

Why Better Cross-Lingual Alignment Fails for Better Cross-Lingual Transfer: Case of Encoders

Yana Veitsman^{1*} Yihong Liu^{2,3} Hinrich Schütze^{2,3}

¹University of Göttingen ²LMU Munich

³Munich Center for Machine Learning (MCML)

Abstract

Cross-lingual alignment is often assumed to improve cross-lingual transfer by bringing representations of different languages closer together. However, improvements in representational alignment do not consistently translate into better downstream performance. We investigate this disconnect using XLM-R models explicitly aligned across four language pairs with *token-level*, *sentence-level*, and *masked-language-modeling* objectives. We evaluate their zero-shot transfer on a token-level task (*part-of-speech tagging*) and a sentence-level task (*sentence classification*), and analyze both representational changes and the gradients induced by the alignment and downstream objectives. We find that embedding-based alignment metrics do not reliably indicate whether alignment will improve or degrade downstream performance. Moreover, alignment and downstream-task gradients are often nearly *orthogonal*, particularly when the alignment objective and downstream task operate at different representational levels. These findings suggest that representation alignment alone is insufficient for assessing cross-lingual transfer, and that the compatibility between alignment and downstream objectives should be considered when designing/evaluating alignment methods.

1 Introduction

Despite their growing multilingual use, large language models (LLMs), including multilingual LLMs, remain heavily English-centric in training data (Blasi et al., 2022; Xu et al., 2025; Qin et al., 2025). Recent multilingual corpora (Nguyen et al., 2024; Penedo et al., 2025; Jumelet et al., 2026) and model families (Üstün et al., 2024; Martins et al., 2025; Hernández-Cano et al., 2025) have expanded linguistic coverage, yet performance remains uneven, particularly for low-resource languages, due

to limited training data (Veitsman and Hartmann, 2025), typological variation (Petrov et al., 2023), and capacity constraints such as the *curse of multilinguality* (Conneau et al., 2020).

A central strategy for improving performance in low-resource languages is *cross-lingual transfer*, in which knowledge learned from one or more source languages is transferred to a target language under zero-shot or few-shot conditions (Nooralahzadeh et al., 2020; Asai et al., 2024). One widely studied approach to facilitating this transfer is *cross-lingual alignment*, which aims to make representations comparable across languages (Wu and Dredze, 2020; Gaschi et al., 2023; Hämmerl et al., 2024; Liu et al., 2025). Initially developed for multilingual encoder models, this principle has motivated a broad range of alignment objectives and training methods (Chi et al., 2021; Philippp et al., 2023; Xhelili et al., 2024; Hämmerl et al., 2024).

Despite its intuitive appeal, the relationship between representational alignment and downstream cross-lingual transfer remains unclear. Explicit alignment can substantially improve intrinsic measures such as embedding similarity and cross-lingual retrieval, yet these improvements do not consistently translate into better downstream performance. In some cases, alignment even degrades performance on tasks such as part-of-speech tagging and named entity recognition (Gaschi et al., 2023; Liu et al., 2024). This raises a fundamental question: *why does better cross-lingual alignment not reliably lead to better cross-lingual transfer?*

Approach and Contributions We investigate this question from two complementary perspectives. First, we examine how explicit alignment changes the multilingual representation space and whether these changes correspond to downstream performance. Second, we study the interaction between the optimization objectives themselves by comparing gradients induced by alignment and

*Work done while at Saarland University. Correspondence to: yana.veitsman@uni-goettingen.de.

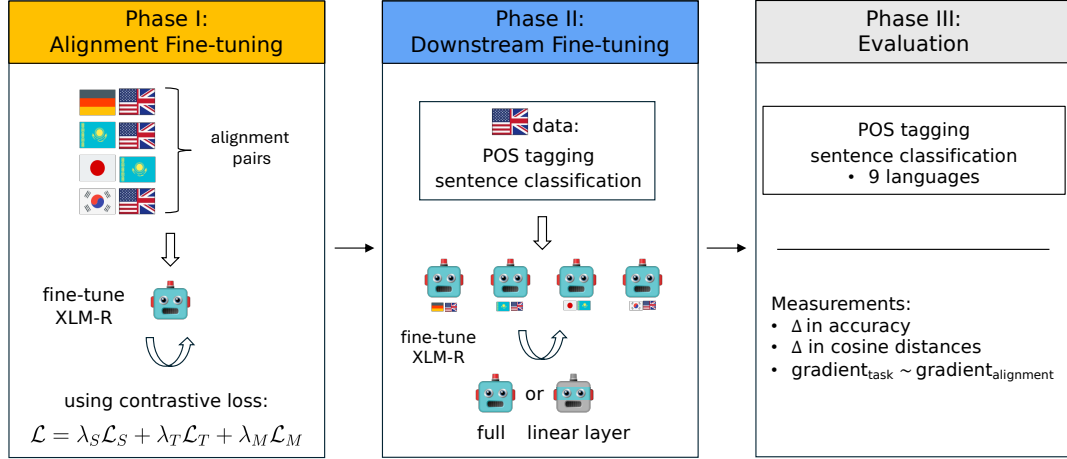


Figure 1: Overview of the experimental setup, including three phases: 1) fine-tuning alignment of XLM-R_{base} model on 4 language pairs using a contrastive loss objective; 2) full or linear layer-only fine-tuning downstream task performance of aligned models using English data for POS TAGGING and SENTENCE CLASSIFICATION; 3) evaluation of the models via Δ in accuracies on the task (zero-shot vs. fine-tuned), Δ in cosine distances between language embeddings, and alignment and task cosine similarity analysis.

downstream-task supervision. Intuitively, even if an alignment objective brings representations closer across languages, the resulting parameter updates may not be compatible with those required by a downstream task. This distinction allows us to separate *representational proximity* from *optimization compatibility*. For an overview of our approach, refer to Figure 1.

Our contributions therefore lie along two axes:

1. We show that improvements in layer-wise cosine-distance and retrieval-based alignment metrics are not reliable predictors of downstream task performance (Section 4.1).
2. We find that gradients induced by alignment and task-specific objectives are largely orthogonal, and that their interaction can conflict with the optimization required for improved downstream performance (Section 4.3).

Overall, our findings demonstrate that the downstream consequences of alignment are better characterized by optimization compatibility—measured here through gradient similarity—than by representational proximity alone.

2 Background and Related Work

We focus on encoder-only Transformer architectures, such as BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). Their cross-lingual generalization is primarily attributed to shared parameters inducing a joint multilingual space (Pires et al.,

2019; Dufter and Schütze, 2020; K et al., 2020). Contextual representations of this space provide a common ground for measuring and defining cross-lingual alignment (Hämmerl et al., 2024).

Cross-Lingual Alignment Cross-lingual alignment refers to the extent to which semantically related inputs from different languages occupy similar regions of this shared space. Two notions, commonly referred to as *strong* and *weak* alignment, can be distinguished (Roy et al., 2020; Hämmerl et al., 2024). Throughout this paper, we use the definition of weak alignment provided in Appendix A.

Alignment can be encouraged at the *token level*, the *sentence level*, or both (Hämmerl et al., 2024). Token-level alignment is particularly relevant to tasks such as part-of-speech tagging and named entity recognition, whereas sentence-level alignment is more directly relevant to sentence classification and cross-lingual retrieval.

