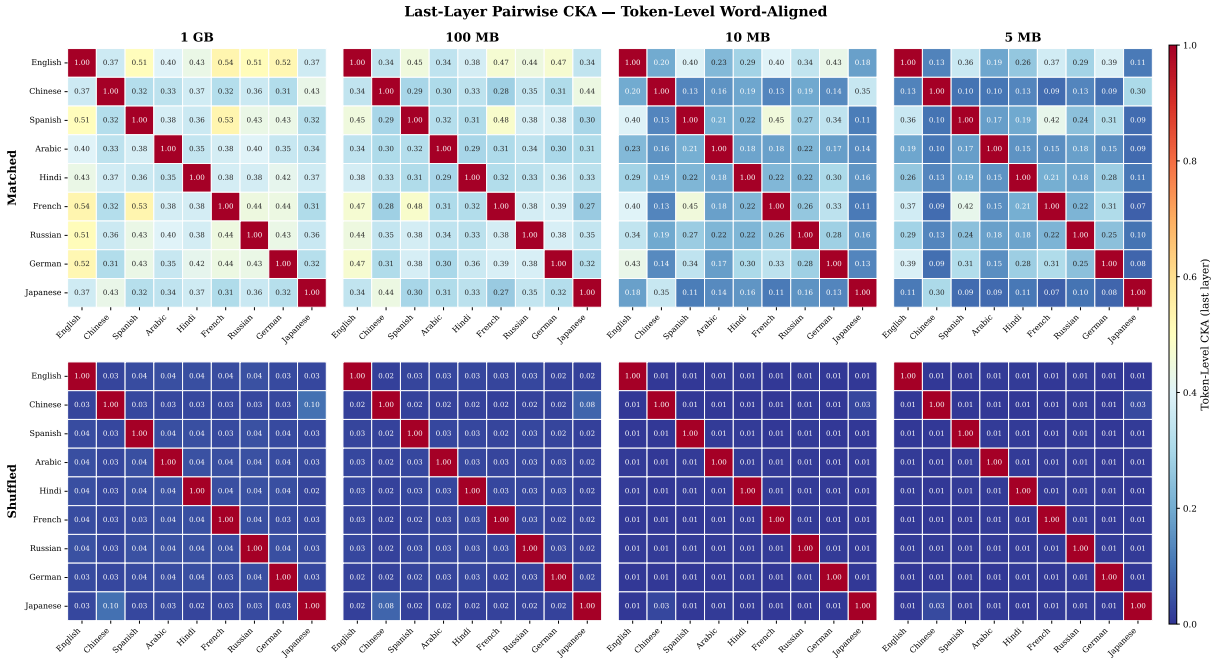

Figure 15: Setting 2 — Token-Level Word-Aligned: same layout as Figure 14. Token-level alignment is lower in absolute terms than sentence-level but shows the same monotonic scaling with training data.

