MAPP: CROSS-LINGUAL TRANSFER THROUGH REPRESENTATION ALIGNMENT WITH PERFORMANCE-GAP PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual representation alignment — training a Large Language Model (LLM) to encode semantically equivalent inputs similarly across languages — is increasingly used to improve LLM performance on low-resource languages, motivated by the observation that a language's performance gap to English correlates with the distance between their hidden states, particularly at middle layers. State-of-the-art alignment methods optimize distance functions or contrastive losses selected *a priori*, all of which are minimized in the limit of complete alignment. We argue that these objectives are only *proxies* for optimal cross-lingual transfer learning by showing that complete alignment both eliminates language identity and yields sub-optimal performance gains. By training oracle models and examining their multilingual representation geometry, we establish the *performance gap* and the *language gap* as separable components of cross-lingual representation difference. We therefore propose MAPP (Multilingual Alignment through Performance-gap Probing), which trains probes on middle-layer hidden representations to predict cross-lingual performance gaps and uses them as reward models to align the LLM. Across three tasks, MAPP yields larger performance gains than state-of-the-art alignment signals while preserving output language identity.

1 INTRODUCTION

LLMs perform unevenly across languages on identical tasks: ARC-Challenge accuracy under Gemma-2-2B falls from 53.7% in English to 26.2% in Marathi (Sin Jing Xuan et al., 2025), and comparable disparities hold across model families and benchmarks (Kargaran et al., 2025); task-specific supervision is also concentrated in a handful of high-resource languages (Grattafiori et al., 2024). Cross-lingual transfer learning is therefore important and practical: it transfers language-specific capabilities the model already exhibits, or behaviors learnable through data exclusive to a certain language, into other languages, using the model itself as the source of supervision rather than collecting annotation per language (Conneau et al., 2020; Hämmerl et al., 2024). Most prior work exploits high-resource-language *outputs*, for instance by translating them into low-resource languages to serve as SFT data (Lai et al., 2023; Li et al., 2023), or adding a KL term that makes equivalent inputs in different languages produce equivalent output distributions (Zheng et al., 2021; Liu et al., 2026).

Beyond outputs, the LLM's internal states can also serve as a source of supervision, building on the assumption that on-task learning in a specific language enforces language-agnostic knowledge inside LLM internal states. This is known as representation alignment. One of the major obstacles to representation alignment is that the **desired performance contrast signal** is always coupled with an **undesired language signal**; moving low-resource language representations toward their high-resource counterparts informs them with knowledge encoded in the high-resource hidden representations, but also makes them more likely to decode tokens in the high-resource language (Deng et al., 2025; Ryu et al., 2026). We show through steering that such a coupled signal not only hurts language identity but also reduces the amount of improvement transferred from the high-resource language (Fig. 3).

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.

We model the representation distance of semantically equivalent inputs in two languages as the sum of two quantities: the *performance gap*, which accounts for the shortfall of the target-language representation from what a competent model would produce, and the *language gap*, which accounts for the residual that does not vanish with competence. We show that the two components are functionally distinct: steering base-model representations along the performance gap raises accuracy from 32.2% to 40.0% with output language unaffected; steering their sum reaches only 35.8% and increases off-language output from 5% to 60%.

We therefore propose MAPP: Multilingual Alignment through Performance-gap Probing, a multilingual transfer learning objective that aims to minimize the performance gap while leaving the language gap unaffected. MAPP involves first training a probe on middle-layer representations to predict the per-sentence performance difference between parallel inputs, measured as the difference in translation log-probability into a shared pivot language, then using the trained probe to supervise the LLM to minimize performance lag of a low-resource language compared to high-resource languages. We provide two variants of the performance-gap probe: a sparse autoencoder fine-tuned so that a subset of its features are language agnostic (Cunningham et al., 2023; Lieberum et al., 2024), and a cross-attention module that scores token-level correspondence between the two sequences. During LLM training, the alignment objective — maximizing probe reward — alternates with next-token prediction on reference-language task data. The probe serves as a reward model at training time rather than an additional component of the LLM, leaving LLM-inference cost unchanged.

We evaluate on three tasks: ARC-Challenge (Clark et al., 2018), MASSIVE slot-filling (FitzGerald et al., 2023), and HellaSwag (Zellers et al., 2019; Lai et al., 2023), each with English alone or six high-resource languages as the supervised languages. We show that MAPP improves transfer to the target languages more than state-of-the-art alignment objectives on all three tested domains, and that MAPP has smaller forgetting in the supervised languages.

2 Related Work

Cross-lingual representation similarity. Multilingual encoders share representations across languages without explicit cross-lingual supervision (Pires et al., 2019). In decoder LLMs, intermediate-layer representations of non-English inputs often resemble those of the dominant pretraining language, typically English (Wendler et al., 2024; Schut et al., 2025; Wu et al., 2025; Zhong et al., 2024), and encode grammatical knowledge (Brinkmann et al., 2025) and concepts (Körner et al., 2026) similarly across languages. Some work also characterizes language-specific information, primarily categorized into **language identity** (which language to decode to) and knowledge accessible in only some languages. Language-sensitive neurons concentrate in early and late layers (Tang et al., 2024; Zhao et al., 2024; Wang et al., 2025), although middle layers also carry language gaps attributed to differences in reasoning capability (Zhao et al., 2025; Lim et al., 2025).

Representation similarity and performance. Cross-lingual representation distance predicts downstream performance: languages closer to **dominant languages** such as English perform better (Kargaran et al., 2025; Li et al., 2025; Al Ali et al., 2026), and at the instance level, weaker alignment to high-resource languages accompanies incorrect predictions (Ravisankar et al., 2026). Related correlations hold for cross-lingual neuron overlap during pretraining (Wang et al., 2024) and SAE activation gaps (Sin Jing Xuan et al., 2025). We show in §3.1 that sentence-level distance correlates more strongly with the performance gap after subtracting a mean *language vector* estimated on a parallel corpus.