Internal Representations Languages occupy a shared embedding space, with deeper layers exhibiting stronger cross-lingual alignment (Mousi et al., 2024). Shared tokens may serve as representational anchors (Hua et al., 2024), and alignment is stronger among related languages, consistent with semantic hubness patterns (Wu et al., 2025).

Contrastive Alignment We focus on improving explicit representation alignment (Wu and Dredze, 2020; Hu et al., 2021) via contrastive learning (Wei et al., 2021; Chi et al., 2021). Contrastive objec-

tives reduce the distance between representations of semantically equivalent inputs (positive pairs) while increasing their distance from unrelated inputs (negative pairs). Applied at the token and sentence levels, they encourage more consistent cross-lingual representations and a more isotropic embedding space. For example, InfoXLM (Chi et al., 2021) maximizes cross-lingual mutual information alongside MLM, while HiCTL (Wei et al., 2021) learns representations at the word, sentence, and context levels. In our work, we use a modification of the HiCTL objective.

Alignment and Transfer The link between alignment and downstream transfer remains insufficiently explained. Explicit contrastive alignment yields inconsistent gains when controlling for model size and seed variance (Wu and Dredze, 2020). Encoder-only models show stronger correlations (Gaschi et al., 2023), but task-specific fine-tuning can reduce alignment, particularly for token-level tasks, possibly due to catastrophic forgetting. Moreover, stronger alignment does not reliably improve downstream performance and can even degrade it, e.g., with transliteration-based objectives (Liu et al., 2024; Xhelili et al., 2024).

3 Experimental Setup

To study how better cross-lingual alignment interacts with the downstream task performance in multilingual encoders, we develop two orthogonal hypotheses, related to either linguistic typology or to downstream task interaction.

H_I: Linguistic Relatedness Building on typological evidence, we expect that zero-shot transfer depends on the linguistic proximity between (a) alignment language pair, (b) the fine-tuning language, and (c) the target language:

Hypothesis I

Zero-shot performance depends on the degree of linguistic relatedness among alignment, fine-tuning, and target languages.

H_{II}: Competing Objectives Previously, Wang et al. (2023) reported interference between the gradients of different languages (both typologically similar and distinct) during pre-training stage. We hypothesize that such misalignment can exist between languages, but also between training objec-

tives themselves. Therefore, we formulate the second hypothesis as follows:

Hypothesis II

Alignment and downstream-task objectives induce parameter updates in opposing directions, limiting joint optimization.

To evaluate the effects of improved representational alignment, we also test the following propositions suggested by prior work: (i) embedding distances alone are insufficient predictors of task performance; (ii) fine-tuning consistently improves downstream task performance; and (iii) different representation levels capture distinct linguistic properties relevant to different tasks. To examine the above stated hypotheses and propositions systematically, our experimental setup consists of three phases: (i) alignment fine-tuning, (ii) downstream task fine-tuning, and (iii) zero-shot evaluation.

3.1 Phase I: Cross-lingual Alignment Fine-Tuning

Model, Data, and Languages For all experiments, we use the 270M XLM-R_{base} (Conneau et al., 2020) model. We perform cross-lingual alignment fine-tuning using OPUS (Tiedemann and Nygaard, 2004) or MultiCCAligned (El-Kishky et al., 2020) datasets, depending on the specific language corpus availability. To allow for better typological diversity, we select German-English, English-Kazakh, Japanese-Kazakh, and Korean-Kazakh language pairs. For brevity, we refer to each checkpoint by the model-language pair abbreviation, for example XLM-R_{de-en} for German-English (and XLM-R_{en-kk}, XLM-R_{ja-kk}, or XLM-R_{kk-ko} for other pairs respectively). We report relevant data and training parameters in Appendix B.

Fine-tuning Objective We adopt the Hierarchical Contrastive Learning objective from Wei et al. (2021) with a number of modifications. The overall loss is a sum of scaled losses of two representation levels ($\mathcal{L}_S$ and $\mathcal{L}_T$) and the regular MLM objective ($\mathcal{L}_M$):

$$\mathcal{L} = \lambda_S \mathcal{L}_S + \lambda_T \mathcal{L}_T + \lambda_M \mathcal{L}_M \quad (1)$$

Sentence-level Loss ($\mathcal{L}_S$) aims to pull sentence of similar semantic content together, while pushing non-similar sentences apart. It is based on the

POS Tagging							
Full Model Fine-tuning							
Model	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
XLM-R_{base}	.72	.72	.72	.72	.72	.72	.72
XLM-R _{de-en}	.01 ± .00	.00 ± .02	.01 ± .01	.01 ± .02	.02 ± .01	.01 ± .01	.02 ± .01
XLM-R _{en-kk}	-.02 ± .01	-.03 ± .02	-.02 ± .01	.01 ± .02	.00 ± .01	.01 ± .02	.00 ± .03
XLM-R _{ja-kk}	.02 ± .02	.00 ± .01	.02 ± .01	.02 ± .02	.03 ± .01	.01 ± .02	.04 ± .01
XLM-R _{kk-ko}	.01 ± .02	.01 ± .03	.01 ± .01	.01 ± .02	.01 ± .02	.00 ± .03	.02 ± .02
Linear Layer-Only Fine-tuning							
Model	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
XLM-R_{base}	.63	.63	.63	.63	.63	.63	.63
XLM-R _{de-en}	.10 ± .00	.01 ± .01	.08 ± .00	.02 ± .00	.04 ± .00	.02 ± .01	-.03 ± .00
XLM-R _{en-kk}	.08 ± .00	-.09 ± .00	.06 ± .00	.01 ± .00	.07 ± .00	.03 ± .00	-.01 ± .00
XLM-R _{ja-kk}	.10 ± .00	-.02 ± .00	.08 ± .00	.02 ± .00	.03 ± .00	-.02 ± .00	-.04 ± .00
XLM-R _{kk-ko}	.06 ± .00	-.02 ± .00	.06 ± .00	.04 ± .00	.01 ± .00	.02 ± .00	-.02 ± .00
Sentence Classification							
Full Model Fine-tuning							
Model	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
XLM-R_{base}	.88	.88	.88	.88	.88	.88	.88
XLM-R _{de-en}	.00 ± .00	-.02 ± .01	.00 ± .01	.00 ± .00	.00 ± .00	-.01 ± .01	.01 ± .00
XLM-R _{en-kk}	.00 ± .00	-.02 ± .01	.00 ± .01	.00 ± .00	.01 ± .01	-.01 ± .01	.00 ± .01
XLM-R _{ja-kk}	.00 ± .00	-.03 ± .02	.00 ± .01	-.01 ± .00	.00 ± .00	.00 ± .00	.00 ± .01
XLM-R _{kk-ko}	.01 ± .01	-.04 ± .02	.00 ± .00	.00 ± .00	.00 ± .01	-.01 ± .00	-.01 ± .00
Linear Layer-Only Fine-tuning							
Model	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
XLM-R_{base}	.72	.72	.72	.72	.72	.72	.72
XLM-R _{de-en}	.01 ± .01	-.09 ± .00	-.05 ± .01	.15 ± .00	.15 ± .00	.14 ± .00	.13 ± .00
XLM-R _{en-kk}	-.04 ± .00	-.20 ± .01	-.04 ± .01	.10 ± .00	.11 ± .00	.10 ± .00	.10 ± .00
XLM-R _{ja-kk}	.02 ± .01	-.06 ± .01	-.06 ± .00	.11 ± .00	.12 ± .00	.10 ± .00	.12 ± .00
XLM-R _{kk-ko}	-.02 ± .01	-.18 ± .00	.01 ± .01	.13 ± .00	.11 ± .00	.11 ± .00	.11 ± .00

Table 1: POS TAGGING (**top**) and SENTENCE CLASSIFICATION (**bottom**) results per aligned model and loss configuration. Each block reports Δ accuracy from XLM-R_{base}, averaged over 3 seeds. **Red** indicates deterioration, **green** indicates improvement, and black indicates inconclusive results.

standard InfoNCE loss (van den Oord et al., 2019):

$$\mathcal{L}_S = \frac{-1}{2B} \sum_{k=1}^{2B} \log \frac{\exp(s(z_k, z_{pos(k)}))}{\sum_{j \neq k} \exp(s(z_k, z_j))}, \quad (2)$$

where B is the batch size, k is an index of the sentence embedding in the parallel dataset, z_k is the k -th embedding of the concatenated tensor, $pos(k)$ is the index of the positive partner of the sentence, and s is the chosen similarity metric (which is, in our case, cosine similarity).

