ALIGNFREEZE: Navigating the Impact of Realignment on the Layers of Multilingual Models Across Diverse Languages

Steve Bakos[1] **Félix Gaschi**[3] **David Guzmán**[2]
Riddhi More[1] **Kelly Chutong Li**[2] **En-Shiun Annie Lee**[1,2]
[1]Ontario Tech University, Canada [2]University of Toronto, Canada
[3]SAS Posos, France
felix@posos.co

Abstract

Realignment techniques are often employed to enhance cross-lingual transfer in multilingual language models, still, they can sometimes degrade performance in languages that differ significantly from the fine-tuned source language. This paper introduces ALIGNFREEZE, a method that freezes either the layers' lower half or upper half during realignment. Through controlled experiments on 4 tasks, 3 models, and in 35 languages, we find that realignment affects all the layers but can be the most detrimental to the lower ones. Freezing the lower layers can prevent performance degradation. Particularly, ALIGNFREEZE improves Part-of-Speech (PoS) tagging performances in languages where full realignment fails: with XLM-R, it provides improvements of more than one standard deviation in accuracy in seven more languages than full realignment.

1 Introduction

Multilingual Language Models (mLMs) like XLM-R (Conneau et al., 2020) or mBERT (Devlin et al., 2019) can perform cross-lingual transfer (Pires et al., 2019; Wu and Dredze, 2019). Once fine-tuned on a specific task in English, these models perform well on that same task when evaluated in other languages. While this can be useful for languages where fine-tuning data might be missing, cross-lingual transfer is often less efficient for languages that differ greatly from English (Pires et al., 2019), which unfortunately are the languages that would benefit the most from such ability.

With an approach similar to building multilingual word embeddings (Lample et al., 2018; Zhang et al., 2017; Artetxe et al., 2018), realignment explicitly re-trains an mLM for multilingual alignment with the hope of improving its cross-lingual transfer abilities. While some work report some level of success (Cao et al., 2020; Zhao et al., 2021; Pan et al., 2021; Wang et al., 2019), systematic evaluations show that realignment does not consistently improve cross-lingual transfer abilities and can significantly degrade them in some cases (Efimov et al., 2023; Wu and Dredze, 2020).

The relative failure of realignment raises the question of whether better multilingual alignment necessarily implies stronger cross-lingual transfer abilities. Previous work has found that mLMs have good multilingual alignment, on top of their cross-lingual transfer abilities (Dou and Neubig, 2021; Ebrahimi et al., 2023), and there even seems to be a strong link between alignment and cross-lingual transfer (Gaschi et al., 2023), although the correlation is not causation and it remains that realignment often fails.

If better alignment is linked to better cross-lingual transfer, we hypothesize that realignment has some adverse effect that induces catastrophic forgetting of other important features of the model.

To better understand this side-effect of realignment and how the different layers are affected, we propose ALIGNFREEZE. In this method, half of the model layers are frozen during realignment. With a simple controlled experiment, we compare the impact on the lower and the upper layers. We find that realignment impacts all layers, but is particularly detrimental on lower layers, namely for a low-level task like PoS tagging.

2 Background on realignment

Realignment explicitly enforces the multilingual alignment of embeddings produced by multilingual models. It trains a multilingual model to produce similar representations for corresponding words in translated sentences. Two resources are needed: a translation dataset and a word alignment tool which, in our experiments, is either FastAlign (Dyer et al., 2013), AwesomeAlign (Dou and Neubig, 2021), or a simple look-up table based on bilingual dictionaries (Lample et al., 2018) as proposed in Gaschi

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 562–586
April 29 - May 4, 2025

et al. (2023).

In our experiments, we use the realignment method proposed by Wu and Dredze (2020), where a contrastive loss maximizes the similarity between the representations of a pair of corresponding words (h and aligned(h)) compared to all other possible pairs of words in a batch ($\mathcal{H}$ of size B) of pairs of translated sentences:

$$\mathcal{L}(\theta) = \frac{1}{2B} \sum_{h \in \mathcal{H}} \log \frac{\exp(\text{sim}(h, \text{aligned}(h))/T)}{\sum_{h' \in \mathcal{H}, h' \neq h} \exp(\text{sim}(h, h')/T)} \quad (1)$$

T is the temperature, a hyperparameter set to 0.1.

3 Methodology

We introduce ALIGNFREEZE, a realignment method that relies on partial freezing to preserve half of the weights of an mLM during realignment. Because full realignment was shown not to work consistently (Wu and Dredze, 2020), we hypothesize that applying realignment on the whole model could trigger some catastrophic forgetting of information useful to downstream cross-lingual tasks. To help mitigate that and better understand the impact of realignment, ALIGNFREEZE freezes half of the layers of the mLM during realignment only.

Freezing Strategies For the sake of simplicity and to reduce the number of experimental runs, we work with only two freezing strategies: 1) *Front-freezing*, which freezes the lower-half layers while the remaining layers are realigned; and 2) *Back-freezing*, which freezes upper-half layers instead.

Assuming that basic linguistic features are encoded in the lower layers while the top ones retain higher-level information (Peters et al., 2018), *Front-freezing* aims to preserve the foundational language understanding captured in the early layers while enabling task-specific adaptation in the later layers. *Back-freezing* seeks to maintain the abstract, high-level representations developed in the deeper layers while fine-tuning the model's basic linguistic features. Our approach intentionally employs a straightforward freezing strategy, not to establish a new state-of-the-art realignment method, but to better understand the conditions under which realignment fails and how to mitigate its failure.

The freezing is applied only during realignment. Thus, ALIGNFREEZE can be described with the following steps: 1) Take a multilingual Language Model (mLM), 2) Freeze half of its layers, 3) train the remaining weights for the realignment loss, 4) unfreeze the frozen layers, 5) perform fine-tuning on the whole model for cross-lingual transfer.

4 Experiment Setup

	Parameters	Values
AlignFreeze	**Freezing Strategies**	no freezing (full), Front Half, Back Half
	Word Alignment Methods	FastAlign (Dyer et al., 2013), AwesomeAlign (Dou and Neubig, 2021), Bilingual Dictionaries (Lample et al., 2018)
Settings	**Tasks**	PoS tagging (34 lang.), NER (34 lang.), NLI (12 lang.)
	Datasets	UD-PoS, NER, XNLI
	Baseline Models	XLM-R, DistilMBERT

Table 1: Summary of the experimental setting.

Datasets *Realignment Dataset*: We use the OPUS-100 dataset (Zhang et al., 2020) for the realignment phase. OPUS-100 is a multilingual parallel corpus that includes sentence pairs across multiple languages.
Downstream Task Dataset: We evaluate multilingual models on three tasks: PoS tagging, Named Entity Recognition (NER), Natural Language Inference (NLI), and Question Answering (QA). For PoS tagging, we use the Universal Dependencies dataset (Zeman et al., 2020), which provides annotated treebanks for a wide range of languages. For NER, we use the WikiANN dataset (Rahimi et al., 2019). For NLI, we use the Cross-lingual Natural Language inference (XNLI) corpus (Conneau et al., 2018). For QA, we use the XQuAD dataset (Artetxe et al., 2020).

Models Following Gaschi et al. (2023), we work with three models: DistilMBERT (Sanh et al., 2019), mBERT (Devlin et al., 2019), and XLM-R Base (Conneau et al., 2020). DistilMBERT is a smaller version of mBERT (Devlin et al., 2019) obtained through distillation (Sanh et al., 2019). DistilMBERT, mBERT, and XLM-R are all Transformer-based masked multilingual models.

Languages We use English as the source language for fine-tuning. We evaluate on 34 languages for PoS-tagging and NER,, 12 for NLI, and 11 for QA. For realignment, we use the 34 available languages for PoS tagging, NER,, NLI, and QA. Using the same setting allows for comparison of results across tasks and also improves the outcome (cf. Appendix C.2). We use all the languages that our resources allow: every language must be present in the translation dataset, the bilingual dictionaries, and one of the downstream datasets. The full list can be found in the subsection B.1.

Further details about the implementation can be found in Appendix B and in the source code[1].