Representation alignment. Representation alignment turns similarity into a training signal, training low-resource representations toward those of parallel high-resource sentences, typically with contrastive objectives such as InfoNCE (Hadsell et al., 2006; van den Oord et al., 2018). A general challenge is separating **transferable task information** from **modality-specific variation**. Multimodal work learns separate representations for shared and modality-specific information (Liang et al., 2023); other work shows that when one representation carries both, alignment should upweight samples with the smallest modality gap and down-weight largely modality-specific ones. Multilingual objectives include contrastive losses on parallel hidden states (Li et al., 2024; Liu & Niehues, 2025), penalties on SAE activation differences (Sin Jing Xuan et al., 2025) and soft-token alignment (He et al., 2026). All encourage complete alignment, although representation dis-

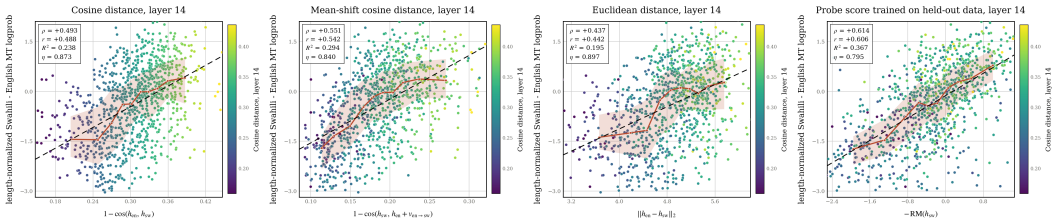

Figure 1: Representation distance is a weak, noisy sentence-level predictor of the performance gap. Each point is one of 1,000 English–Swahili TED2020 pairs; the vertical axis is g (Equation 1) with Chinese as the pivot language. Horizontal axes: *(a)* cosine distance at layer 14; *(b)* cosine distance after subtracting the mean language offset v from the English representation; *(c)* Euclidean distance; *(d)* a trained probe. Red lines are binned means with interquartile bands.

tance carries language difference as well as shared knowledge; InfoNCE alignment collapses the output language (§6.2) and can lower target-language accuracy, whereas probing and closing only performance-related differences gives larger gains (§6.1).

3 Motivating a Performance-driven Alignment Signal

This section motivates an alignment signal that separates performance from language differences. Representation distance predicts performance across languages but only weakly across sentences within a language (§3.1). We train an **oracle model** with enough target-language task SFT to perform as well as in English, and a **base model** that sees only a subset of that data and so underperforms on test items that need unseen knowledge. The **performance gap** Δ_{perf} is the base–oracle representation difference for the same query; the **language gap** Δ_{lang} is the oracle's representation difference for the same query in two languages (§3.2). Δ_{perf} is largest relative to Δ_{lang} in the middle layers (Fig. 2), which prior work finds most language agnostic and task relevant (Wendler et al., 2024; Wu et al., 2025; Zhao et al., 2024). Steering only Δ_{perf} improves accuracy most, while objectives that close both gaps lose language identity and gain less (§3.3).

3.1 Representation distance is a weak per-sentence predictor of performance

We analyze the correlation between representation distance and performance across languages on machine translation tasks, where performance for different input languages is directly comparable through the log probability of the same ground-truth completion token sequence, namely the human-annotated translation in the shared output language. Given a target-language sentence x_t, its reference-language translation x_r and their translation y_o into a shared output (pivot) language, we force-decode y_o from each source and compare length-normalized log-probabilities $\ell(y \mid x) = \log p(y \mid x)/|y|$:

$$g(t, r; o) = \frac{\ell(y_o \mid x_r) - \ell(y_o \mid x_t)}{\ell(y_o \mid x_r)}. \tag{1}$$

We present the correlation of several distance metrics with the performance gap g. Cosine and Euclidean distance are measured between $h_r^{(i)}$ and $h_t^{(i)}$, the representations of x_r and x_t at layer 14, mean-pooled over tokens in the input sentence. Following mean-shift alignment of language subspaces (Xiao et al., 2025) and evidence that languages differ largely in their subspace means (Chang et al., 2022; Xu et al., 2023), we estimate a language gap by computing the mean displacement between the two languages across 1,000 held-out sentence pairs, namely $v = \frac{1}{N}\sum_i (h_r^{(i)} - h_t^{(i)})$ with $N = 1000$. Subtracting it from the hidden representations of the reference sentences before computing cosine distance yields the mean-shifted cosine distance (Fig. 1) that accounts for language difference.

Cosine distance reaches only $\rho = 0.493$ ($R^2 = 0.238$; Fig. 1a) and Euclidean distance $\rho = 0.437$ ($R^2 = 0.195$; Fig. 1c). Mean-shifting raises ρ to 0.551 (Fig. 1b), so much of this noise is language

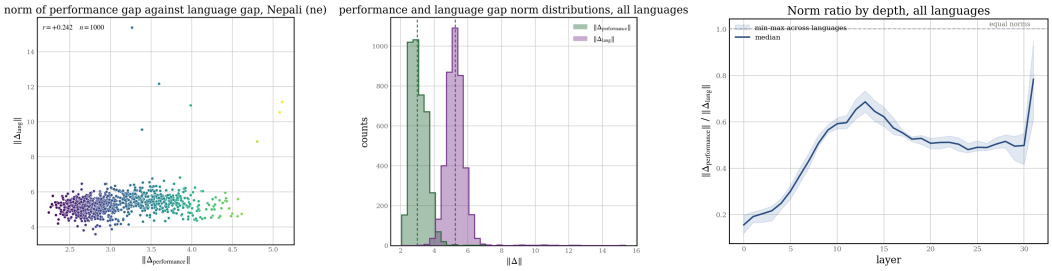

Figure 2: The language gap persists at performance parity. **Left:** per-sentence $\|\Delta_{\text{perf}}\|$ against $\|\Delta_{\text{lang}}\|$ at layer 14, pooled over four target languages ($n = 4{,}000$). **Middle:** marginal distributions; dashed lines mark medians. **Right:** median $\|\Delta_{\text{perf}}\|/\|\Delta_{\text{lang}}\|$ by layer, pooled over the same languages; shading spans the per-language minimum and maximum.