Token-level ($\mathcal{L}_T$) Loss assumes that tokens often occur in similar contexts. So, it uses a token-level bag-of-words $\mathcal{B}$ from concatenated parallel sentences as a positive set and $\mathcal{S} \subset \mathcal{V} \setminus \mathcal{B}$, where $\mathcal{V}$ is the model vocabulary, as a negative token subset:

$$\mathcal{L}_T = \frac{-1}{|\mathcal{B}|} \sum_{t=1}^{|\mathcal{B}|} \log \frac{\exp(s(q_i, e(w_t)))}{\exp(s(q_i, e(w_t))) + \sum_{w_j \in \mathcal{S}} \exp(s(q_i, e(w_j)))}, \quad (3)$$

where e is the embedding lookup function, q is the query token embedding, t is an index in $\mathcal{B}$, and w_t is the respective token at that index, and s is the chosen similarity metric.

Loss Scaling Factor λ_k . To evaluate the effect of each loss component independently, we apply a loss scaling factor as defined in Liu et al. (2024). We enumerate all non-zero combinations of loss coefficients:

$$\lambda_k \in \{0, 1\}, \quad \sum_k \lambda_k > 0 \quad (4)$$

This yields a total of 7 loss configurations $\mathcal{C}$ per alignment pair.

$$C = (\lambda_S, \lambda_T, \lambda_M) \in \{0, 1\}^3 : (\lambda_S, \lambda_T, \lambda_M) \neq (0, 0, 0) \quad (5)$$

For visual clarity, we denote these configurations in the form of S1T0M1, where letters corresponds to loss type followed by the coefficient applied along with that specific loss.

Sentence Retrieval and Embedding Distances

To verify the effectiveness of the alignment performed in Phase I, we use two widely adopted measures: *sentence retrieval recall* and *cosine-based embedding distances* (Hämmerl et al., 2024). Sentence retrieval is evaluated on the FLORES-200 parallel dataset (Team et al., 2022) using mean-pooled representations from the 8th layer, while token- and sentence-level embedding distances are measured using MUSE bilingual dictionaries (Lample et al., 2018) and FLORES-200 (Team et al., 2022), respectively. We compare embeddings under three conditions: (i) the base model (XLM-R_{base}), (ii) the aligned models (XLM-R_{de-en}, XLM-R_{en-kk}, XLM-R_{kk-ko}, and XLM-R_{ja-kk}), and (iii) the downstream fine-tuned models. For this analysis, representations are mean-pooled from the 4th, 8th, and 12th layers. Evaluation across these layers is done so the changes within the layers that are supposed to be more or less language-specific can be easily captured.

3.2 Phase II: Downstream Fine-tuning

After alignment fine-tuning, we separately fine-tune all checkpoints for token-level (POS TAGGING) and sentence-level (SENTENCE CLASSIFICATION) tasks on respective English language data. For POS TAGGING task, we use Universal Dependency treebanks (Nivre et al., 2020). For SENTENCE CLASSIFICATION, we employ SIB-200 (Adelani et al., 2024), which enhances the FLORES-200 dataset (Team et al., 2022) with classification labels.

3.2.1 Fine-tuning Regimen

Encoder fine-tuning commonly follows one of two strategies: (i) *full fine-tuning*, where the entire model is updated, and (ii) *linear-only fine-tuning*, where only the classification layer is trained while the encoder backbone remains frozen (Devlin et al., 2019; Peters et al., 2019). This comparison facilitates the investigation of our H_{II}. In case the hypothesis is correct, we expect that linear-only fine-tuning will preserve the representations induced by

the alignment objective, whereas full fine-tuning would amplify the interference between the objectives.

Gradient Similarity Similar to a multi-task learning setup, objectives of alignment and task may produce *conflicting gradients* (Sener and Koltun, 2018; Wang et al., 2023). Building on this insight, we analyze gradient interactions at both task and language levels. Specifically, we examine (i) gradients between downstream fine-tuning task and alignment objectives, (ii) per-language task gradients after alignment using zero-shot evaluation, and (iii) per-language alignment gradients computed via the MLM objective on FLORES-200 (Team et al., 2022).

3.3 Phase III: Zero-shot Transfer

After downstream task fine-tuning, we perform zero-shot evaluation of each checkpoint and report the change in classification accuracy for downstream task. The set of evaluation languages for both tasks includes both the languages that were and were not present in the alignment enhancement phase:

$$L_{POS} = L_{SENT} = \{de, es, fa, fr, hi, ja, kk, ko, ru\} \quad (6)$$

4 Results

We observe that even though cross-lingual alignment improves after Phase I, this improvement does not directly translate to better performance, depending on the fine-tuning regimen and loss configuration. We attribute this to the gradient conflict between the alignment and the downstream task training objectives.

4.1 Embedding Distances

After Phase I, we evaluate changes in token- and sentence-level embedding distances and sentence-level cosine similarity and retrieval. Despite substantial cross-lingual variance, consistent layer-, model-, and objective-specific patterns emerge.

Distances Across Layers Layer-wise results (Figure 5 in Appendix) support prior distinctions between relatively language-agnostic and more language- or task-specific layers (Pires et al., 2019; Tanti et al., 2021; Choenni and Shutova, 2022). Distances generally decrease in the 8th layer for both aligned and downstream fine-tuned models, except under the S0T1M0 configuration (Figure 3).