5 Results and Discussion

Finding 1: Full realignment fails in many cases. As already observed by previous work (Wu and Dredze, 2020; Efimov et al., 2023; Gaschi et al., 2023), full realignment isn't always successful. Table 2 shows that realignment provides, on average, a significant improvement over fine-tuning with DistilMBERT, but the improvement is smaller with mBERT and even more so with XLM-R, especially for NLI and QA where it even degrades the results. Figure 1 and Table 2 also show that the outcome of full realignment varies a lot by language. For PoS-tagging with mBERT and distilMBERT, the majority of languages see a significant increase in accuracy. But with XLM-R, only 11 see a significant increase and one (Farsi) even undergoes a significant decrease of 2 points. For NLI, full realignment fails almost systematically with XLM-R, since 8 languages over 12 see a significant decrease in accuracy with realignment, while there can be as many significant increases and decreases for NER with XLM-R.

Finding 2: ALIGNFREEZE (front) mitigates some of the failures of realignment. Freezing the lower layers during realignment often improves results for cases where full realignment fails. Table 2 shows that it brings an average improvement over full realignment with XLM-R for PoS-tagging and NLI, with 0.4 percent increases for both, but not for NER or QA, although the standard deviation is higher for QA making the results less conclusive. But more importantly, for PoS tagging, all languages are positively or neutrally impacted by front-freezing. And with XLM-R, the improvement is significant for 7 more languages than full realignment. On Figure 1, while Farsi (fa) and Hebrew (he) undergo a significant decrease with full realignment for PoS tagging, they do not with ALIGNFREEZE and even benefit from a 1-point improvement in the case of Hebrew. There are other languages, like Slovakian (sk), Polish (pl), and Hindi (hi) where full realignment provides a smaller improvement than front-freezing. Similarly to PoS tagging, front-freezing with mBERT for NER reduces the number of languages that suffer from realignment (from 19 to 1), but this is not the case with XLM-R. Contrary to PoS tagging and NER, NLI and QA do not benefit much from realignment, but front-freezing allows to reduce the number of languages for which realignment is detrimental for NLI.

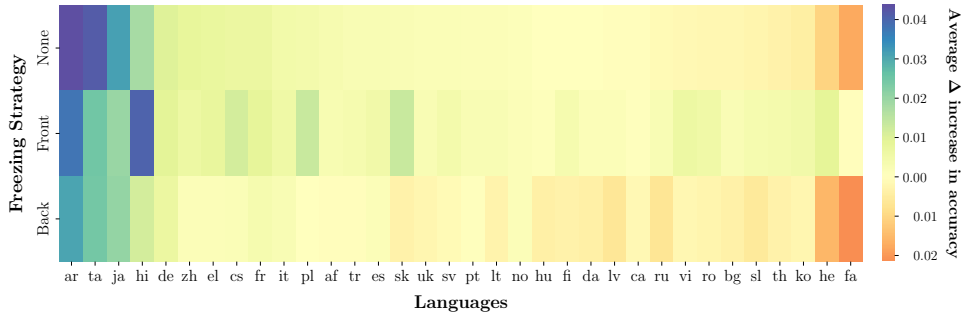
Finding 3: Realignment impacts the entire model, but it seems detrimental to the lower layers while it can be beneficial to the upper ones. Front-freezing can mitigate some failure cases of full realignment, thus realignment can have a detrimental effect on the lower layers. On the other hand, back-freezing seems to have a less important impact on realignment. Table 2 shows that back-freezing does not significantly improve over full realignment, and Figure 1 suggests that it provides worse results than any other alignment method for PoS tagging and NLI. The only exception is QA, for which back-freezing seems to improve over full realignment for distilMBERT and mBERT, but this improvement is not significant compared to the high variance of the results. This contradicts Gaschi et al. (2023) who hypothesized that since realignment appears to work better on smaller models, realignment might only have an impact on the upper layers of the model. Our results show that realignment impacts all layers and seems to be the most detrimental to the lower ones.

5.1 Generalized Recommendations for Practitioners using ALIGNFREEZE

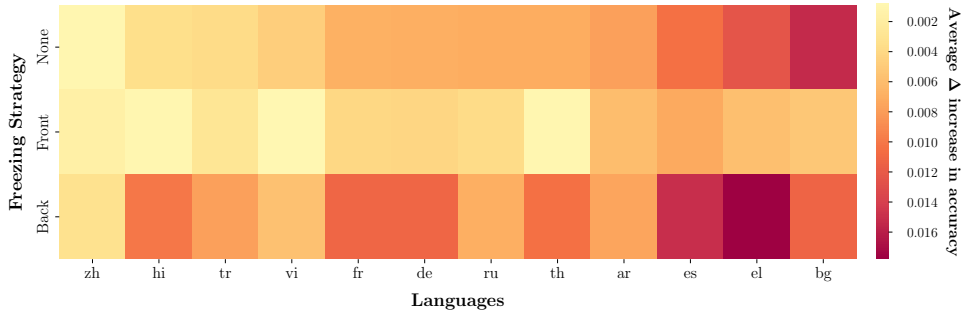
Full realignment should be used for smaller models and low-level tasks. As already suggested by previous work (Gaschi et al., 2023), full realignment works better for smaller models like DistilMBERT and the technique proves beneficial for tasks involving lower-level linguistic features, as evidenced by more consistent improvements in PoS tagging, compared to NLI QA, or even NER (Table 2). This finding is relevant for researchers and organizations facing computational constraints. ALIGNFREEZE and full realignment enable the enhancement of smaller, resource-efficient models, achieving competitive results without large-scale models or extensive computational resources.

ALIGNFREEZE improves upon full realignment for PoS-tagging. Table 2 shows that ALIGNFREEZE is never detrimental to cross-lingual transfer and improves results for more languages than full realignment. For NLI, while ALIGNFREEZE still provides better results than full realignment,

[1] https://github.com/posos-tech/multilingual-alignment-and-transfer/tree/main/scripts/2025_naacl



(a) Variation of the accuracy with realignment with XLM-R Base for the PoS tagging task.



(b) Variation of the accuracy with realignment with XLM-R Base for the NLI task.

Figure 1: Variation of the accuracies with realignment with XLM-R Base for the PoS tagging and NLI tasks. Languages are sorted by the improvement brought by full realignment. The average increase in accuracy is computed over 5 runs. Numerical values and results for other models can be found in Appendix C.

	PoS (34 lang.)			**NER** (34 lang.)			**NLI** (12 lang.)			**QA** (11 lang.)			Total (91)	
	acc.	#↓	#↑	acc.	#↓	#↑	acc.	#↓	#↑	F1	#↓	#↑	#↓	#↑
DistilMBERT														
Fine-tuning Only	$73.8_{\pm 0.6}$	-	-	$82.5_{\pm 0.3}$	-	-	$60.1_{\pm 0.3}$	-	-	$38.1_{\pm 0.6}$	-	-	-	-
Full realignment	$77.6_{\pm 0.3}$	0	31	$84.7_{\pm 0.2}$	3	21	$61.6_{\pm 0.2}$	3	5	$39.3_{\pm 1.2}$	2	5	8	62
ALIGNFREEZE (front)	$76.2_{\pm 0.2}$	0	34	$84.0_{\pm 0.5}$	1	21	$61.6_{\pm 0.1}$	1	8	$37.4_{\pm 0.8}$	4	2	6	65
ALIGNFREEZE (back)	$77.4_{\pm 0.1}$	0	30	$83.7_{\pm 0.7}$	4	17	$61.9_{\pm 0.2}$	1	6	$39.1_{\pm 1.0}$	2	5	7	58
mBERT														
Fine-tuning Only	$77.0_{\pm 0.5}$	-	-	$85.7_{\pm 0.3}$	-	-	$66.3_{\pm 0.6}$	-	-	$57.1_{\pm 0.4}$	-	-	-	-
Full realignment	$79.6_{\pm 0.4}$	1	32	$86.4_{\pm 0.3}$	19	4	$67.4_{\pm 0.4}$	0	8	$52.9_{\pm 0.7}$	11	0	31	44
ALIGNFREEZE (front)	$79.2_{\pm 0.2}$	0	32	$\mathbf{86.7}_{\pm 0.2}$	1	6	$67.7_{\pm 0.2}$	0	10	$55.3_{\pm 0.7}$	9	0	10	48
ALIGNFREEZE (back)	$79.3_{\pm 0.3}$	1	30	$86.5_{\pm 0.6}$	12	6	$67.5_{\pm 0.3}$	0	10	$53.7_{\pm 0.6}$	11	0	24	46
XLM-R Base														
Fine-tuning Only	$80.9_{\pm 0.1}$	-	-	$84.9_{\pm 0.4}$	-	-	$\mathbf{73.9}_{\pm 0.2}$	-	-	$\mathbf{61.2}_{\pm 0.4}$	-	-	-	-
Full realignment	$81.3_{\pm 0.1}$	1	11	$85.3_{\pm 0.2}$	8	8	$73.2_{\pm 0.2}$	8	0	$59.4_{\pm 0.7}$	10	0	27	19
ALIGNFREEZE (front)	$\mathbf{81.7}_{\pm 0.2}$	0	18	$84.8_{\pm 0.3}$	11	4	$73.6_{\pm 0.2}$	6	0	$59.1_{\pm 0.5}$	10	0	27	22
ALIGNFREEZE (back)	$80.9_{\pm 0.2}$	7	4	$84.9_{\pm 0.1}$	13	7	$72.9_{\pm 0.3}$	11	0	$58.0_{\pm 1.1}$	11	0	42	11
Total of #↓ and #↑ by task		/102			/102			/36			/33		/273	
Full realignment	-	2	74	-	30	33	-	11	13	-	6	6	64	125
ALIGNFREEZE (front)	-	0	84	-	13	31	-	7	18	-	9	2	43	135
ALIGNFREEZE (back)	-	8	64	-	29	30	-	12	16	-	11	10	73	115