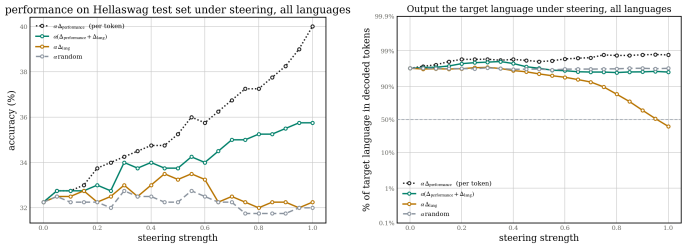

Figure 3: Closing the performance gap improves accuracy while preserving the output language. We subtract α times each gap, unnormalized, from base-model layer-14 activations at target-language prompt tokens. *(Left)* task accuracy; *(Right)* share of target-language output tokens.

difference. A trained probe (§4.2) does better still, with held-out $\rho = 0.614$ (Fig. 1d), suggesting that a learned readout can separate the performance signal that raw distances mix with language.

3.2 Disentangling language and performance gaps with oracle models

For a parallel pair (x_r, x_t), let h_r^*, h_t^* and $h_r^{\text{base}}, h_t^{\text{base}}$ be the mean-pooled oracle (§3) and base-model representations. With English as reference and Swahili, Azerbaijani, Icelandic and Tagalog as targets, and assuming a negligible English gap ($h_r^{\text{base}} \approx h_r^*$), the cross-lingual displacement decomposes as

$$h_t^{\text{base}} - h_r^{\text{base}} = \underbrace{h_t^{\text{base}} - h_t^*}_{\Delta_{\text{perf}}} + \underbrace{h_t^* - h_r^*}_{\Delta_{\text{lang}}} + \underbrace{h_r^* - h_r^{\text{base}}}_{\approx 0} = \Delta_{\text{perf}} + \Delta_{\text{lang}}. \tag{2}$$

Δ_{perf} is the target-language performance deficit and Δ_{lang} the language difference that remains at matched performance. At layer 14, the language gap has the larger median norm ($\|\Delta_{\text{lang}}\| = 5.2$ vs. $\|\Delta_{\text{perf}}\| = 3.0$; Fig. 2, left and middle). Middle-layer representation differences, which prior alignment work treats as performance signal, are thus dominated by language, and minimizing the full distance also closes a larger language component that ought to be kept.

The median ratio $\|\Delta_{\text{perf}}\|/\|\Delta_{\text{lang}}\|$ peaks near layer 13 (Fig. 2, right), consistent with prior middle-layer alignment (Liu & Niehues, 2025) and motivating our use of layer 14. Even there, the language gap is about 1.4 times the performance gap, so choosing a middle layer reduces but does not remove the language component of a distance-based signal.

3.3 Steering to close the performance and language gaps

We steer αd, $d \in \{\Delta_{\text{perf}}, \Delta_{\text{lang}}, \Delta_{\text{perf}} + \Delta_{\text{lang}}\}$, into the base model's layer-14 activations over the target-language input span (Fig. 3). Closing Δ_{perf} raises accuracy monotonically from 32.2% to 40.0% while target-language output stays flat or rises slightly. Closing Δ_{lang} gives only a transient gain, gone by $\alpha = 1$, and cuts target-language output from 95% to 40%; a norm-matched random

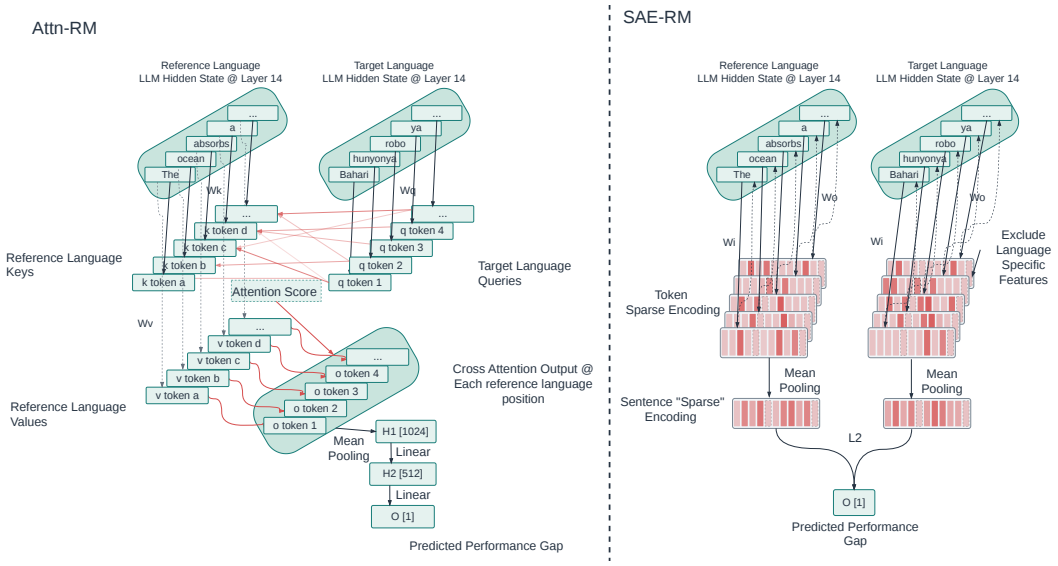

Figure 4: MAPP fits a probe on translation-based performance measurements, then uses it to supervise alignment. *(Left)* A reference sentence and its target-language translation each force-decode a shared pivot-language translation under the frozen LLM. Their length-normalized log-probabilities provide probe supervision. *(Right)* LLM training alternates between alignment on bare parallel sentences using the frozen probe and task fine-tuning on reference-language data.

direction matches its accuracy without disrupting the output language. Closing both gaps reaches only 35.75%, less than half the gain from Δ_{perf} alone, and also reduces target-language output. We therefore want a signal that closes Δ_{perf} while preserving Δ_{lang}.

4 MAPP: Multilingual Alignment through Performance-gap Probing

The results in §3 motivate an alignment objective that closes Δ_{perf} while preserving Δ_{lang}. MAPP approximates this objective with a probe trained on translation-based performance measurements; the probe then rewards input pairs with low expected performance difference. We consider two architectures for the probe: a cross-attention probe trained from scratch and a sparse-autoencoder (SAE) probe fine-tuned such that sparse encoding distance of parallel inputs predicts performance difference. We then alternate reference-language task fine-tuning with alignment training that minimizes the probe's predicted performance difference (Fig. 4). Gradients pass through the frozen probe into the LLM; the probe is not needed for inference with the LLM, leaving inference cost unchanged.