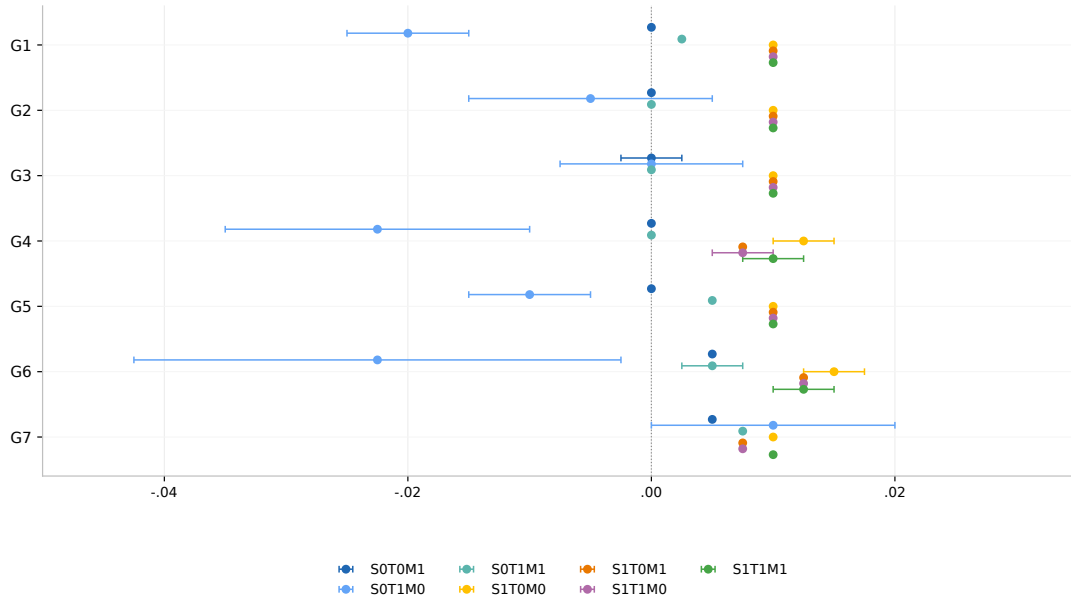


Figure 2: Group-averaged cosine similarity changes across loss configurations. Points show the mean Δ cosine similarity relative to $\text{XLM-R}_{\text{base}}$, and whiskers show the averaged standard deviation within each group. Group definitions and detailed results are provided in Appendix, Table 8.

Changes in lower and final layers, which are more heterogeneous, likely reflect language- and task-specific adaptation.

Cosine Similarities Sentence-level similarity gains are generally modest, partly because some baseline pairs are already highly similar (Figure 2, details in Appendix, Table 8). For example, Spanish–English reaches approximately .98. Most improvements remain below .02, with the smallest gains observed for non-aligned pairs involving English and different scripts (Group 3: hin–eng, zho–eng, spa–eng, and rus–eng). This pattern suggests that alignment fine-tuning may partially shift these languages away from the English-centered representational space. Sentence retrieval results exhibit the same tendency (see Tables 5 and 6 in Appendix).

4.2 Downstream Task Performance

We find that downstream task performance is rather driven by the specific loss configuration and fine-tuning regimen, than a particular language alignment pair (Table 1).

Loss Configuration Some loss combinations generally hurt the performance across tasks, languages, and fine-tuning modes. For example, in the case of the *S0T1M0* configuration, performance degrades in both fine-tuning modes and tasks. On the other hand, configurations that include a com-

bination of sentence-level loss with either or both token-level and MLM objectives, tend to improve the downstream accuracy more often. Moreover, inclusion of sentence-level loss also positively affects POS TAGGING, specifically in the linear layer fine-tuning case.

Fine-tuning Regimen Choice of the fine-tuning regimen is also impactful: full model fine-tuning tends to affect task solving abilities negatively, while the linear layer-only fine-tuning shows more robust improvements. For example, $\text{XLM-R}_{\text{de-en}}$ experiences up to 15% accuracy gain on SENTENCE CLASSIFICATION in the case of *S1T0M0* and *S1T0M1* configurations. Consistent improvements are also seen in POS TAGGING, $\text{XLM-R}_{\text{en-kk}}$ becomes up to 7% more accurate.

Thus, although the similarity metrics showcase improved alignment, the impact on downstream performance is more fine-grained, particularly so more within the applied loss and training mode, than language pairs.

4.3 Gradient Similarity

To test H_{II} , we calculate the cosine similarity between task and alignment objectives. Here we observe a clear separation: gradients for SENTENCE CLASSIFICATION are generally more positively aligned with the alignment objective, whereas POS TAGGING gradients frequently approach zero or be-

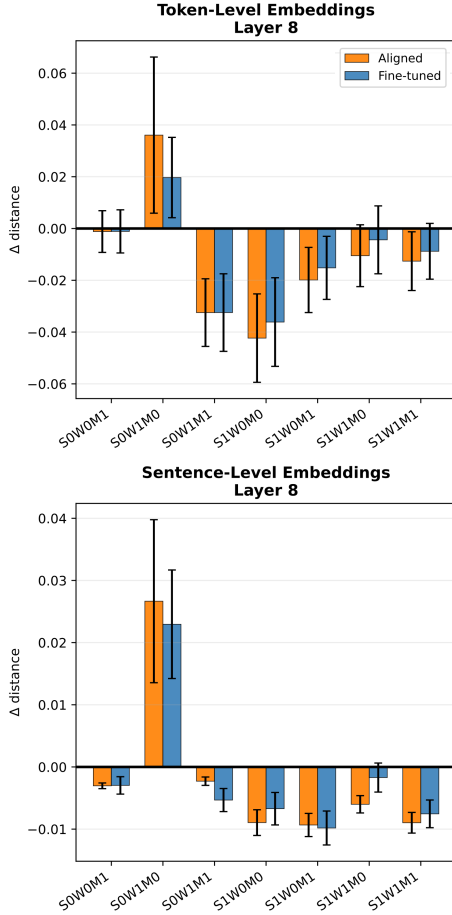


Figure 3: Δ in cosine distance from XLM-R_{base} for embeddings averaged across models and languages and grouped by loss configuration. Full figure, including layers 4 and 12, is available in the Figure 5 in Appendix.

come negative, indicating orthogonality or conflict (Figure 4). These results suggest fundamental differences in how alignment objectives interact with token- vs. sentence-level supervision.

Language pair and loss configuration refine this picture. Alignment pairs that exclude English show higher gradient similarity and thus reduced task conflict. On the other hand, English-involving alignments exhibit stronger task separation, with consistently positive cosine values for SENTENCE CLASSIFICATION and negative ones for POS TAGGING. Adding a sentence-level alignment loss increases similarity for sentence-level tasks as expected, whereas the MLM-only setup shows moderate synergy with SENTENCE CLASSIFICATION but remains weakly aligned with token-level objectives. Overall, the findings confirm that alignment and downstream tasks differ substantially in optimization behavior, depending on representation level and alignment configuration.

5 Discussion

We confirm that improving cross-lingual representational alignment is not, by itself, sufficient to improve zero-shot downstream performance. Instead, the effect of alignment fine-tuning depends on which representation level is optimized, how compatible the alignment and downstream objectives are, and whether subsequent fine-tuning preserves or modifies the aligned encoder.

Alignment fine-tuning consistently changes the multilingual representation space, particularly in the 8th layer, where token- and sentence-level distances generally decrease. Sentence-level cosine similarities and retrieval performance also improve for many language pairs. Nevertheless, these representational improvements do not consistently translate into higher downstream accuracy. Some configurations produce clearer cross-lingual similarity while having little effect on task performance, and the token-only S0T1M0 objective frequently deteriorates performance despite explicitly optimizing token-level alignment.

Takeaway 1

Embedding similarity and sentence-retrieval metrics are insufficient predictors of downstream task performance.

Gradient analysis provides a more direct explanation for the discrepancy between representational alignment and downstream performance. Gradients from SENTENCE CLASSIFICATION are more positively aligned with the preceding alignment objective, whereas gradients from POS TAGGING are frequently close to orthogonal or negatively aligned. Thus, the two downstream tasks interact differently with the same aligned representation space: sentence-level supervision can often reuse directions encouraged during alignment fine-tuning, while token-level supervision more frequently requires updates that compete with them. This conclusion is supported by the comparison of downstream accuracy results in the fully and linear layer-only tuned models, where tuning across layers seems to amplify the interference.