Table 2: Average accuracy of all target languages for PoS tagging, NER, and XNLI with all models and realignment approaches. The number of languages for which realignment provides an increase above one standard deviation is reported (#↑) as well as the number of languages for which it provides a decrease of more than one standard deviation (#↓), the remaining languages see no significant change. The results shown are for the bilingual dictionary aligner. Results are averaged over five runs. ± indicates the standard deviation.

it can still be detrimental to cross-lingual transfer in some languages. This suggests ALIGNFREEZE is most effective when applied to tasks relying on syntactic and morphological information preserved in the frozen layers.

Cross-lingual transfer is hard to predict The variability in effectiveness across languages, models, and tasks highlights the importance of tailored approaches in multilingual NLP. In a truly zero-shot context, it seems hard to determine the right method for cross-lingual transfer, as shown by our

results and previous work (Schmidt et al., 2023; Yarmohammadi et al., 2021). If evaluation data is available in the target language, practitioners should try all methods available to improve cross-lingual transfer, as results vary a lot by setting.

6 Conclusion

This study introduces ALIGNFREEZE, a method using partial freezing to improve cross-lingual transfer in multilingual language models. Our experiments demonstrate that ALIGNFREEZE effectively mitigates the failure cases of partial realignment by preserving pre-trained knowledge in the lower layers.

When it comes to cross-lingual transfer, there does not seem to be any "silver bullet" (Yarmohammadi et al., 2021) method that works for all languages, models, and tasks. Like realignment itself, and other cross-lingual approaches, ALIGNFREEZE can help for some situations but not others. ALIGNFREEZE can at least be useful for cross-lingual PoS-tagging with XLM-R.

ALIGNFREEZE helps better understand how realignment works. It impacts all layers and can be most detrimental to the lower ones, which is more visible on low-level tasks like PoS-tagging, that might be encoded in lower layers (Peters et al., 2018). Realignment probably fails simply because it is applied to the whole model without hindrance, which explains ALIGNFREEZE relative success but also the results of other methods based on adapters like MAD-X (Pfeiffer et al., 2020).

7 Ethics and Limitations

7.1 Limitations

We worked with the languages available in the datasets we used, but this led to high-resource languages and European languages being over-represented. To evaluate the effectiveness of cross-lingual transfer and realignment, the accuracy was averaged over all languages for a given task and model. Using the average to analyze the results has its risks, as different sets of languages can then potentially lead to different conclusions. However, the average remains convenient for our analysis and it was completed with some language-wise analysis as in Figures 1b and 1a. Moreover, detailed results are provided in Appendix C.5 for the interested reader.

The experiments of this paper could be extended to more tasks and more models. PoS tagging, NER,, NLI, and QA were chosen for their differences. PoS tagging is a more low-level task looking at word categories while NLI deals with understanding. Moreover, partial realignment works well for PoS tagging, whereas it provides weaker results with NLI (Gaschi et al., 2023). NER is chosen to complement this analysis with a task that is word-level, like PoS tagging, and semantic, like NLI. QA is chosen because it is a more difficult semantic tasks, like NLI, but is also a word-level one, like NER and PoS-tagging. The choice of model was based on a similar approach. XLM-R Base is the largest mLM that we could train with our experimental setting while DistilMBERT offered a smaller alternative, and mBERT some middle ground. XLM-R was shown not to benefit too much from realignment, while DistilMBERT observes a large performance increase and can sometimes match XLM-R with the help of realignment (Gaschi et al., 2023).

Throughout this paper, realignment is applied to encoder-only Language Models like DistilMBERT or XLM-R. While the literature on realignment also focuses on encoders (Cao et al., 2020; Zhao et al., 2021; Efimov et al., 2023; Wu and Dredze, 2020), realignment could be extended to more recent decoder-only generative multilingual models like Bloom (Scao et al., 2023) or XGLM (Lin et al., 2022). However, these models are often intended to be used in a zero-shot or few-shot fashion, and Ahuja et al. (2023) showed that cross-lingual transfer with fine-tuning of XLM-R largely outperforms prompt-based approaches with generative models on classification tasks.

This study experiments only with two simple freezing strategies: front-freezing and back-freezing. More granular freezing strategies could be designed to better understand the role of each layer. However, we experimented with several other approaches, but the results were not conclusive enough to include in the paper. Freezing half of the model does influence realignment, though the overall impact is already relatively minor. More granular freezing strategies led to even smaller variations (See Appendix C.3 for some results).

Some languages seem to benefit more from realignment than others. This study shows that freezing the bottom half of the layers during realignment might help with some languages that do not benefit from full realignment. However, ALIGNFREEZE, like full realignment, does not work for all languages, and it is still hard to determine in advance

which language will benefit or not from realignment. This issue can be explored through a regression analysis of our realignment results, but the regressor we trained overfitted on language-specific features and wasn't generalizing across languages, which defeats its purpose (cf. Appendix C.4). Further research is needed to better understand what makes realignment fail under some conditions and succeed in others, but it might need larger-scale experiments to get conclusive results.

7.2 Ethics statement

The resources we relied on limited our choice of languages. While working with 35 languages in total, this work contributes to the overexposure of European languages in the scientific literature. However, our work demonstrates that realignment can have a very different impact depending on the language and proposes new ways to improve cross-lingual transfer. While our conclusions will not directly impact the speakers of low-resource languages, they pave the way for potentially useful applications.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

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

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

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Abteen Ebrahimi, Arya D. McCarthy, Arturo Oncevay, John E. Ortega, Luis Chiruzzo, Gustavo Giménez-Lugo, Rolando Coto-Solano, and Katharina Kann. 2023. Meeting the needs of low-resource languages: The value of automatic alignments via pretrained models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3912–3926, Dubrovnik, Croatia. Association for Computational Linguistics.

Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. 2023. *The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer*, page 51–67. Springer Nature Switzerland.

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.

Félix Gaschi, François Plesse, Parisa Rastin, and Yannick Toussaint. 2022. Multilingual transformer encoders: a word-level task-agnostic evaluation. *Preprint*, arXiv:2207.09076.

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

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. 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.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023. One for all & all for one: Bypassing hyperparameter tuning with model averaging for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12186–12193, Singapore. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

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.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Kennedy Ajede Chika, et al. 2020. Universal dependencies 2.6.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

A Related Works

Pre-trained multilingual language models have become the predominant approach for cross-lingual transfer tasks. Word alignment methods that depend on these models have also been proposed (Jalili Sabet et al., 2020; Nagata et al., 2020). Current realignment methods are typically applied to a multilingual pre-trained model before fine-tuning in a single language (usually English) and applying to other languages on tasks such as Natural Language Inference (NLI) (Conneau et al., 2018), Named Entity Recognition (NER) (Rahimi et al., 2019), Part-of-speech tagging (PoS) (Zeman et al., 2020), or Question Answering (QA) (Artetxe et al., 2020). This process is intended to enhance the model's ability to generalize to other languages for these tasks.