4.1 Training the probe on a parallel MT task

The translation-based performance gap measurement in Equation 1 requires three-way parallel data: a reference sentence x_r, its target-language translation x_t, and a shared output language translation y_o. Aligning the LLM hidden representations with the performance gap directly would limit the multilingual alignment to domains where such triples are available, which are very limited in size and representativeness. We thus fit a probe to predict the performance gap using only the two input sequences' representations, namely H_r and H_t, without access to the output language. We can then use it as a proxy for performance differences to allow alignment training with only bitext. As a general training objective for the probe, we minimize

$$\mathcal{L}_{\text{probe}}(\phi) = \mathbb{E}_{(x_r,\, x_t,\, y_o)}\Big[\,\text{Huber}\big(f_\phi(H_r, H_t),\ g(t, r; o)\big)\Big], \tag{3}$$

where $f_\phi(H_r, H_t)$ denotes the probe's prediction and $g(t, r; o)$ denotes the ground-truth performance gap. The Huber loss ensures that the heavy tail of g, produced by near-zero log-probabilities

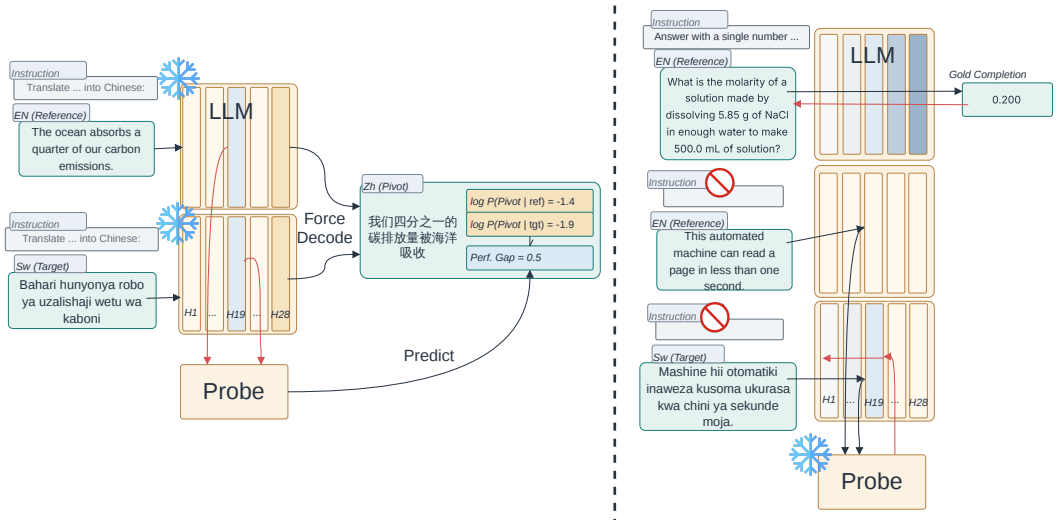

Figure 5: Both probes score a parallel pair from its hidden states alone. *(Left)* Cross-attention probe: target tokens attend to reference tokens; comparisons in a shared value space are pooled into a score regressed onto g. *(Right)* SAE probe: the score is the negative distance between mean-pooled sparse codes after masking language-specific features. Translation performance weights the alignment objective used to fine-tune the SAE.

in its denominator, does not dominate the fit. We also provide a variant that is based on fine-tuning sparse autoencoders in §4.2 that takes a different training objective adapted from traditional sparse autoencoder training schemes.

The probe is trained to predict the performance gap from just the input representation pair, thus learning a **representation relationship** among multilingual parallel inputs that signifies performance. The probe's output is language agnostic by design: when a single probe is trained on a mixture of language triples, it needs to read the performance gap out of the representation pairs regardless of the input language identity. The probe is then used to provide supervision for the LLM when prompted with the same question in a high-resource language and a low-resource language, where the probe's output is a predicted score for the LLM on the low-resource question, which is useful when ground-truth labels do not exist for direct SFT.

4.2 Probe Architectures

Let $H_r = (h_{r,j})_{j=1}^{n_r}$ and $H_t = (h_{t,j})_{j=1}^{n_t}$ denote the hidden states of the tokens of the reference sentence x_r and its target-language translation x_t, respectively. Both probes map these two sequences of hidden states to a scalar $f_\phi(H_r, H_t)$ as a prediction of $g(t, r; o)$. They differ in the underlying architecture:

Cross-attention probe. The cross-attention module attn_ϕ uses target tokens as queries and reference tokens as keys and values. Its value projection V_ϕ is also applied to target tokens, placing each target token and its attended reference content in the same comparison space:

$$A = \mathrm{attn}_\phi(H_t, H_r), \quad \hat{g} = f_\phi(H_r, H_t) = \rho_\phi\left(\frac{1}{|\mathcal{I}_t|}\sum_{j\in\mathcal{I}_t}\psi_\phi(A_j)\right), \tag{4}$$

where A_j, the attention output for target token j, is passed through a small feed-forward network ψ_ϕ, mean-pooled over target tokens, and mapped by a second network ρ_ϕ to a scalar predicted performance gap. The set $\mathcal{I}_t$ excludes target padding and BOS positions. The target influences the prediction through its attention queries, without an explicit comparison between target and reference value vectors. Applying ψ_ϕ before pooling lets the probe extract features from each target-conditioned readout before aggregating across the sentence. We train $\hat{g}$ against the measured performance gap

using Equation 3. The uniform-attention ablation in Table 5 tests the contribution of learned reference selection.

Sparse-autoencoder probe. The probe scores a parallel pair by the negative distance between its pooled sparse codes, excluding language-specific features: $f_\phi(H_r, H_t) = -\|m \odot (c_r - c_t)\|_2$. Here c_r and c_t are the SAE codes averaged over the content tokens of the reference and target sentences, respectively, and m is a binary mask that zeroes language-specific features (Appendix A). We initialize from Goodfire's pretrained SAE for layer 19 of Llama-3.1-8B-Instruct (Goodfire, 2025) and fine-tune it on parallel representations from the frozen SFT model with an additional multilingual alignment term in the loss function:

$$\begin{aligned}\mathcal{L}_{\text{SAE}} = \mathcal{L}_{\text{rec}} + \lambda_0 \mathcal{L}_0 + \lambda_1 \mathcal{L}_1 \\ + \lambda_{\text{align}}\, \mathbb{E}_i\Big[w_i \|m \odot (c_r^{(i)} - c_t^{(i)})\|_2\Big] + \lambda_{\text{anc}}\, \mathbb{E}_i\Big[\|c_r^{(i)} - \bar{c}_r^{(i)}\|_2^2\Big].\end{aligned} \tag{5}$$

Here $\mathcal{L}_{\text{rec}}$ is the token-level reconstruction error, while $\mathcal{L}_0$ and $\mathcal{L}_1$ penalize the number and decoder-norm-weighted magnitude of active features. The alignment term penalizes the SAE for producing different encodings for semantically equivalent inputs across languages, through measuring the L_2 distance between sparse encodings for x_r and x_t. The anchor loss keeps reference codes near their pretrained values $\bar{c}_r$, encouraging target codes to move toward them.