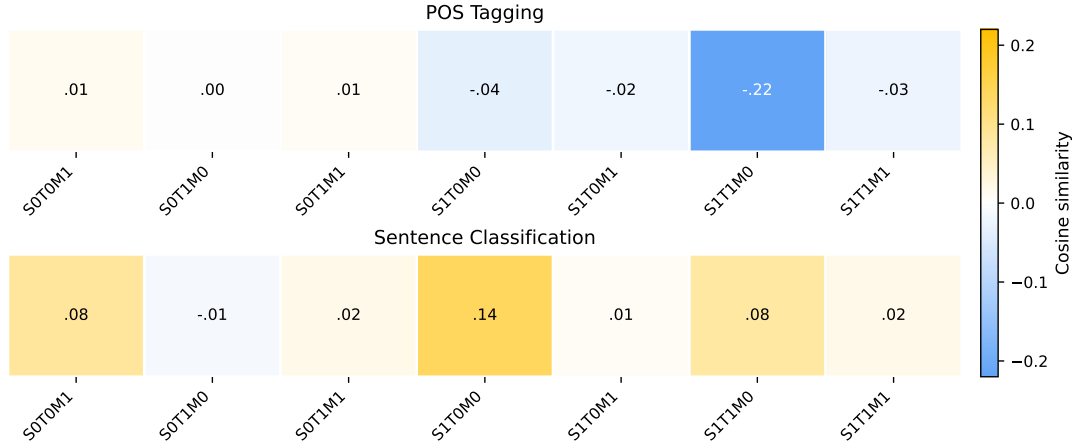


Figure 4: Cosine similarity between gradients induced by the alignment objective and those induced by each downstream task: POS TAGGING (**top**) and SENTENCE CLASSIFICATION (**bottom**), averaged over four alignment pairs and grouped by loss configuration. Detailed results per language pair with corresponding standard deviations are available in Appendix, Table 7.

Takeaway II

Compatibility between alignment and downstream objectives depends on the task and representation level.

More broadly, the results emphasize that the usefulness of alignment depends not only on how close representations become, but also on which information is preserved and whether the downstream objective can exploit them without undoing the alignment itself.

6 Conclusion

We show that improved cross-lingual representational alignment does not necessarily translate into better downstream transfer: embedding similarity can indicate stronger alignment even when downstream performance remains unchanged or deteriorates. Our gradient analysis provides a complementary explanation for this disconnect, showing that alignment and downstream objectives are often weakly aligned or conflicting, particularly when they operate at different representational levels. We further find that linear-only fine-tuning more consistently preserves the benefits of alignment than full-model fine-tuning. Overall, our results suggest that representational alignment should be treated as a diagnostic model property rather than a task-independent proxy for transfer, and highlight the importance of considering the compatibility between alignment and downstream objectives when designing cross-lingual alignment methods.

7 Limitations

Model Architecture Our analysis is constrained by its reliance on a single Transformer-based encoder model, which limits the generality of the findings. Different encoder and decoder-only architectures exhibit varying multilingual capabilities, and extending the study would help test whether the observed dynamics hold more broadly.

Pre-training and Tokenization Disparities We do not examine the effect of pre-training language distributions, data quality, or volume on the effectiveness of the cross-lingual alignment. Additionally, we do not account for potential tokenization disparities, which might further drive linguistic inequality.

Training Objective Limitations We test only one specific contrastive learning objective. Other design choices, for example, replacing the token-level alignment with a carefully designed word-level alignment using bilingual dictionaries, may be of interest for future work.

8 Acknowledgements

YV thanks Vera Demberg, Mark RoFin, and Oleh Bakumenko for useful discussions and feedback on the draft. YL and HS were supported by DFG (grant SCHU 2246/14-1).

References

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. [SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian’s, Malta. Association for Computational Linguistics.
- Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. [BUFFET: Benchmarking large language models for few-shot cross-lingual transfer](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.
- Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. [Systematic inequalities in language technology performance across the world’s languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.
- Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. [InfoXLM: An information-theoretic framework for cross-lingual language model pre-training](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.
- Rochelle Choenni and Ekaterina Shutova. 2022. [Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology](#). *Computational Linguistics*, 48(3):635–672.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- 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.
- Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. [CCAligned: A massive collection of cross-lingual web-document pairs](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.
- Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. [Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3020–3042, Toronto, Canada. Association for Computational Linguistics.
- Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. [Understanding cross-lingual Alignment—A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.
- Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 82 others. 2025. [Apertus: Democratizing open and compliant llms for global language environments](#). *Preprint*, arXiv:2509.14233.
- Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. [Explicit alignment objectives for multilingual bidirectional encoders](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.
- Tianze Hua, Tian Yun, and Ellie Pavlick. 2024. [mOthello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models?](#) In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1585–1598, Mexico City, Mexico. Association for Computational Linguistics.
- Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrama, Francesca Padovani, Francois Meyer, Hai Hu, Julien Etxaniz, Laurent Prevot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, and 7 others. 2026. [BabyBabelLM: A](#)

- multilingual benchmark of developmentally plausible training data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3297–3329, Rabat, Morocco. Association for Computational Linguistics.
- Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. [Cross-Lingual Ability of Multilingual BERT: An Empirical Study](#). In *Proc. of the International Conference on Learning Representations*.
- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ran- zato, Ludovic Denoyer, and Hervé Jégou. 2018. [Word translation without parallel data](#). In *International Conference on Learning Representations*.
- Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024. [TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.
- Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Ayyoob ImaniGooghari, Orgest Xhelili, Haotian Ye, Chunlan Ma, François Yvon, and Hinrich Schütze. 2025. [How transliterations improve crosslingual alignment](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2417–2433, Abu Dhabi, UAE. Association for Computational Linguistics.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. [Eurollm: Multilingual language models for europe](#). *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.
- Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. 2024. [Exploring alignment in shared cross-lingual spaces](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6326–6348, Bangkok, Thailand. Association for Computational Linguistics.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. [Zero-shot cross-lingual transfer with meta learning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Al- lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2025. The fineweb datasets: decanting the web for the finest text data at scale. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. [To tune or not to tune? adapting pretrained representations to diverse tasks](#). In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.
- Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. [Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. [A survey of multilingual large language models](#). *Patterns*, 6(1):101118.
- Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. [LAREQA: Language-agnostic answer retrieval from a multilingual pool](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

- Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 525–536, Red Hook, NY, USA. Curran Associates Inc.
- Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. [On the language-specificity of multilingual BERT and the impact of fine-tuning](#). In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad and 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, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation.
- Jörg Tiedemann and Lars Nygaard. 2004. [The OPUS corpus - parallel and free: <http://logos.uio.no/opus>](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. [Aya model: An instruction fine-tuned open-access multilingual language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. [Representation learning with contrastive predictive coding](#). *Preprint*, arXiv:1807.03748.
- Yana Veitsman and Mareike Hartmann. 2025. [Recent advancements and challenges of Turkic Central Asian language processing](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 309–324, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schuetze. 2023. [GradSim: Gradient-based language grouping for effective multilingual training](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4631–4646, Singapore. Association for Computational Linguistics.
- Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. [On learning universal representations across languages](#). In *International Conference on Learning Representations*.
- Shijie Wu and Mark Dredze. 2020. [Do explicit alignments robustly improve multilingual encoders?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.
- Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. [The semantic hub hypothesis: Language models share semantic representations across languages and modalities](#). In *International Conference on Learning Representations*.
- Orgest Xhelili, Yihong Liu, and Hinrich Schuetze. 2024. [Breaking the script barrier in multilingual pre-trained language models with transliteration-based post-training alignment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11283–11296, Miami, Florida, USA. Association for Computational Linguistics.
- Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin XU, Yuqi Ye, and Hanwen Gu. 2025. [A survey on multilingual large language models: corpora, alignment, and bias](#). *Frontiers of Computer Science*, 19:1911362.