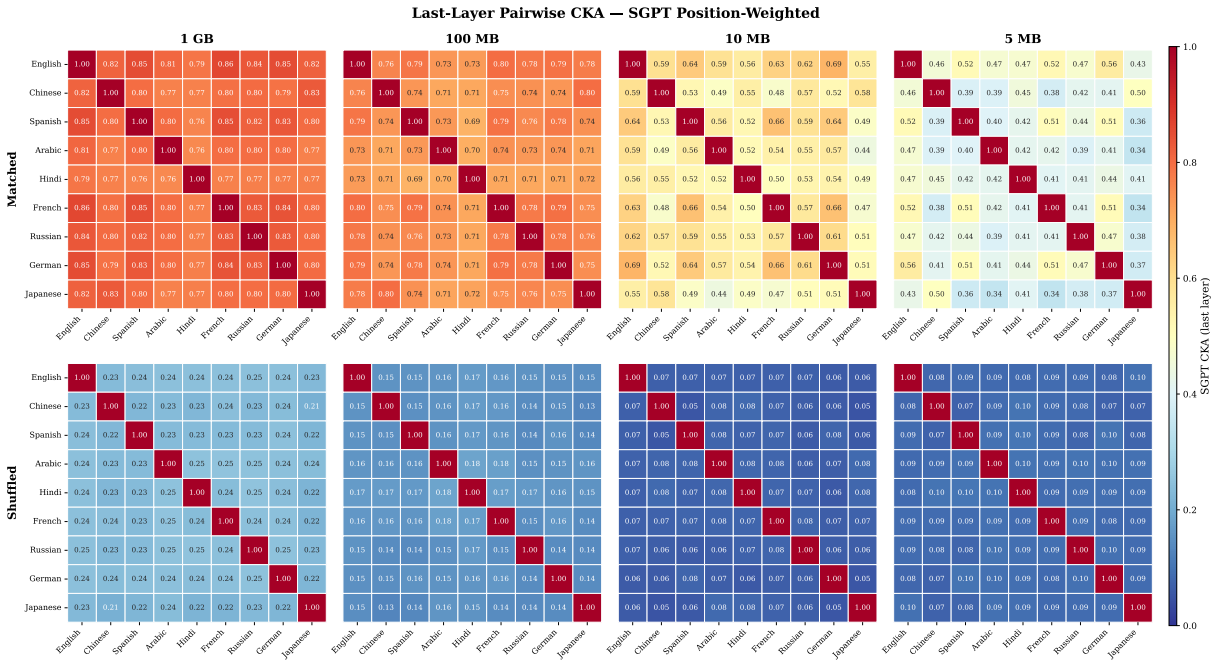

Figure 16: Setting 3 — SGPT Position-Weighted Pooling: same layout as Figure 14. SGPT pooling yields the highest absolute matched values across all scales, while preserving the monotonic size effect.

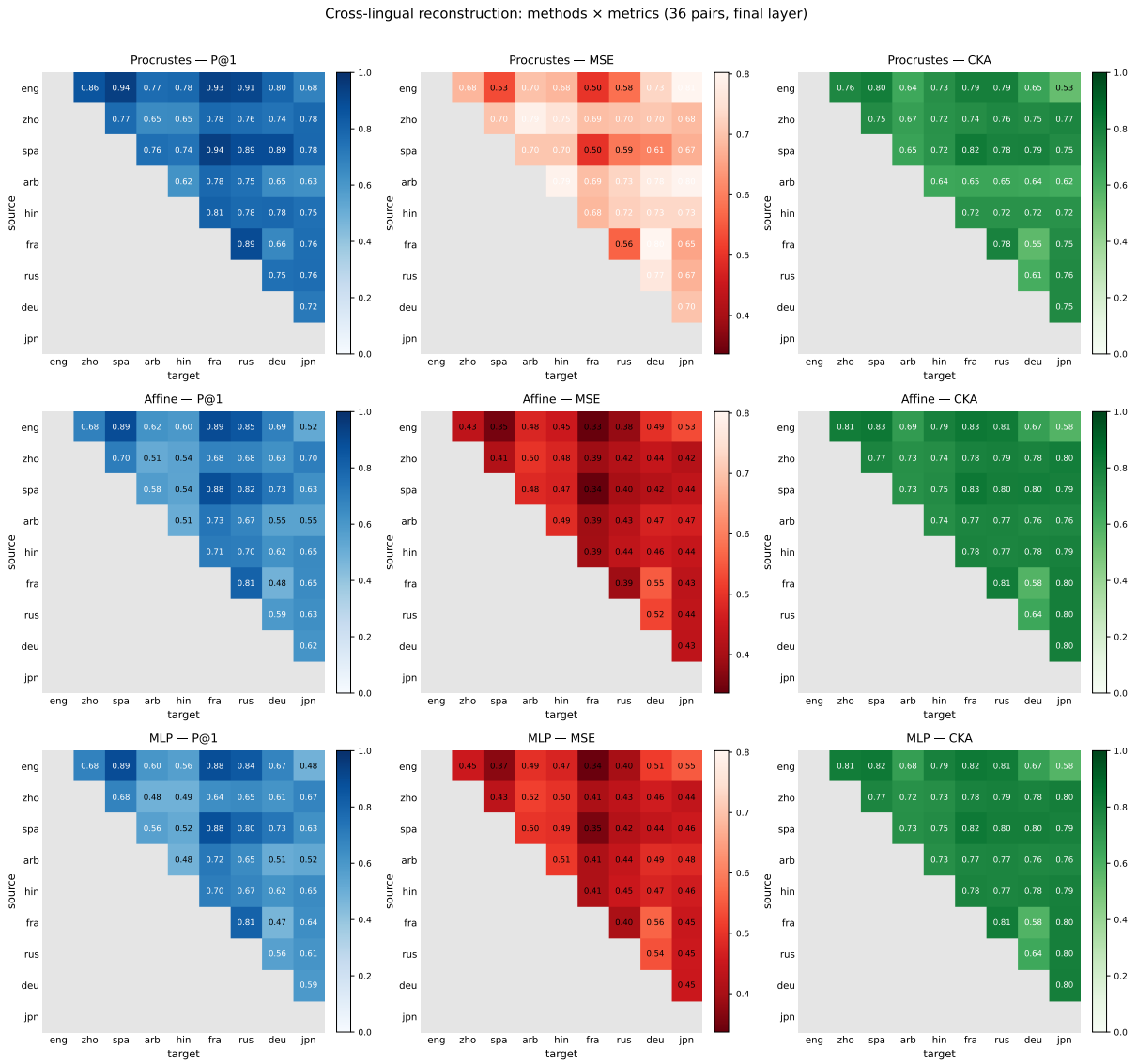

Figure 17: **Method×metric heatmaps across layers and pairs.** Rows: methods (Procrustes, Affine, MLP). Columns: metrics (P@1, MSE, CKA). Each cell aggregates across 36 language pairs by layer.