Note the weight w_i applied to the alignment term. It accounts for the fact that the sparse autoencoder is fine-tuned with an **imperfect LLM**, whose representation pairs H_r, H_t sometimes still have large Δ_{perf}. Ideally, the SAE should be fine-tuned on representation pairs that only differ in language, thus learning to encode $c_r^* = c_t^*$. In practice the representations have non-negligible Δ_{perf}, however, and the SAE would instead learn to encode $c_r^* = c_t^{\text{base}}$, erasing the useful performance signal that the sparse encoding distance is supposed to carry. We address this with sample reweighting, an established approach to learning from noisy or unreliable supervision (Ren et al., 2018; Jiang et al., 2018). We define the weight for pair i as

$$w_i = \tfrac{1}{Z} \exp\big(\min(g_i, 0)\big), \tag{6}$$

where g_i is the performance gap (Equation 1) of pair i and Z normalizes the weights to mean one. This weighting upweights pairs with smaller Δ_{perf}, and downweights pairs where Δ_{perf} dominates, therefore training the SAE to be **language agnostic** but not **performance agnostic**. The fine-tuned SAE thus provides signal for performance difference through its sparse encoding distance.

4.3 Alternating optimization

After fitting the probe, MAPP alternates reference-language task fine-tuning with alignment on parallel sentences (Fig. 4, right). Writing H_r^θ and H_t^θ for the parallel representations under the current LLM parameters θ, and θ_0 for the pre-alignment anchor parameters, the alignment objective minimizes

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(x_r, x_t)}\Big[\max\big(0, -f_\phi(\text{sg}[H_r^\theta], H_t^\theta)\big) + \lambda_{\text{KL}} \frac{1}{|\mathcal{C}_t|} \sum_{i \in \mathcal{C}_t} \text{KL}\Big(p_\theta(\cdot \mid x_t^{<i}) \,\Big\|\, p_{\theta_0}(\cdot \mid x_t^{<i})\Big)\Big], \tag{7}$$

where $\mathcal{C}_t$ are the token positions of the target sentence. We back-propagate through the target-language pass only, treating the reference pass as fixed within each update. Alignment gradients pass through the frozen probe and target-language representations, reaching only the layers up to the probe's layer; the KL penalty and task fine-tuning also train the layers above it.

5 Experimental Setup

5.1 Datasets

We evaluate on slot filling (MASSIVE; FitzGerald et al., 2023), commonsense sentence completion (HellaSwag; Zellers et al., 2019) and science question answering (ARC-Challenge; Clark et al., 2018). MASSIVE is scored by span-level slot F1 under greedy decoding, ARC-C and HellaSwag by accuracy, ranking candidates by length-normalized log-likelihood. The six high-resource supervised

Table 1: MAPP yields larger performance transfer from reference language to target language across three distinct benchmarks. When the English-SFT model is aligned to English, both MAPP probes outperform both distance objectives on all three tasks, with gains over SFT of up to 7.0 points. With alignment to a mixture of six high-resource languages (§5), MAPP again outperforms both baselines on every task, with gains over SFT of up to 3.8 points. Columns: *en*, English; *6-high*, mean over the six high-resource supervised languages (§5); *tr*, mean over the target languages. Scores are accuracy (ARC-C, HellaSwag) and slot F1 (MASSIVE); bold marks the best *tr* score within rows 3–7 and within rows 8–12.

		ARC-C			MASSIVE			HellaSwag		
	Setting	en	6-high	tr	en	6-high	tr	en	6-high	tr
1	Base model	40.7	34.8	30.8	–	–	–	77.3	60.7	49.2
2	Low-resource target lang SFT	64.0	52.6	42.6	78.8	68.4	75.8	–	–	–
3	English-only SFT	61.3	49.3	36.6	83.6	62.7	40.8	81.7	61.6	44.5
4	3 + English InfoNCE alignment	55.3	43.1	32.4	81.3	60.0	44.7	78.7	59.1	45.6
5	3 + English SAE-L_2 alignment	53.3	38.2	33.3	84.5	61.2	43.4	81.7	63.1	50.8
6	3 + En MAPP-SAE	57.7	47.3	39.1	84.5	64.9	**46.6**	81.6	62.7	**51.5**
7	3 + En MAPP-attn	59.3	47.4	**39.2**	83.5	64.2	45.9	81.7	62.3	51.0
8	6 high-resource lang SFT	65.7	53.2	38.6	85.2	75.6	53.7	79.3	64.6	46.3
9	8 + 6-lang InfoNCE alignment	53.7	51.5	41.0	82.2	73.3	54.9	74.1	63.2	46.1
10	8 + 6-lang SAE-L_2 alignment	59.0	50.3	39.5	84.6	74.6	55.4	78.5	64.1	47.7
11	8 + 6-lang MAPP-SAE	62.0	52.4	**42.4**	83.9	76.3	55.8	78.9	64.1	48.8
12	8 + 6-lang MAPP-attn	59.3	51.3	41.9	84.6	75.6	**57.0**	78.8	64.9	**48.9**

languages are English, German, Spanish, French, Russian and Chinese. Probes are trained on parallel triples from the FLORES-200 development split (NLLB Team, 2024) and TED2020 (Reimers & Gurevych, 2020), with distinct reference and output languages drawn from these six. Targets are Azerbaijani, Icelandic, Swahili and Tagalog for MASSIVE, and Swedish, Danish, Hungarian, Romanian, Ukrainian, Slovak, Catalan, Bengali and Tamil for ARC-C and HellaSwag. Alignment uses English-centric Tatoeba bitext (Tiedemann, 2020); for six-language supervision, both target and non-English reference languages are aligned to English.