Realignment can be performed in different ways. Cao et al. (2020) minimizes the l2 distance between translated pairs. But some regularization is needed to prevent the representations from collapsing, which can be done through an additional loss term (Cao et al., 2020; Zhao et al., 2021) or using contrastive learning (Wu and Dredze, 2020). Since the alignment is done at the word level between contextualized representations, an alignment tool is needed to obtain translated pairs to realign. Most methods employ the statistical tool FastAlign (Dyer et al., 2013). However neural-based tools can be used like AwesomeAlign (Dou and Neubig, 2021), which are indeed shown to work better for low-resource languages, although they come at a larger computational cost (Ebrahimi et al., 2023). A bilingual dictionary can also be used as a look-up table but extracts fewer pairs of words (Gaschi et al., 2023). Empirically, it was however shown that realignment has inconsistent results when evaluated across several tasks and languages (Efimov et al., 2023; Wu and Dredze, 2020).

The failure of realignment questions the very link between multilingual alignment and cross-lingual transfer (Gaschi et al., 2022). Realignment can increase multilingual alignment, but it might also be detrimental to some monolingual or even multilingual features learned by the model. To alleviate this, Gaschi et al. (2023) tried to optimize the realignment loss jointly with the fine-tuning loss, but they did not report improved performances.

Due to its black-box nature, it is not straightforward to determine what role each layer of an mLM plays, but Peters et al. (2018) empirically showed, for ELMo, that the lower layers might encapsulate more lower-level information like syntax while the top ones relate to semantics. In a multilingual setting, Wu and Dredze (2019) showed that freezing the lower layers of mBERT during fine-tuning can increase its cross-lingual performances.

B Additional Experimental details

B.1 Languages

For PoS tagging and NER, because we used languages that were available simultaneously in the dataset but also in the different resources used for that task (bilingual dictionaries and the translation dataset), we worked with the following 34 languages: Afrikaans, Arabic, Bulgarian, Catalan, Chinese, Czech, Danish, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Italian, Japanese, Korean, Latvian, Lithuanian, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Tamil, Thai, Turkish, Ukrainian, and Vietnamese.

For NLI, due to similar constraints, we worked with the following 12 languages: Arabic, Bulgarian, Chinese, French, German, Greek, Hindi, Russian, Spanish, Thai, Turkish, and Vietnamese.

B.2 Model Settings

For both experiments, we reused the experimental setup from Gaschi et al. (2023). All experiments were run with 5 random seeds and performed using Nvidia A40 GPUs.
We train up to 5 epochs for PoS-tagging and NER and 2 epochs for NLI, with a learning rate of 2e-5, batch size of 32 for training and evaluation, and a maximum length of 200 for the source and target. For realignment, we use a maximum length of 96 and a batch size of 16.

B.3 Word alignment tools

We employ three word alignment methods: FastAlign (Dyer et al., 2013), AwesomeAlign (Dou and Neubig, 2021), and Bilingual Dictionaries (Lample et al., 2018). From a translation dataset, pairs were extracted either using a bilingual dictionary, following Gaschi et al. (2022), with FastAlign or AwesomeAlign. For FastAlign, alignments were generated in both directions and then symmetrized using the grow-diag-final-and heuristic provided by FastAlign, following Wu and Dredze (2020). In all extraction methods, only one-to-one alignments were retained, and trivial cases where both words

	PoS-tagging	NLI	NER	QA
train (en)	12,570	392,702	20,029	288,132
Afrikaans	425	-	1,002	-
Arabic	856	5010	10,000	4,317
Bulgarian	1,117	5010	10,005	
Catalan	1,863	-	10,001	-
Chinese	501	5010	10,378	3,831
Czech	10,163	-	10,001	-
Danish	565	-	10,000	-
Finnish	1,000	-	10,000	-
French	416	5010	10,000	-
German	977	5010	10,000	3,405
Greek	478	5010	10,001	7,035
Hebrew	509	-	10,000	-
Hindi	1,685	5010	1,000	5,195
Hungarian	451	-	10,004	-
Italian	485	-	10,000	-
Japanese	546	-	11,724	-
Korean	989	-	10,002	-
Latvian	1,828	-	10,002	-
Lithuanian	687	-	10,000	-
Norwegian	1,939	-	10,000	-
Persian	1,456	-	10,000	-
Polish	2,218	-	10,018	-
Portuguese	1,208	-	10,002	-
Romanian	734	-	10,000	4,174
Russian	612	5010	10,000	4,109
Slovak	1,061	-	10,001	-
Slovenian	790	-	10,018	-
Spanish	429	5010	10,000	3,391
Swedish	1,000	-	10,000	-
Tamil	125	-	1,000	-
Thai	1,031	5010	13,125	11,093
Turkish	1,000	5010	10,001	3,839
Ukrainian	915	-	10,000	-
Vietnamese	800	5010	10,000	3,550

Table 3: Size of the datasets (in number of samples) in the Universal Dependencies, NLI, NER, and QA tasks.

were identical were discarded, also following Wu and Dredze (2020).

We use the three aligners for PoS tagging, but only the bilingual dictionaries for NLI, QA, and NER, because it takes longer to train on NLI than PoS tagging and to avoid performing too many unnecessary experiments. The approach based on bilingual dictionaries is preferred, as it is the aligner that provided the best results in Gaschi et al. (2023). Ultimately, the main part of the paper only reports the results with the bilingual dictionary, results with other aligners for PoS tagging are left at the end of the Appendix for the interested reader but do not impact our conclusions.

B.4 Statistics about the datasets used

The size of the datasets used for training and evaluating are reported in Table 3.

B.5 Scientific artefacts used

Here is a list of the scientific artifacts used[2]:

- The code for realignment comes from Gaschi et al. (2023) and has MIT License
- the weights of DistilMBERT (Sanh et al., 2019) have License Apache-2.0
- the weights of XLM-R Base (Conneau et al., 2020) have MIT License
- The OPUS-100 dataset (Zhang et al., 2020) does not have a known license, but it is a filtering of the OPUS corpus (Tiedemann, 2009) which is itself the compilation of many translation datasets which are, to the best of our knowledge, free to be redistributed.
- The Universal Dependencies dataset (Zeman et al., 2020) is also a compilation of several datasets, which all have, to the best of our knowledge, open-source licenses.
- The XNLI corpus (Conneau et al., 2018) has a dedicated license but is nevertheless freely available for "typical machine learning use", which is the case in this paper.
- The WikiANN dataset (Rahimi et al., 2019) doesn't have a known license to the best of our knowledge. It is thus assumed to be free to use.
- The XQuAD dataset (Artetxe et al., 2020) has a the License CC-BY-SA-4.0, which allows its usage.
- FastAlign (Dyer et al., 2013) has Apache-2.0 license
- AWESOME-align (Dou and Neubig, 2021) has BSD 3-Clause License
- The bilingual dictionaries (Lample et al., 2018) have an "Attribution-NonCommercial 4.0 International" license that allows non-commercial use as is the case here

The scientific artifacts were thus used consistently with the intended use, as all identified licenses are open-source or authorize non-commercial use.

[2]It does not include all the resources that are leveraged by those artifacts like specific Python packages.

We cannot guarantee that the data we use do not contain personally identifying information or offensive content. However, this paper is not redistributing the data in any way and is simply using it for experiments. Nevertheless, we looked at randomly sampled elements of our datasets to verify their relevance and did not find any offensive or identifying content.

C Additional Results

C.1 Filtering data does not improve results

We hypothesized a direct correlation between the quality of the realignment results on the downstream tasks and the quality of the OPUS-100 dataset. To evaluate this, we employed a Quality Estimation (QE) model (Rei et al., 2022) to selectively filter out sentence pairs below a predefined quality threshold. Since the OPUS-100 dataset contains significantly more sentences than needed for the realignment steps, the filtering should not affect the amount of data seen during realignment. Subsequently, we conducted experiments using this curated dataset to assess the impact of data quality on realignment results on the downstream tasks. Contrary to expectations, Figure 2 shows that, on average, using a higher quality dataset filtered by a QE model has little impact on the final results.