A Strong and Weak Cross-Lingual Alignment

We formalize *strong* and *weak* cross-lingual alignment as follows. Let $\mathcal{S} = \{L_1, \dots, L_n\}$ be the set of languages and $\mathcal{X}_L$ the input space for language L . An encoder

$$f_\theta : \bigcup_{L \in \mathcal{S}} \mathcal{X}_L \rightarrow \mathbb{R}^d \quad (7)$$

maps inputs into a shared representation space, and $\text{sim} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ measures embedding similarity. We define a semantic equivalence relation $x \sim x'$ for pairs expressing the same meaning and polarity; non-equivalent $x \not\sim x'$ includes unrelated or opposite content. Let $\mathcal{D}_{\text{pos}}$ and $\mathcal{D}_{\text{neg}}$ denote distributions over equivalent and non-equivalent pairs.

Strong Alignment. The encoder f_θ satisfies (δ, ϵ) -*strong alignment* between L_a and L_b if, with probability at least $1 - \epsilon$ over $(x, x') \sim \mathcal{D}_{\text{pos}}$ and $(x, x'') \sim \mathcal{D}_{\text{neg}}$,

$$\text{sim}(f_\theta(x), f_\theta(x')) \geq \text{sim}(f_\theta(x), f_\theta(x'')) + \delta. \quad (8)$$

This enforces that equivalent pairs remain more similar than non-equivalent ones by at least a margin δ .

Weak Alignment. To relax the binary equivalence assumption, we introduce a graded semantic similarity function

$$\sigma : \left(\bigcup_L \mathcal{X}_L \right) \times \left(\bigcup_L \mathcal{X}_L \right) \rightarrow [-1, 1], \quad (9)$$

with $\sigma(x, x') = 1$ for perfect translations and values near -1 for antonyms. We instantiate sim using cosine similarity:

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \|v\|}. \quad (10)$$

For a distribution $\mathcal{D}$ over arbitrary cross-lingual pairs, f_θ exhibits γ -weak alignment if

$$\text{Cov}_{(x, x') \sim \mathcal{D}} [\sigma(x, x'), \text{sim}(f_\theta(x), f_\theta(x'))] \geq \gamma, \quad (11)$$

or equivalently,

$$\text{Corr}_{(x, x') \sim \mathcal{D}} [\sigma(x, x'), \text{sim}(f_\theta(x), f_\theta(x'))] \geq \rho, \quad (12)$$

with $\rho = \gamma / \sqrt{\text{Var}[\sigma] \text{Var}[\text{sim}]}$. Weak alignment thus requires embedding similarity to increase with semantic relatedness without assuming strict translation equivalence.

B Hyperparameters and Computational Resources

Parameter	Value
Learning rate	$1e - 5$
Batch size	16
Max sequence length	128
Word negatives	512
Temperature	0.1
Number of epochs	1
Number of steps	$\approx 6,000$
Number of sentences*	$\approx 100,000$
Seed	42

Table 2: Hyperparameters used during alignment enhancement (Phase I). * each language pair uses 100,000 sentences, except English-Kazakh which only has 80,000 sentences available.

In Phase I (alignment fine-tuning), we use a single V100 NVIDIA GPU_{xt} to tune each variant of $\text{XLM-R}_{\text{base}}$ for each language pair (7 loss configurations per alignment pair, a total of 28 models). Total amount of GPU_h spent on alignment fine-tuning across models is ≈ 275 , with individual model tuning time varying from 5 to 22 GPU_h.

Parameter	Value
Batch size	16
Backbone learning rate (full model)	$1e - 4$
Classifier learning rate (full model)	$1e - 3$
Learning rate (linear layer only)	$5e - 3$
Number of epochs	15
Weight decay	0.01
Max sequence length	256
Seeds	15, 33, 42

Table 3: Hyperparameters used for full model and linear layer only fine-tuning (Phase II) for POS-Tagging.

Parameter	Value
Batch size	16
Learning rate (full model)	$2e - 5$
Learning rate (linear layer only)	$3e - 3$
Number of epochs	15
Max sequence length	256
Seeds	15, 33, 42

Table 4: Hyperparameters used for full model and linear layer only fine-tuning (Phase II) for Sentence Classification.

In Phase II for **POS Tagging fine-tuning**, we use a single V100 NVIDIA GPU to fine-tune each variant of aligned $\text{XLM-R}_{\text{base}}$, i.e. $\text{XLM-R}_{\text{de-en}}$, $\text{XLM-R}_{\text{en-kk}}$, $\text{XLM-R}_{\text{ja-kk}}$, $\text{XLM-R}_{\text{kk-ko}}$ for POS TAGGING as well as the baseline ($\text{XLM-R}_{\text{base}}$) itself. We fine-tune the whole model as well as the linear layer only. Total amount of GPU_h spent on POS TAGGING fine-tuning across models is ≈ 30 . In Phase II for **Sentence Classification fine-tuning** we use 1xV100 GPU to fine-tune each variant of aligned $\text{XLM-R}_{\text{base}}$, i.e. $\text{XLM-R}_{\text{de-en}}$, $\text{XLM-R}_{\text{en-kk}}$, $\text{XLM-R}_{\text{ja-kk}}$, $\text{XLM-R}_{\text{kk-ko}}$ for SENTENCE CLASSIFICATION as well as the baseline ($\text{XLM-R}_{\text{base}}$) itself. We fine-tune the whole model as well as the linear layer only. Total amount of GPU_h spent on SENTENCE CLASSIFICATION fine-tuning across models is ≈ 8 .