5.2 Baselines

Besides reference-language SFT, we compare two distance objectives. InfoNCE (Liu & Niehues, 2025; van den Oord et al., 2018) contrasts mean-pooled layer-16 target representations with their English translations against English negatives. SAE-L_2, adapted from Sin Jing Xuan et al. (2025), minimizes the distance between mean-pooled codes of Goodfire's pretrained layer-19 SAE under the same language-feature mask as MAPP. Since alignment assumes no target-language task data, models trained on such data (available for MASSIVE and ARC-C) serve only as *ceilings*.

5.3 MAPP training

All models are Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with LoRA (Hu et al., 2022); alignment starts from each setting's SFT checkpoint (Table 1). The probes read layer 16 (attention) and layer 19 (SAE; Goodfire, 2025); hyperparameters are in Appendix A.

6 Results and Analysis

6.1 MAPP yields stronger performance transfer to low-resource languages without task data

Across the three tested domains, MAPP brings a larger performance gain compared to state-of-the-art alignment objectives, such as InfoNCE and L_2 distance in the sparse encoding space. In the setting where the LLM receives English-only supervision and alignment between English and target languages, the best MAPP variant improves over SFT by 2.6 accuracy points on ARC-C, 5.8

Table 2: MAPP yields considerably higher target-language generation percentage after alignment, demonstrating that multilingual alignment with MAPP preserves language identity. MAPP-attn retains 93.5% and 94.4% target-language tokens on ARC-C and HellaSwag, compared with 22.3% and 26.6% for InfoNCE, reducing off-target generation by over 90% on both tasks. Bold marks the best score in each column within each task.

Method	sv	da	hu	ro	uk	sk	ca	bn	ta	Mean
ARC-C										
InfoNCE	24.7	29.3	17.6	30.6	5.6	17.5	37.4	7.9	29.9	22.3
SAE-L_2	73.5	84.5	48.1	92.2	89.3	92.5	73.6	**100.0**	**100.0**	83.7
MAPP-SAE	73.2	75.9	62.9	87.0	69.4	85.6	78.9	66.2	31.0	70.0
MAPP-attn	**91.7**	**94.8**	**79.8**	**94.2**	**95.0**	**94.4**	**91.9**	**100.0**	**100.0**	**93.5**
HellaSwag										
InfoNCE	21.1	21.7	30.4	45.6	31.0	19.6	33.7	32.6	4.0	26.6
SAE-L_2	95.3	95.3	76.7	**94.7**	96.6	92.9	95.3	**100.0**	**100.0**	94.1
MAPP-SAE	95.2	95.1	75.2	93.7	96.9	92.6	**95.4**	**100.0**	**100.0**	93.8
MAPP-attn	**95.4**	**96.0**	**77.4**	93.4	**98.7**	**93.1**	**95.4**	**100.0**	**100.0**	**94.4**

F1 on MASSIVE and 7.0 accuracy points on HellaSwag. With six high-resource languages, the corresponding gains are 3.8, 3.3 and 2.6 points. The large margin that MAPP brings demonstrates that MAPP turns parallel text into a stronger transfer signal than pure representation distance.

ARC-C shows that the alignment objective can determine whether transfer improves at all. Starting from English-only SFT at 36.6%, InfoNCE and SAE-L_2 reduce target-language accuracy to 32.4% and 33.3%; MAPP-SAE and MAPP-attn raise it to 39.1% and 39.2%. The same parallel data supports positive transfer when alignment targets predicted performance differences.

6.2 MAPP PRESERVES LANGUAGE IDENTITY

MAPP-attn avoids the output-language collapse induced by InfoNCE (Table 2). It retains 93.5% of decoded tokens in the target language on ARC-C and 94.4% on HellaSwag, against 22.3% and 26.6% for InfoNCE. This cuts off-target generation by over 90% on both tasks. On ARC-C, MAPP-attn also matches or exceeds SAE-L_2 in every evaluated language, raising mean retention from 83.7% to 93.5%. The SAE probe preserves less language identity: its retention reaches 70.0% on ARC-C and 93.8% on HellaSwag, above InfoNCE but below SAE-L_2 (83.7% and 94.1%). MAPP-attn therefore delivers the strongest combination of performance transfer and language preservation. Appendix D further shows that MAPP moves target-language representations more like direct target-language supervision than either distance objective does (oracle travel 0.57–0.61 vs. 0.36–0.49 with six languages).

7 CONCLUSION

Existing cross-lingual representation alignment has treated the entire distance between a target-language representation and its English counterpart as the quantity to close. We showed that this distance is a weak per-sentence predictor of performance and that it combines a performance gap, which alignment training should close, with a language gap that persists even in a model equally competent in both languages; steering and training along the combined distance forgo part of the available gain and erode output language identity. MAPP instead learns to minimize just the performance gap: a probe fit to a translation-based measure of the performance gap supervises the LLM in alternation with task fine-tuning, and the probe is discarded at inference. Across ARC-Challenge, MASSIVE and HellaSwag, with English or six high-resource languages as supervision, MAPP transfers better than InfoNCE and sparse-autoencoder L_2 alignment while preserving supervised performance and language identity. Our evidence comes from a single model family, estimates the performance gap through translation rather than the downstream task, relies on an oracle model for the decomposition, and requires fitting a probe and alternating two training phases; extending the probe to other models and to task-specific competence signals is a natural next step.

AI USE STATEMENT

We used AI tools to assist with code implementation, identifying related work, and polishing the language of author-written text. AI tools were not used to generate or synthesize research ideas or to draft manuscript sections. The authors retain full responsibility for the implementation, references, and content of the manuscript.

REPRODUCIBILITY STATEMENT

All base models, pretrained components and datasets we build on are public. Section 5 specifies the base model, tasks and metrics, the corpora and languages used for probe training and alignment, and the baselines. The performance gap that supervises the probe is defined in Equation 1, and §4 gives the probe objective, both probe architectures and the alternating alignment objective (Equations 3–7). Appendix A (Table 3) lists all hyperparameters, training schedules, data sizes, prompt formats, the language-feature mask procedure, decoding settings and random seeds, and Appendices B, C and D detail the oracle parity test, the probe ablations and the representation-travel analysis.