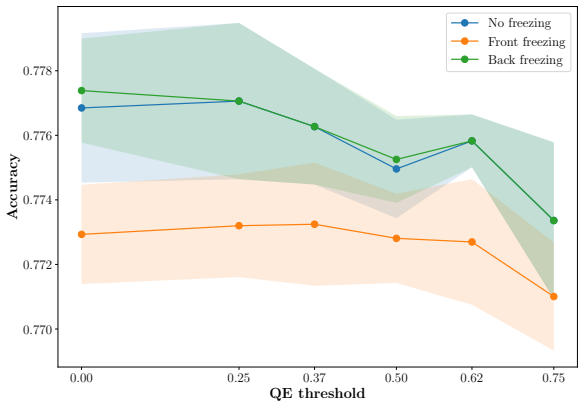


Figure 2: Average accuracy for DistilMBERT when filtering the dataset for different percentiles of QE for the PoS tagging task.

C.2 Discussion on the amount of languages in realignment

In this paper, realignment is performed with 34 languages for all tasks, despite the downstream evaluation being possible in only 12 of those languages for NLI. In preliminary experiments, realignment was only performed on those 12 languages for NLI, and the whole set of 34 languages was used for PoS tagging and NER. However, we eventually chose to use the same realignment step for both tasks, for a more controlled experiment, which means that we used 34 languages for NLI. As Table 4 shows, realigning on 34 languages provides better results for all models except DistilMBERT.

	12 languages			34 languages		
	acc.	#↓	#↑	acc.	#↓	#↑
DistilMBERT						
Fine-tuning Only	$60.1_{\pm 0.3}$	-	-	$60.1_{\pm 0.3}$	-	-
Full realignment	$63.1_{\pm 0.2}$	1	9	$61.6_{\pm 0.2}$	3	5
ALIGNFREEZE (front)	$62.7_{\pm 0.3}$	0	11	$61.6_{\pm 0.1}$	1	8
ALIGNFREEZE (back)	$63.1_{\pm 0.2}$	0	10	$61.9_{\pm 0.2}$	1	6
mBERT						
Fine-tuning Only	$66.3_{\pm 0.6}$	-	-	$66.3_{\pm 0.6}$	-	-
Full realignment	$66.9_{\pm 0.7}$	0	4	$67.4_{\pm 0.4}$	0	8
ALIGNFREEZE (front)	$66.7_{\pm 0.4}$	0	2	$67.7_{\pm 0.2}$	0	10
ALIGNFREEZE (back)	$67.0_{\pm 0.7}$	0	4	$67.5_{\pm 0.3}$	0	10
XLM-R Base						
Fine-tuning Only	$\mathbf{73.9}_{\pm 0.2}$	-	-	$\mathbf{73.9}_{\pm 0.2}$	-	-
Full realignment	$72.9_{\pm 0.1}$	11	0	$73.2_{\pm 0.2}$	8	0
ALIGNFREEZE (front)	$73.4_{\pm 0.1}$	9	0	$73.6_{\pm 0.2}$	6	0
ALIGNFREEZE (back)	$73.2_{\pm 0.3}$	11	0	$72.9_{\pm 0.3}$	11	0

Table 4: Results of various realignment methods on NLI when using either 12 or 34 languages when performing realignment.

The evidence may be too anecdotal to conclude that using more languages for realignment generally provides better results. It might depend greatly on the alignment method used. Because we use an in-batch contrastive loss, adding languages increases diversity in the batch which might help the realignment work better. More extensive experiments in that regard are left for future work.

C.3 Additional results with more granular methods

Table 5 shows the results of more granular strategies applied to PoS-tagging with DistilMBERT. While this combination and task and model is the one for which we observe the larger improvement with realignment, we do not observe any significantly interesting pattern for more granular freezing strategies. We tested two types of strategies: (1) freezing all layers except one during realignment (middle section of the table) and (2) freezing only one layer during realignment (bottom section of the table). While the first scenario shows some variation across layers, the number of languages that significantly benefit from these realignment strategies is lower than full realignment or front-freezing. For single-layer freezing, there isn't much variation across layers, and the results are very close to full realignment. This can be explained by the fact that

by freezing only a single layer, we are not making as much as a difference from full realignment than when freezing half of the model.

	PoS-tagging (34 lang.)		
	acc.	#↓	#↑
Baselines			
Fine-tuning Only	$73.8_{\pm 0.6}$	-	-
Full realignment	$77.6_{\pm 0.3}$	0	31
ALIGNFREEZE (front)	$76.2_{\pm 0.2}$	0	**34**
ALIGNFREEZE (back)	$77.4_{\pm 0.1}$	0	30
Single-layer realignment			
Layer 0	$75.0_{\pm 0.3}$	0	18
Layer 1	$76.5_{\pm 0.2}$	1	25
Layer 2	$76.5_{\pm 0.2}$	1	25
Layer 3	$76.3_{\pm 0.3}$	0	24
Layer 4	$76.4_{\pm 0.3}$	0	29
Layer 5	$75.6_{\pm 0.2}$	0	27
Layer 6	$73.6_{\pm 0.3}$	4	1
Single-layer freezing			
Layer 0	$77.7_{\pm 0.2}$	0	30
Layer 1	$77.5_{\pm 0.2}$	0	31
Layer 2	$77.7_{\pm 0.1}$	0	31
Layer 3	$\mathbf{77.8}_{\pm 0.2}$	0	32
Layer 4	$77.5_{\pm 0.1}$	0	29
Layer 5	$77.7_{\pm 0.2}$	0	30
Layer 6	$77.7_{\pm 0.2}$	0	30

Table 5: Average accuracy of all target languages for PoS-tagging for distilMBERT with more granular freezing strategies. Refer to Table 2 for more details on the notations.

C.4 Realignment performance prediction

Some languages seem to benefit more than others from realignment. We performed a regression analysis using a random forest classifier to predict the ability to perform cross-lingual transfer from language-related and realignment-related features.

Prediction target : the target variable for our regression model was the change in the model's accuracy with and without realignment for a given language. In other words, we compare the cross-lingual accuracy in a given language with and without realignment.

Input features : as input features, we used various categorical features indicating the realignment method used: the aligner used (Fastalign, AWESOME-align, or bilingual dictionary), the freeze location (front or back freezing), and the freezing status (whether there is or isn't freezing). The language-related features are lang2vec distances from English (Littell et al., 2017) (featural, syntactic, genetic, inventory, geographic, and phonological), word order, script type, and the language itself.

feature	importance
Lang2vec distance	0.546
Language	0.251
Script type	0.077
Freeze location	0.053
Aligner	0.053
Freezing status	0.011
Word order	0.008

Table 6: Feature importance of various features of the random forest regressor applied to realignment results.

The random forest uses 30 estimators, with warm-start, bootstrapping, and the mean squared error as the splitting criterion. We perform the regression on the realignment results with Full realignment and ALIGNFREEZE (front and back) for PoS-tagging with distilmBERT, because it is the configuration for which we have the higher variance in results and the larger amount of data points (all aligners were used). We also remove outliers using interquantile range method (IQR).

The fitted regressor has an R^2 score of 0.7126 and a mean squared error of 0.0001. The features' importance, aggregated by categories, is reported in Table 6. While it seems that the lang2vec distances with English can largely help predict the effectiveness of realignment, this regression analysis has many limitations. First of all, while the R^2 score is adequate, attempts at generalizing the regressor to unseen languages provided poor results. The issue probably is that there aren't enough data points compared to the number of input features. The regressor overfits on language-related features because the language itself is a good predictor of the accuracy since results do not vary a lot across different seeds of realignment methods.

In conclusion, realignment appears more effective for languages distant from English. However, since our regressor doesn't fully generalize to unseen languages, these findings should be interpreted with caution. We believe that additional data points are needed to draw more definitive conclusions, as the experiments in this paper provide a limited dataset.