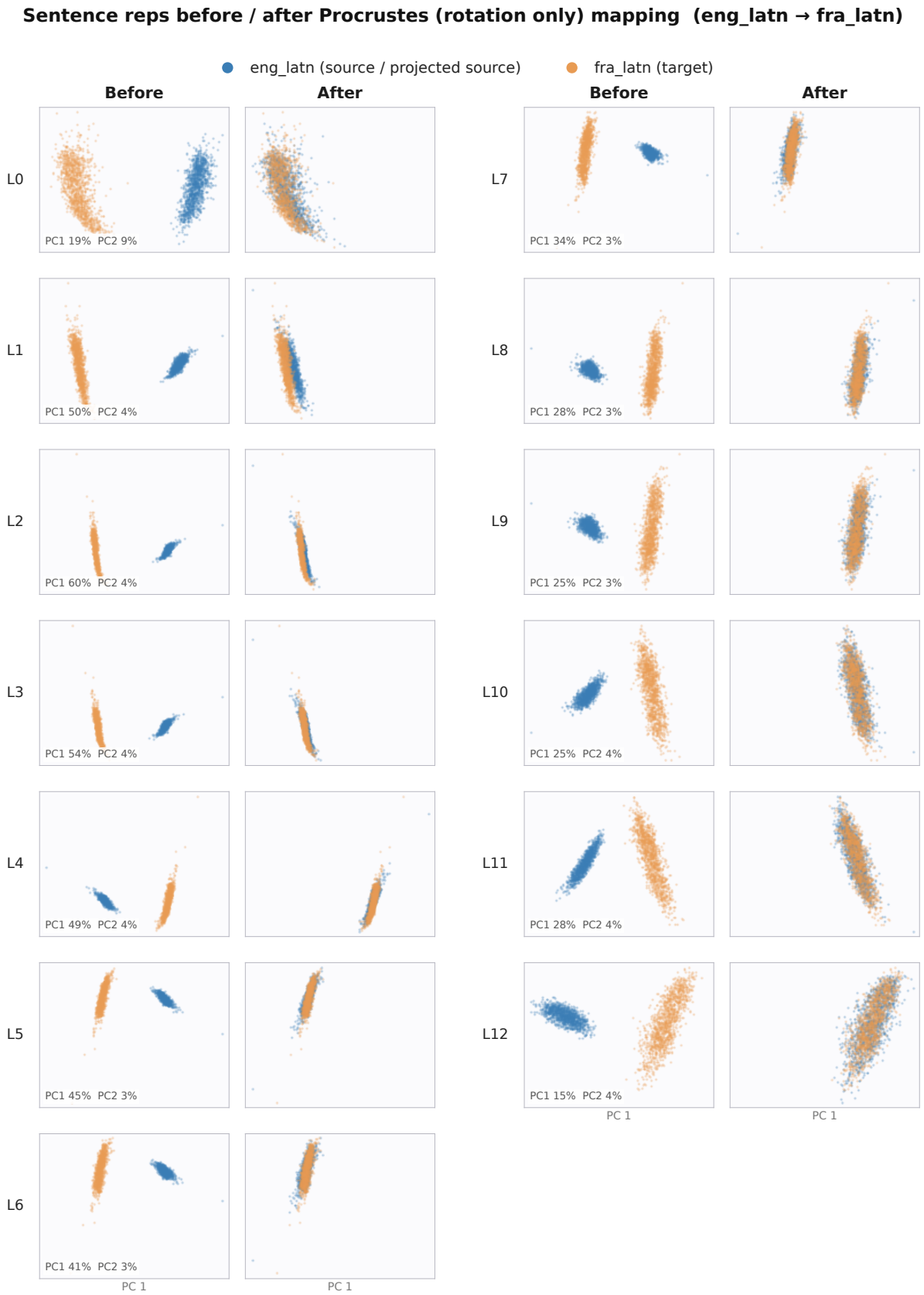

Figure 18: **Procrustes mapping, eng–fra** (peak retrieval layer; 2D PCA over the joint test set). English source (○), French target (●), and Procrustes-mapped English (×). Lines connect each English source to its matched French target.