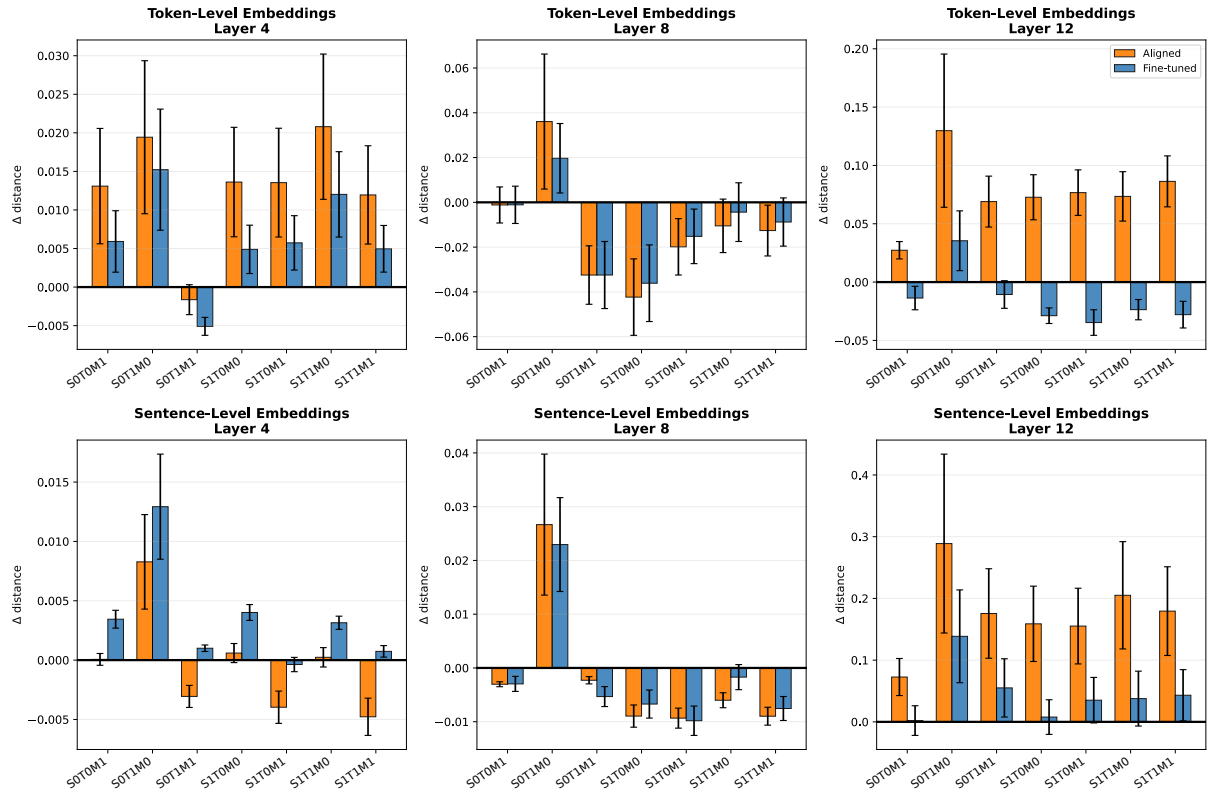


Figure 5: Δ in cosine distance from XLM- R_{base} for embeddings averaged across models and languages and grouped by loss configuration across 4th, 8th, and 12th layers.

C Evaluation Details

Lang Pair	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
Group 1: Aligned Languages With English							
deu-eng	0.00 ± 0.00	−0.09 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
kaz-eng	0.04 ± 0.01	−0.67 ± 0.22	−0.03 ± 0.07	0.06 ± 0.02	0.05 ± 0.02	0.03 ± 0.05	0.04 ± 0.04
kor-eng	0.02 ± 0.01	−0.58 ± 0.22	−0.01 ± 0.03	0.02 ± 0.01	0.01 ± 0.02	−0.01 ± 0.04	−0.02 ± 0.04
jpn-eng	0.01 ± 0.02	−0.49 ± 0.09	−0.03 ± 0.02	0.02 ± 0.01	0.01 ± 0.01	0.02 ± 0.01	0.01 ± 0.01
Group 2: Aligned Languages Without English							
jpn-kaz	0.02 ± 0.01	−0.51 ± 0.13	−0.02 ± 0.03	0.04 ± 0.02	0.05 ± 0.02	0.03 ± 0.02	0.03 ± 0.03
kor-kaz	0.03 ± 0.01	−0.41 ± 0.13	0.00 ± 0.03	0.04 ± 0.01	0.04 ± 0.01	0.03 ± 0.02	0.02 ± 0.02
Group 3: Other Languages With English							
hin-eng	0.03 ± 0.01	−0.48 ± 0.15	−0.03 ± 0.03	0.03 ± 0.01	0.02 ± 0.01	0.01 ± 0.02	0.00 ± 0.03
zho-eng	0.02 ± 0.00	−0.32 ± 0.08	0.00 ± 0.01	0.02 ± 0.01	0.02 ± 0.00	0.02 ± 0.01	0.02 ± 0.01
spa-eng	0.00 ± 0.00	−0.06 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
rus-eng	0.00 ± 0.00	−0.29 ± 0.15	0.00 ± 0.01	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Group 4: Languages of the Same Script, Non-Aligned							
lit-spa	0.02 ± 0.00	−0.30 ± 0.12	0.01 ± 0.02	0.04 ± 0.01	0.04 ± 0.01	0.04 ± 0.01	0.03 ± 0.01
hin-kas	0.00 ± 0.01	−0.02 ± 0.01	−0.02 ± 0.02	0.06 ± 0.02	0.03 ± 0.01	0.06 ± 0.04	0.04 ± 0.01
spa-fra	0.00 ± 0.00	−0.06 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
rus-bul	0.00 ± 0.00	−0.05 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Group 5: Languages of the Same Script, Aligned							
kaz-bul	0.06 ± 0.00	−0.46 ± 0.17	0.02 ± 0.04	0.08 ± 0.01	0.08 ± 0.01	0.06 ± 0.05	0.06 ± 0.02
kaz-rus	0.05 ± 0.01	−0.45 ± 0.18	0.03 ± 0.03	0.08 ± 0.02	0.08 ± 0.01	0.06 ± 0.05	0.07 ± 0.01
deu-spa	0.01 ± 0.00	−0.17 ± 0.07	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
deu-fra	0.01 ± 0.00	−0.13 ± 0.07	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.00	0.01 ± 0.01	0.01 ± 0.00
Group 6: Languages of Different Scripts, Non-Aligned							
spa-rus	0.00 ± 0.00	−0.15 ± 0.05	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
spa-kas	0.01 ± 0.01	−0.01 ± 0.01	−0.01 ± 0.01	0.05 ± 0.01	0.03 ± 0.01	0.05 ± 0.02	0.03 ± 0.01
heb-fra	0.03 ± 0.00	−0.36 ± 0.10	0.01 ± 0.02	0.04 ± 0.01	0.04 ± 0.01	0.04 ± 0.01	0.03 ± 0.01
hin-zho	0.03 ± 0.02	−0.41 ± 0.05	−0.02 ± 0.05	0.03 ± 0.03	0.04 ± 0.03	0.02 ± 0.04	0.02 ± 0.03
Group 7: Languages of Different Scripts, Aligned							
deu-heb	0.01 ± 0.00	−0.26 ± 0.09	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.01
kaz-hin	0.06 ± 0.01	−0.42 ± 0.16	0.02 ± 0.05	0.09 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.08 ± 0.01
jpn-rus	0.03 ± 0.01	−0.31 ± 0.04	0.00 ± 0.02	0.05 ± 0.01	0.05 ± 0.01	0.04 ± 0.02	0.04 ± 0.01
kor-fra	0.06 ± 0.01	−0.53 ± 0.17	0.01 ± 0.05	0.07 ± 0.02	0.08 ± 0.01	0.05 ± 0.05	0.04 ± 0.04

Table 5: R@5 Δ from baseline (XLM-R_{base}), averaged over all aligned models and grouped by loss configuration. Color coding indicates the averaged performance change magnitude: **red** for degradations, **green** for gains (adjusted for standard deviation).