C.5 Full Results

This section contains the detailed results of the experiments of this paper:

- Realignment results for PoS tagging with DistilMBERT in Table 7

- Realignment results for NER with DistilMBERT in Table 8
- Realignment results for NLI with DistilMBERT in Table 9
- Realignment results for QA with DistilMBERT in Table 10
- Realignment results for PoS tagging with mBERT in table 11
- Realignment results for NER with mBERT in Table 12
- Realignment results for NLI with mBERT in table 13
- Realignment results for QA with mBERT in Table 14
- Realignment results for PoS tagging with XLM-R in Table 15
- Realignment results for NER with XLM-R in Table 16
- Realignment results for NLI with XLM-R in Table 17
- Realignment results for QA with XLM-R in Table 18
- Results of filtering for different percentiles of QE for NLI with DistilMBERT in Table 19
- Results of filtering for different percentiles of QE for PoS tagging with DistilMBERT and FastAlign aligner in Table 20
- Results of filtering for different percentiles of QE for PoS tagging with DistilMBERT and AwesomeAlign aligner in Table 21
- Results of filtering for different percentiles of QE for PoS tagging with DistilMBERT and bilingual dictionary aligner in Table 22
- Results of single-layer realignment for PoS tagging with DistilMBERT and bilingual dictionary aligner in Table 23
- Results of single-layer freezing for PoS tagging with DistilMBERT and bilingual dictionary aligner in Table 24

	FT Only	**vanilla realignment**			**ALIGNFREEZE with front-freezing**			**ALIGNFREEZE with back-freezing**		
	-	FA	AA	BD	FA	AA	BD	FA	AA	BD
Afrikaans	85.5$_{\pm 0.2}$	**86.4**$_{\pm 0.3}$	**86.4**$_{\pm 0.3}$	85.6$_{\pm 0.4}$	86.2$_{\pm 0.2}$	86.3$_{\pm 0.3}$	86.1$_{\pm 0.3}$	86.0$_{\pm 0.2}$	86.0$_{\pm 0.3}$	85.4$_{\pm 0.1}$
Arabic	51.7$_{\pm 1.7}$	63.9$_{\pm 0.5}$	63.6$_{\pm 0.3}$	**66.6**$_{\pm 0.5}$	63.3$_{\pm 0.5}$	63.0$_{\pm 0.5}$	65.0$_{\pm 0.6}$	63.5$_{\pm 0.6}$	62.8$_{\pm 0.7}$	65.3$_{\pm 0.3}$
Bulgarian	85.0$_{\pm 0.5}$	87.4$_{\pm 0.2}$	**87.6**$_{\pm 0.3}$	**87.6**$_{\pm 0.4}$	87.1$_{\pm 0.3}$	87.3$_{\pm 0.2}$	87.2$_{\pm 0.3}$	87.2$_{\pm 0.2}$	**87.6**$_{\pm 0.2}$	87.5$_{\pm 0.2}$
Catalan	86.6$_{\pm 0.4}$	87.8$_{\pm 0.2}$	88.1$_{\pm 0.2}$	**88.4**$_{\pm 0.1}$	87.6$_{\pm 0.3}$	87.8$_{\pm 0.2}$	88.2$_{\pm 0.1}$	87.9$_{\pm 0.2}$	88.2$_{\pm 0.2}$	88.1$_{\pm 0.2}$
Chinese	64.3$_{\pm 1.4}$	66.2$_{\pm 0.5}$	66.3$_{\pm 0.6}$	**67.4**$_{\pm 0.7}$	66.6$_{\pm 0.5}$	66.3$_{\pm 0.4}$	67.3$_{\pm 0.6}$	66.2$_{\pm 0.7}$	66.3$_{\pm 0.7}$	66.7$_{\pm 0.5}$
Czech	79.1$_{\pm 0.7}$	84.6$_{\pm 0.3}$	84.7$_{\pm 0.4}$	**85.3**$_{\pm 0.5}$	83.7$_{\pm 0.3}$	84.0$_{\pm 0.2}$	84.3$_{\pm 0.3}$	84.4$_{\pm 0.3}$	84.8$_{\pm 0.2}$	85.1$_{\pm 0.2}$
Danish	87.8$_{\pm 0.3}$	88.1$_{\pm 0.1}$	88.2$_{\pm 0.2}$	88.3$_{\pm 0.2}$	88.5$_{\pm 0.2}$	**88.7**$_{\pm 0.2}$	**88.7**$_{\pm 0.2}$	87.9$_{\pm 0.1}$	87.9$_{\pm 0.1}$	88.0$_{\pm 0.2}$
Finnish	82.3$_{\pm 0.8}$	84.5$_{\pm 0.4}$	84.1$_{\pm 0.4}$	84.1$_{\pm 0.3}$	84.7$_{\pm 0.4}$	84.7$_{\pm 0.2}$	**84.8**$_{\pm 0.2}$	83.9$_{\pm 0.2}$	83.6$_{\pm 0.5}$	83.9$_{\pm 0.3}$
French	85.4$_{\pm 0.2}$	86.5$_{\pm 0.2}$	86.5$_{\pm 0.2}$	**86.6**$_{\pm 0.1}$	86.5$_{\pm 0.3}$	86.5$_{\pm 0.2}$	**86.6**$_{\pm 0.2}$	86.2$_{\pm 0.3}$	86.4$_{\pm 0.3}$	86.2$_{\pm 0.2}$
German	87.4$_{\pm 0.4}$	88.6$_{\pm 0.1}$	88.5$_{\pm 0.1}$	**89.0**$_{\pm 0.2}$	88.2$_{\pm 0.2}$	88.2$_{\pm 0.1}$	88.4$_{\pm 0.1}$	88.3$_{\pm 0.2}$	88.4$_{\pm 0.1}$	88.6$_{\pm 0.3}$
Greek	74.9$_{\pm 1.2}$	78.8$_{\pm 0.8}$	78.6$_{\pm 0.7}$	80.1$_{\pm 0.5}$	77.7$_{\pm 0.6}$	78.1$_{\pm 0.5}$	77.9$_{\pm 0.6}$	78.3$_{\pm 0.9}$	78.6$_{\pm 0.5}$	**80.3**$_{\pm 0.4}$
Hebrew	62.3$_{\pm 0.9}$	64.3$_{\pm 0.6}$	64.0$_{\pm 1.0}$	65.2$_{\pm 0.1}$	64.7$_{\pm 0.9}$	64.8$_{\pm 0.6}$	**65.6**$_{\pm 0.6}$	64.2$_{\pm 0.9}$	63.6$_{\pm 1.1}$	65.2$_{\pm 0.4}$
Hindi	60.7$_{\pm 3.2}$	**67.5**$_{\pm 3.0}$	64.8$_{\pm 1.3}$	65.9$_{\pm 3.3}$	65.9$_{\pm 1.8}$	63.2$_{\pm 2.0}$	63.8$_{\pm 2.2}$	66.7$_{\pm 3.3}$	63.8$_{\pm 2.3}$	67.0$_{\pm 2.7}$
Hungarian	79.1$_{\pm 0.2}$	81.3$_{\pm 0.6}$	81.1$_{\pm 0.4}$	**81.9**$_{\pm 0.3}$	80.9$_{\pm 0.5}$	80.9$_{\pm 0.1}$	81.4$_{\pm 0.1}$	80.8$_{\pm 0.6}$	80.6$_{\pm 0.3}$	81.5$_{\pm 0.4}$
Italian	85.0$_{\pm 0.4}$	85.4$_{\pm 0.2}$	85.6$_{\pm 0.1}$	85.9$_{\pm 0.1}$	85.7$_{\pm 0.2}$	85.7$_{\pm 0.2}$	**86.0**$_{\pm 0.2}$	85.2$_{\pm 0.2}$	85.4$_{\pm 0.2}$	85.5$_{\pm 0.1}$
Japanese	47.8$_{\pm 2.1}$	51.4$_{\pm 0.9}$	53.0$_{\pm 1.5}$	52.7$_{\pm 2.0}$	49.8$_{\pm 0.5}$	49.8$_{\pm 1.5}$	49.4$_{\pm 1.4}$	50.8$_{\pm 1.4}$	50.9$_{\pm 2.0}$	**53.4**$_{\pm 1.7}$
Korean	55.4$_{\pm 2.7}$	58.8$_{\pm 1.1}$	59.9$_{\pm 1.9}$	61.8$_{\pm 1.0}$	59.6$_{\pm 1.5}$	60.2$_{\pm 1.4}$	**63.0**$_{\pm 1.3}$	59.6$_{\pm 0.6}$	60.6$_{\pm 1.7}$	62.5$_{\pm 0.8}$
Latvian	69.5$_{\pm 2.0}$	76.9$_{\pm 0.3}$	**77.3**$_{\pm 0.2}$	76.2$_{\pm 0.6}$	75.3$_{\pm 0.3}$	76.0$_{\pm 0.3}$	75.3$_{\pm 0.1}$	76.1$_{\pm 0.4}$	76.7$_{\pm 0.5}$	76.0$_{\pm 0.2}$
Lithuanian	71.6$_{\pm 1.8}$	76.6$_{\pm 0.6}$	**78.0**$_{\pm 0.4}$	76.3$_{\pm 0.7}$	76.3$_{\pm 0.4}$	77.0$_{\pm 0.5}$	75.9$_{\pm 0.3}$	75.8$_{\pm 0.3}$	77.3$_{\pm 0.4}$	75.9$_{\pm 0.6}$
Norwegian	88.7$_{\pm 0.4}$	90.2$_{\pm 0.2}$	**90.3**$_{\pm 0.2}$	90.1$_{\pm 0.2}$	89.5$_{\pm 0.4}$	89.5$_{\pm 0.3}$	89.5$_{\pm 0.3}$	89.9$_{\pm 0.4}$	90.1$_{\pm 0.2}$	90.0$_{\pm 0.3}$
Persian	72.6$_{\pm 0.7}$	72.2$_{\pm 0.7}$	71.9$_{\pm 0.4}$	72.2$_{\pm 0.6}$	**74.1**$_{\pm 0.3}$	73.3$_{\pm 0.3}$	73.8$_{\pm 0.4}$	72.1$_{\pm 0.4}$	72.2$_{\pm 0.2}$	71.9$_{\pm 0.8}$
Polish	79.7$_{\pm 0.3}$	83.4$_{\pm 0.3}$	**83.6**$_{\pm 0.2}$	83.5$_{\pm 0.3}$	83.3$_{\pm 0.4}$	83.5$_{\pm 0.2}$	83.5$_{\pm 0.3}$	82.9$_{\pm 0.3}$	83.3$_{\pm 0.1}$	83.0$_{\pm 0.3}$
Portuguese	83.0$_{\pm 0.3}$	83.5$_{\pm 0.1}$	83.4$_{\pm 0.1}$	**84.1**$_{\pm 0.1}$	83.5$_{\pm 0.2}$	83.5$_{\pm 0.1}$	83.9$_{\pm 0.0}$	83.5$_{\pm 0.2}$	83.5$_{\pm 0.1}$	83.7$_{\pm 0.2}$
Romanian	80.0$_{\pm 0.5}$	83.5$_{\pm 0.2}$	**83.8**$_{\pm 0.3}$	83.4$_{\pm 0.5}$	83.1$_{\pm 0.3}$	83.4$_{\pm 0.2}$	83.0$_{\pm 0.4}$	82.9$_{\pm 0.4}$	83.6$_{\pm 0.1}$	83.0$_{\pm 0.3}$
Russian	81.5$_{\pm 0.6}$	84.0$_{\pm 0.4}$	83.8$_{\pm 0.5}$	**84.9**$_{\pm 0.3}$	84.0$_{\pm 0.4}$	84.0$_{\pm 0.5}$	84.2$_{\pm 0.4}$	83.9$_{\pm 0.5}$	83.8$_{\pm 0.3}$	84.6$_{\pm 0.3}$
Slovak	78.2$_{\pm 0.8}$	84.5$_{\pm 0.3}$	84.6$_{\pm 0.4}$	**85.0**$_{\pm 0.6}$	83.7$_{\pm 0.6}$	84.0$_{\pm 0.3}$	84.3$_{\pm 0.6}$	84.2$_{\pm 0.4}$	84.6$_{\pm 0.3}$	84.9$_{\pm 0.3}$
Slovenian	79.6$_{\pm 0.5}$	83.6$_{\pm 0.3}$	**83.8**$_{\pm 0.3}$	**83.8**$_{\pm 0.3}$	83.2$_{\pm 0.5}$	83.7$_{\pm 0.2}$	83.6$_{\pm 0.3}$	83.2$_{\pm 0.4}$	83.6$_{\pm 0.3}$	83.5$_{\pm 0.3}$
Spanish	84.4$_{\pm 0.4}$	85.5$_{\pm 0.1}$	85.6$_{\pm 0.1}$	85.7$_{\pm 0.2}$	**85.8**$_{\pm 0.2}$	**85.8**$_{\pm 0.2}$	85.7$_{\pm 0.2}$	85.3$_{\pm 0.2}$	85.6$_{\pm 0.2}$	85.5$_{\pm 0.1}$
Swedish	89.2$_{\pm 0.4}$	90.0$_{\pm 0.2}$	**90.1**$_{\pm 0.2}$	90.0$_{\pm 0.2}$	89.8$_{\pm 0.1}$	89.8$_{\pm 0.1}$	89.8$_{\pm 0.1}$	89.7$_{\pm 0.4}$	89.9$_{\pm 0.1}$	90.0$_{\pm 0.2}$
Tamil	51.9$_{\pm 1.0}$	54.6$_{\pm 1.2}$	55.5$_{\pm 0.7}$	**55.8**$_{\pm 0.7}$	54.7$_{\pm 0.7}$	55.4$_{\pm 0.1}$	54.7$_{\pm 0.9}$	53.3$_{\pm 0.8}$	54.5$_{\pm 0.4}$	54.3$_{\pm 0.8}$
Thai	31.4$_{\pm 6.0}$	52.7$_{\pm 0.8}$	52.9$_{\pm 1.4}$	**55.2**$_{\pm 0.7}$	49.8$_{\pm 0.8}$	51.3$_{\pm 0.9}$	51.7$_{\pm 0.6}$	51.3$_{\pm 1.5}$	52.0$_{\pm 1.4}$	54.9$_{\pm 0.6}$
Turkish	70.0$_{\pm 0.7}$	71.0$_{\pm 0.4}$	70.4$_{\pm 0.3}$	70.4$_{\pm 0.5}$	**71.4**$_{\pm 0.3}$	70.9$_{\pm 0.3}$	71.3$_{\pm 0.3}$	70.7$_{\pm 0.3}$	70.3$_{\pm 0.7}$	70.2$_{\pm 0.5}$
Ukrainian	81.4$_{\pm 0.3}$	84.9$_{\pm 0.3}$	**85.0**$_{\pm 0.4}$	**85.0**$_{\pm 0.2}$	84.4$_{\pm 0.5}$	84.6$_{\pm 0.2}$	84.4$_{\pm 0.3}$	84.5$_{\pm 0.2}$	84.7$_{\pm 0.1}$	84.9$_{\pm 0.3}$
Vietnamese	57.5$_{\pm 0.8}$	56.4$_{\pm 0.4}$	56.9$_{\pm 0.6}$	57.7$_{\pm 0.4}$	58.9$_{\pm 0.4}$	58.8$_{\pm 0.5}$	**59.6**$_{\pm 0.6}$	56.5$_{\pm 0.5}$	56.9$_{\pm 0.9}$	57.3$_{\pm 0.6}$
Average	73.8$_{\pm 0.6}$	77.2$_{\pm 0.2}$	77.3$_{\pm 0.2}$	**77.7**$_{\pm 0.3}$	77.0$_{\pm 0.3}$	77.1$_{\pm 0.2}$	77.3$_{\pm 0.2}$	76.8$_{\pm 0.1}$	77.0$_{\pm 0.1}$	77.5$_{\pm 0.1}$