REFERENCES

Adnan Al Ali, Kathy Hämmerl, Jindřich Libovický, and Alexander Fraser. Predicting multilingual classification and translation performance of LLMs with cross-lingual alignment – is English enough? *arXiv preprint arXiv:2608.03446*, 2026.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 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*, pp. 6131–6150, 2025.

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*, pp. 119–136, 2022.

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

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

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 4563–4608, 2025.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 4277–4302, 2023.

Goodfire. Llama-3.1-8B-Instruct-SAE-l19, 2025. Sparse autoencoder for layer 19 of Llama-3.1-8B-Instruct. `https://huggingface.co/Goodfire/Llama-3.1-8B-Instruct-SAE-l19`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1735–1742, 2006.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. Understanding cross-lingual Alignment—a survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10922–10943, 2024.

Jiayi He, Jungsoo Park, Wei Xu, and Alan Ritter. Soft token alignment for cross-lingual reasoning. *arXiv preprint arXiv:2606.26466*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2304–2313, 2018.

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*, pp. 27001–27023, 2025.

Felicia Körner, Max Müller-Eberstein, Anna Korhonen, and Barbara Plank. 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*, pp. 3149–3169, 2026.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 318–327, 2023.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8058–8076, 2024.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*, 2023.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Language ranker: A metric for quantifying LLM performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28186–28194, 2025.

Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, 2023.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*, 2025.

Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 15979–15996, 2025.

Yihong Liu, Raoyuan Zhao, Michael A. Hedderich, and Hinrich Schütze. Crosslingual on-policy self-distillation for multilingual reasoning. *arXiv preprint arXiv:2605.09548*, 2026.

NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024.

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

Kartik Ravisankar, HyoJung Han, Sarah Wiegreffe, and Marine Carpuat. Can you map it to English? the role of cross-lingual alignment in the multilingual performance of LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 4854–4872, 2026.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4512–4525, 2020.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4334–4343, 2018.

Sangwon Ryu, Yihong Liu, Mingyang Wang, Yunsu Kim, Jungseul Ok, Gary Geunbae Lee, and Hinrich Schütze. Understanding LLM behavior in multi-target cross-lingual summarization. *arXiv preprint arXiv:2606.01252*, 2026.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual LLMs think in English? *arXiv preprint arXiv:2502.15603*, 2025.

Richmond Sin Jing Xuan, Jalil Huseynov, and Yang Zhang. Uncovering cross-linguistic disparities in LLMs using sparse autoencoders. *arXiv preprint arXiv:2507.18918*, 2025.

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*, pp. 5701–5715, 2024.

Jörg Tiedemann. The Tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 1174–1182, 2020.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12159–12173, 2024.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 5075–5094, 2025.

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*, pp. 15366–15394, 2024.

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 (ICLR)*, 2025.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Mahani Aljunied, Lidong Bing, Noura Al Moubayed, and Yu Rong. Analyzing LLMs' knowledge boundary cognition across languages through the lens of internal representations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 24099–24115, 2025.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3692–3702, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. When less language is more: Language-reasoning disentanglement makes LLMs better multilingual reasoners. In *Advances in Neural Information Processing Systems*, 2025.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, 2024.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3403–3417, 2021.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond English-centric LLMs: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*, 2024.

A Hyperparameters

Table 3 lists the training settings. Initial SFT uses 11,514 MASSIVE utterances per supervised language, 1,081–1,116 translated ARC-C questions per supervised language, or 39,905 translated HellaSwag items. The target-language SFT ceilings use 46,056 MASSIVE utterances and 9,899 ARC-C questions. Initial SFT checkpoints are selected on supervised-language development data; subsequent alignment checkpoints are evaluated after every epoch and selected as described in §5.

MASSIVE prompts use the Llama-3.1 chat template with a fixed system instruction, and outputs list entities as "slot: value". ARC-C and HellaSwag use the question or context alone, with each candidate appended as a continuation. Alignment uses bare sentence pairs.

Each MAPP round contains one alignment epoch followed by two task epochs. An alignment epoch is a fixed budget of 1,000 micro-batches, with one optimizer update per micro-batch. SAE-L_2 follows this schedule. InfoNCE instead interleaves 1,000 contrastive and 1,000 task micro-batches per epoch. Its contrastive batches provide seven in-batch negatives per example, supplemented by negatives queued across gradient-accumulation steps.

The feature mask is shared by SAE-L_2 and the MAPP SAE probe. For MASSIVE, it removes 2,463 of 65,536 features in the English-only setting and 1,978 in the six-language setting.

B Oracle Parity

We compare the oracle's per-item correctness in English and in each target language on the same 1,000 HellaSwag items per language (Table 4). The oracle answers 99.6% of items correctly in

Table 3: Hyperparameters shared across tasks unless otherwise specified.

Stage	Setting	Value
Model	base model	Llama-3.1-8B-Instruct, bf16 weights
	LoRA	$r = 8$, $\alpha = 16$, dropout 0.1; all attention and MLP projections
	maximum length	1,024 tokens
SFT	learning rate	5×10^{-4}; inverse square-root schedule; 500 warmup steps
	batch size	4 micro-batch, 32 accumulation steps; effective batch 128
	epochs	At most 30; early-stopping patience 5
Alternation	rounds	5; one alignment epoch and two task epochs per round
	alignment batch	4 sentence pairs per update; 1,000 updates per epoch
	task batch	Effective batch 128
	learning rate	5×10^{-6}, constant, for both phases
	KL coefficient	0.05, toward the SFT checkpoint during alignment
	early stopping	Patience 20 epoch evaluations
InfoNCE	layer, temperature	16, 0.1
	contrastive batch	8; additional negatives queued across accumulation steps
	pooling	Mean over all positions
	schedule	5 epochs; 1,000 contrastive and 1,000 task micro-batches per epoch
	learning rate	5×10^{-6}, constant
Attention probe	layer	16
	attention	Width 2,048; 8 heads; QK-normalized softmax
	token network, head	MLP 2,048–1,024–1,024; MLP 1,024–512–256–1
	dropout	0.4
	loss	Huber regression onto pivot-averaged g (Equation 1)
	learning rate	3×10^{-5}
	training	At most 30 epochs; token network frozen for the first 5; selection by development Pearson r
SAE probe	autoencoder	Goodfire layer 19; 65,536 features; JumpReLU bandwidth 0.01
	fine-tuning	2,000 steps; 16 pairs per step; learning rate 6×10^{-6}
	loss weights	$\lambda_0 = 1$, $\lambda_1 = 1$, $\lambda_{\text{align}} = 1000$, $\lambda_{\text{anc}} = 50$
	pair weight	$\exp(\min(g, 0))$, normalized to mean 1
	MASSIVE pairs	15,985 (English-only); 71,782 (six languages)
Feature mask	test	Per-feature permutation test; 20,000 permutations; Benjamini–Hochberg $q = 0.05$
	data	More competently translated half of 300 Tatoeba development sentences per language
Evaluation	MASSIVE decoding	Greedy; at most 32 tokens
	seeds	Evaluation subsets: 0; training: 1