Lang Pair	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
Group 1: Aligned Languages With English							
deu-eng	0.00 ± 0.00	−0.07 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
kaz-eng	0.03 ± 0.01	−0.64 ± 0.23	−0.03 ± 0.05	0.04 ± 0.02	0.04 ± 0.02	0.02 ± 0.04	0.02 ± 0.02
kor-eng	0.01 ± 0.00	−0.52 ± 0.23	−0.01 ± 0.02	0.01 ± 0.01	0.00 ± 0.01	−0.01 ± 0.03	−0.01 ± 0.03
jpn-eng	0.00 ± 0.01	−0.42 ± 0.08	−0.02 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.01
Group 2: Aligned Languages Without English							
jpn-kaz	0.02 ± 0.01	−0.46 ± 0.13	−0.01 ± 0.03	0.03 ± 0.01	0.04 ± 0.01	0.02 ± 0.02	0.02 ± 0.01
kor-kaz	0.03 ± 0.01	−0.34 ± 0.11	0.01 ± 0.02	0.03 ± 0.01	0.03 ± 0.01	0.03 ± 0.01	0.02 ± 0.01
Group 3: Other Languages With English							
hin-eng	0.02 ± 0.00	−0.43 ± 0.16	−0.02 ± 0.02	0.02 ± 0.01	0.02 ± 0.01	0.01 ± 0.01	0.00 ± 0.02
zho-eng	0.01 ± 0.00	−0.26 ± 0.08	0.00 ± 0.01	0.02 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
spa-eng	0.00 ± 0.00	−0.04 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
rus-eng	0.00 ± 0.00	−0.23 ± 0.13	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Group 4: Languages of the Same Script, Non-Aligned							
lit-spa	0.02 ± 0.00	−0.25 ± 0.10	0.00 ± 0.01	0.03 ± 0.01	0.03 ± 0.00	0.03 ± 0.00	0.02 ± 0.01
hin-kas	0.00 ± 0.01	−0.04 ± 0.01	−0.03 ± 0.03	0.07 ± 0.03	0.04 ± 0.02	0.07 ± 0.05	0.04 ± 0.01
spa-fra	0.00 ± 0.00	−0.04 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
rus-bul	0.00 ± 0.00	−0.03 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Group 5: Languages of the Same Script, Aligned							
kaz-bul	0.05 ± 0.01	−0.42 ± 0.19	0.01 ± 0.03	0.06 ± 0.01	0.06 ± 0.01	0.04 ± 0.04	0.05 ± 0.01
kaz-rus	0.04 ± 0.01	−0.42 ± 0.19	0.01 ± 0.02	0.06 ± 0.01	0.06 ± 0.01	0.04 ± 0.04	0.05 ± 0.01
deu-spa	0.00 ± 0.00	−0.12 ± 0.06	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
deu-fra	0.00 ± 0.00	−0.10 ± 0.06	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
Group 6: Languages of Different Scripts, Non-Aligned							
spa-rus	0.00 ± 0.00	−0.12 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
spa-kas	0.00 ± 0.01	−0.02 ± 0.02	−0.02 ± 0.02	0.06 ± 0.01	0.03 ± 0.01	0.06 ± 0.03	0.04 ± 0.01
heb-fra	0.03 ± 0.00	−0.29 ± 0.10	0.01 ± 0.02	0.03 ± 0.01	0.03 ± 0.00	0.03 ± 0.01	0.03 ± 0.01
hin-zho	0.03 ± 0.01	−0.37 ± 0.06	−0.01 ± 0.04	0.03 ± 0.02	0.03 ± 0.02	0.02 ± 0.03	0.02 ± 0.03
Group 7: Languages of Different Scripts, Aligned							
deu-heb	0.00 ± 0.00	−0.21 ± 0.07	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
kaz-hin	0.04 ± 0.01	−0.39 ± 0.16	0.01 ± 0.04	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.06 ± 0.01
jpn-rus	0.01 ± 0.01	−0.26 ± 0.04	0.00 ± 0.02	0.03 ± 0.01	0.03 ± 0.01	0.02 ± 0.01	0.03 ± 0.00
kor-fra	0.04 ± 0.00	−0.50 ± 0.17	0.01 ± 0.03	0.05 ± 0.01	0.05 ± 0.01	0.03 ± 0.04	0.02 ± 0.03

Table 6: R@10 Δ from baseline (XLM-R_{base}), averaged over all aligned models and grouped by loss configuration. Color coding indicates the averaged performance change magnitude: **red** for degradations, **green** for gains (adjusted for standard deviation).

Model	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
POS Tagging							
XLM-R _{de-en}	.00	.00	−.01	−.10	−.02	−.78	−.03
XLM-R _{en-kk}	−.02	−.03	−.02	−.02	−.06	−.08	−.10
XLM-R _{ja-kk}	.08	.04	.08	−.01	.08	.00	.10
XLM-R _{kk-ko}	−.01	−.02	−.02	−.01	−.08	−.01	−.07
Sentence Classification							
XLM-R _{de-en}	.03	−.05	−.03	.12	−.03	.12	−.03
XLM-R _{en-kk}	.12	−.03	.11	.18	.03	.11	.01
XLM-R _{ja-kk}	.06	.01	−.02	.13	.01	.04	.07
XLM-R _{kk-ko}	.13	.02	.01	.12	.02	.05	.02

Table 7: Cosine similarity between task gradients for POS TAGGING (**top**) and SENTENCE CLASSIFICATION (**bottom**) and respective alignment objective by model (aligned and then fine-tuned on the task). **Red** indicates negative similarity.

Lang Pair	S0T0M1	S0T1M0	S0T1M1	S1T0M0	S1T0M1	S1T1M0	S1T1M1
Group 1: Aligned Languages With English							
deu-eng	.00 ± .00	-.01 ± .00	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
kaz-eng	.00 ± .00	-.03 ± .01	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
kor-eng	.00 ± .00	-.02 ± .01	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
jpn-eng	.00 ± .00	-.02 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
Group 2: Aligned Languages Without English							
jpn-kaz	.00 ± .00	-.02 ± .01	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
kor-kaz	.00 ± .00	.01 ± .01	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
Group 3: Other Languages With English							
hin-eng	.00 ± .00	-.02 ± .01	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
zho-eng	.00 ± .00	.01 ± .01	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
spa-eng	.00 ± .01	.00 ± .00	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
rus-eng	.00 ± .00	.01 ± .01	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
Group 4: Languages of the Same Script, Non-Aligned							
lit-spa	.00 ± .00	-.01 ± .00	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
hin-kas	.00 ± .00	-.08 ± .05	.00 ± .00	.02 ± .01	.01 ± .00	.01 ± .01	.01 ± .01
spa-fra	.00 ± .00	.00 ± .00	.00 ± .00	.01 ± .00	.00 ± .00	.00 ± .00	.01 ± .00
rus-bul	.00 ± .00	.00 ± .00	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
Group 5: Languages of the Same Script, Aligned							
kaz-bul	.00 ± .00	-.02 ± .01	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
kaz-rus	.00 ± .00	-.02 ± .01	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
deu-spa	.00 ± .00	.00 ± .00	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
deu-fra	.00 ± .00	.00 ± .00	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
Group 6: Languages of Different Scripts, Non-Aligned							
spa-rus	.00 ± .00	-.01 ± .00	.00 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
spa-kas	.01 ± .00	-.10 ± .06	.00 ± .01	.03 ± .01	.02 ± .00	.02 ± .00	.02 ± .01
heb-fra	.01 ± .00	.01 ± .01	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
hin-zho	.00 ± .00	.01 ± .01	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
Group 7: Languages of Different Scripts, Aligned							
deu-heb	.01 ± .00	.01 ± .01	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
kaz-hin	.00 ± .00	.01 ± .01	.00 ± .00	.01 ± .00	.00 ± .00	.00 ± .00	.01 ± .00
jpn-rus	.00 ± .00	.01 ± .01	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00
kor-fra	.01 ± .00	.01 ± .01	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00

Table 8: Cosine similarity Δ from XLM-R_{base}, averaged over all aligned models and grouped by loss configuration. Color coding indicates the averaged similarity change magnitude: **red** for decreased similarity (accounting for standard deviation), **green** for gains from +.01 and above, black for negligible improvements or inconclusive measurements.