Table 7: PoS tagging average accuracy results across 5 seeds using DistilMBERT by freezing strategy, language, and aligner. Aligner names: FA - FastAlign, AA - AWESOME-align, BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Afrikaans	90.9$_{\pm 0.5}$	**92.0**$_{\pm 0.3}$	91.8$_{\pm 0.3}$	91.9$_{\pm 0.4}$
Arabic	65.2$_{\pm 0.9}$	68.0$_{\pm 2.6}$	64.7$_{\pm 1.9}$	**69.4**$_{\pm 2.7}$
Bulgarian	89.4$_{\pm 0.3}$	**89.9**$_{\pm 0.2}$	89.7$_{\pm 0.2}$	89.6$_{\pm 0.4}$
Catalan	91.6$_{\pm 0.1}$	**91.7**$_{\pm 0.0}$	**91.7**$_{\pm 0.0}$	91.6$_{\pm 0.2}$
Chinese	76.8$_{\pm 0.4}$	**78.1**$_{\pm 0.5}$	77.8$_{\pm 0.4}$	77.2$_{\pm 0.8}$
Czech	91.7$_{\pm 0.4}$	**92.6**$_{\pm 0.2}$	92.4$_{\pm 0.2}$	92.4$_{\pm 0.1}$
Danish	93.2$_{\pm 0.4}$	**93.8**$_{\pm 0.1}$	93.7$_{\pm 0.2}$	93.6$_{\pm 0.2}$
Finnish	90.8$_{\pm 0.6}$	91.1$_{\pm 0.1}$	**91.3**$_{\pm 0.3}$	91.0$_{\pm 0.3}$
French	86.7$_{\pm 0.2}$	**87.2**$_{\pm 0.3}$	86.8$_{\pm 0.2}$	87.0$_{\pm 0.2}$
German	92.3$_{\pm 0.2}$	92.4$_{\pm 0.3}$	**92.8**$_{\pm 0.2}$	92.5$_{\pm 0.3}$
Greek	87.6$_{\pm 0.3}$	**88.6**$_{\pm 0.2}$	88.5$_{\pm 0.3}$	88.3$_{\pm 0.4}$
Hebrew	81.5$_{\pm 0.1}$	81.0$_{\pm 0.4}$	**82.1**$_{\pm 0.3}$	80.4$_{\pm 0.1}$
Hindi	**77.6**$_{\pm 0.5}$	76.4$_{\pm 1.0}$	77.5$_{\pm 0.9}$	75.5$_{\pm 1.3}$
Hungarian	88.8$_{\pm 0.4}$	89.9$_{\pm 0.1}$	**90.0**$_{\pm 0.3}$	89.7$_{\pm 0.2}$
Italian	91.2$_{\pm 0.2}$	**91.6**$_{\pm 0.1}$	91.5$_{\pm 0.1}$	91.5$_{\pm 0.2}$
Japanese	62.5$_{\pm 1.0}$	**70.0**$_{\pm 1.1}$	67.5$_{\pm 0.8}$	67.6$_{\pm 2.4}$
Korean	74.0$_{\pm 0.3}$	75.8$_{\pm 0.5}$	**76.0**$_{\pm 0.3}$	74.8$_{\pm 0.6}$
Latvian	85.9$_{\pm 0.3}$	85.9$_{\pm 0.1}$	**86.2**$_{\pm 0.1}$	85.5$_{\pm 0.2}$
Lithuanian	87.5$_{\pm 0.8}$	87.5$_{\pm 0.5}$	**87.8**$_{\pm 0.5}$	87.3$_{\pm 0.5}$
Norwegian	89.6$_{\pm 0.3}$	**90.4**$_{\pm 0.4}$	90.1$_{\pm 0.4}$	90.1$_{\pm 0.4}$
Persian	64.2$_{\pm 0.6}$	**67.4**$_{\pm 0.8}$	65.9$_{\pm 0.5}$	66.6$_{\pm 1.7}$
Polish	90.6$_{\pm 0.3}$	**91.2**$_{\pm 0.2}$	91.1$_{\pm 0.1}$	**91.2**$_{\pm 0.2}$
Portuguese	87.0$_{\pm 0.3}$	87.0$_{\pm 0.2}$	86.6$_{\pm 0.3}$	**87.2**$_{\pm 0.4}$
Romanian	85.3$_{\pm 0.4}$	85.7$_{\pm 0.4}$	85.9$_{\pm 0.3}$	**86.3**$_{\pm 0.3}$
Russian	**84.2**$_{\pm 0.4}$	83.7$_{\pm 0.3}$	84.2$_{\pm 0.3}$	83.1$_{\pm 0.3}$
Slovak	89.9$_{\pm 0.5}$	**90.9**$_{\pm 0.2}$	90.6$_{\pm 0.1}$	90.7$_{\pm 0.2}$
Slovenian	90.5$_{\pm 0.4}$	**91.1**$_{\pm 0.2}$	90.8$_{\pm 0.2}$	90.9$_{\pm 0.1}$
Spanish	84.8$_{\pm 0.6}$	85.3$_{\pm 0.5}$	84.3$_{\pm 0.4}$	**86.3**$_{\pm 0.3}$
Swedish	86.8$_{\pm 3.0}$	86.3$_{\pm 1.8}$	86.3$_{\pm 2.2}$	**88.0**$_{\pm 1.2}$
Tamil	72.8$_{\pm 1.2}$	73.1$_{\pm 0.8}$	**74.2**$_{\pm 0.6}$	71.7$_{\pm 1.0}$
Thai	23.1$_{\pm 4.6}$	**69.3**$_{\pm 2.9}$	51.8$_{\pm 14.4}$	42.4$_{\pm 15.9}$
Turkish	85.4$_{\pm 0.6}$	86.2$_{\pm 0.2}$	**86.2**$_{\pm 0.4}$	86.0$_{\pm 0.2}$
Ukrainian	87.7$_{\pm 0.5}$	**88.2**$_{\pm 0.7}$	87.8$_{\pm 0.5}$	87.8$_{\pm 0.6}$
Vietnamese	77.3$_{\pm 0.5}$	81.1$_{\pm 0.6}$	78.9$_{\pm 0.3}$	**81.7**$_{\pm 0.4}$
Average	82.5$_{\pm 0.3}$	**84.7**$_{\pm 0.2}$	84.0$_{\pm 0.5}$	83.7$_{\pm 0.7}$