English and 97.8–99.4% in the target languages. An exact McNemar test on the discordant pairs finds no significant difference for Icelandic ($p = 0.75$), but rejects equality for Basque, Marathi and Nepali ($p \leq 0.002$), where the oracle is 1.6–1.8 points less accurate than in English. Because a significance test at this sample size rejects even sub-two-point differences, and a non-rejection is not evidence of equality, we also test equivalence directly: pooled over the four languages, the gap between English and target-language accuracy is 1.35 points, with a 90% confidence interval of $[0.85, 1.83]$ from $10{,}000$ bootstrap resamples of the $1{,}000$ items, which lies within a margin of ± 2 points.

C Ablations

Table 5 isolates what the probe architecture contributes. Replacing learned attention with uniform weights — the sentence-level comparison that cross-attention reduces to when every target token attends equally — costs 1.60 F1, and mean-pooling before comparison costs 2.00. Both say that

Table 4: Oracle correctness on 1,000 HellaSwag items per language. b counts items answered correctly in English only and c items answered correctly in the target language only; no item is answered incorrectly in both. χ^2 is the continuity-corrected McNemar statistic and p the exact McNemar p-value.

Target	English acc.	target acc.	b	c	χ^2	p
Swahili	99.6	99.4	6	4	0.10	0.754
Azerbaijani	99.6	98.0	20	4	9.38	0.002
Icelandic	99.6	97.8	22	4	11.12	0.001
Tagalog	99.6	97.8	22	4	11.12	0.001

Table 5: Probe architecture and placement for the attention probe with six high-resource alignment languages: MASSIVE transfer slot F1 ("tr") and its change from the full probe (Table 1, row 12). Six high-resource language SFT scores 53.7.

Probe	tr	Δ
Attention probe (full, layer 16)	**56.99**	–
Uniform attention	55.39	-1.60
Mean-pooled MLP	54.99	-2.00
Top layer (31)	55.79	-1.20
Bottom layer (8)	54.09	-2.90

token-level correspondence between the two sequences is doing real work, not that a larger scoring function is. Placement reproduces the depth profile of Fig. 2: the bottom layer is worst (-2.90), the top layer intermediate (-1.20), and the middle layer best, aligning with the trend of the relative scale $\|\Delta_{\mathrm{perf}}\|/\|\Delta_{\mathrm{lang}}\|$ among layers.

D MAPP IS A BETTER PROXY FOR DIRECT SUPERVISION FROM LABELED TARGET-LANGUAGE TASK DATA

We also directly contrast the **representation change** induced by the different training objectives, and compare the extent to which these induced representation changes resemble Δ_{perf} and Δ_{lang}. Let h_t^{base}, h_t^* and h_t^M denote the mean-pooled layer-16 representations of a target-language sentence under the base model, target-language SFT oracle and model M, respectively. For each parallel ARC-C pair, the oracle move is $u = h_t^* - h_t^{\mathrm{base}}$, and the language-switch move is $v = h_r^{\mathrm{base}} - h_t^{\mathrm{base}}$. We fit the model's displacement $d_M = h_t^M - h_t^{\mathrm{base}}$ as

$$(t_{\mathrm{oracle}}, t_{\mathrm{lang}}) = \arg\min_{a,b} \sum_i \|d_M^{(i)} - a u^{(i)} - b v^{(i)}\|_2^2, \tag{8}$$

where i indexes evaluated sentence pairs. This decomposition indicates how closely a method's representation shift matches the shift that direct target-language supervision would induce. The oracle training objective, namely SFT on target-language task data, is defined to yield coefficient $(1, 0)$, serving as the reference for an "optimal move".

Both MAPP probes retain more oracle-directed movement than either distance objective (Table 6). With English-only supervision, MAPP retains oracle travel of 0.42–0.46, compared with 0.21–0.31 for distance alignment, from an SFT starting point of 0.49. With six languages, MAPP extends SFT's oracle travel from 0.56 to 0.57–0.61, whereas InfoNCE and SAE-L_2 reduce it to 0.36 and 0.49. MAPP thus better preserves or extends the component of representation change associated with direct target-language supervision.

Table 6: Representation travel from different alignment methods measured against direct in-language supervision (t_{oracle}) and mean language shift (t_{lang}). Both MAPP probes yield greater oracle travel than either distance objective in both supervision settings; with six languages, MAPP-attn reaches 0.61, against 0.36 for InfoNCE and 0.49 for SAE-L_2. Bold marks the best score in each column within rows 3–7 and within rows 8–12 (lowest for t_{lang}, highest otherwise).

	Setting	oracle travel t_{oracle}	language travel t_{lang}	$t_{\text{oracle}}/t_{\text{lang}}$	lang-ID probe
1	Base model	0.00	0.00	–	99.8
2	Low-resource target lang SFT	1.00	0.00	–	99.8
3	English-only SFT	**0.49**	**0.02**	**24.5**	**99.8**
4	3 + English InfoNCE alignment	0.31	0.41	0.8	92.4
5	3 + English SAE-L_2 alignment	0.21	0.10	2.1	96.7
6	3 + En MAPP-SAE	0.42	0.17	2.5	97.1
7	3 + En MAPP-attn	0.46	0.05	9.2	99.6
8	6 high-resource lang SFT	0.56	**0.04**	**14.0**	**99.8**
9	8 + 6-lang InfoNCE alignment	0.36	0.43	0.8	83.4
10	8 + 6-lang SAE-L_2 alignment	0.49	0.12	4.1	87.1
11	8 + 6-lang MAPP-SAE	0.57	0.19	3.0	94.9
12	8 + 6-lang MAPP-attn	**0.61**	0.07	8.7	98.7