Table 8: NER accuracy results across 5 seeds using distilMBert by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Arabic	59.2$_{\pm 0.3}$	59.3$_{\pm 0.6}$	**59.8**$_{\pm 0.4}$	59.2$_{\pm 0.5}$
Bulgarian	63.4$_{\pm 0.3}$	63.6$_{\pm 0.4}$	**64.0**$_{\pm 0.2}$	63.8$_{\pm 0.5}$
Chinese	63.9$_{\pm 0.8}$	63.4$_{\pm 0.1}$	**64.1**$_{\pm 0.5}$	63.4$_{\pm 0.5}$
French	**70.1**$_{\pm 0.6}$	68.7$_{\pm 0.6}$	69.4$_{\pm 0.3}$	69.1$_{\pm 0.2}$
German	65.7$_{\pm 0.2}$	64.8$_{\pm 0.3}$	**66.1**$_{\pm 0.5}$	65.5$_{\pm 0.6}$
Greek	60.8$_{\pm 0.4}$	62.0$_{\pm 0.9}$	**62.9**$_{\pm 0.5}$	61.6$_{\pm 0.5}$
Hindi	54.1$_{\pm 0.6}$	54.9$_{\pm 1.0}$	55.3$_{\pm 0.3}$	**55.6**$_{\pm 0.7}$
Spanish	**70.0**$_{\pm 0.3}$	69.4$_{\pm 0.3}$	69.8$_{\pm 0.2}$	**70.0**$_{\pm 0.3}$
Thai	36.1$_{\pm 0.5}$	47.1$_{\pm 1.7}$	42.0$_{\pm 1.4}$	**47.4**$_{\pm 1.2}$
Turkish	57.0$_{\pm 0.5}$	58.7$_{\pm 0.5}$	58.1$_{\pm 0.6}$	**58.7**$_{\pm 0.9}$
Vietnamese	57.6$_{\pm 2.3}$	64.3$_{\pm 0.3}$	63.9$_{\pm 0.7}$	**65.0**$_{\pm 0.6}$
Average	60.1$_{\pm 0.2}$	61.6$_{\pm 0.2}$	61.6$_{\pm 0.1}$	**61.9**$_{\pm 0.2}$

Table 9: XNLI average accuracy results across 5 seeds using DistilMBERT by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Arabic	37.4$_{\pm 0.7}$	38.1$_{\pm 1.2}$	38.3$_{\pm 1.2}$	**38.8**$_{\pm 1.2}$
Chinese	35.7$_{\pm 0.9}$	36.8$_{\pm 1.3}$	36.3$_{\pm 1.6}$	**38.3**$_{\pm 1.7}$
German	49.5$_{\pm 1.6}$	49.8$_{\pm 1.5}$	49.9$_{\pm 1.4}$	**51.0**$_{\pm 1.0}$
Greek	32.4$_{\pm 1.0}$	33.9$_{\pm 1.5}$	33.4$_{\pm 0.7}$	**34.9**$_{\pm 1.6}$
Hindi	29.4$_{\pm 0.9}$	29.6$_{\pm 0.8}$	30.1$_{\pm 0.4}$	**30.2**$_{\pm 0.8}$
Romanian	44.2$_{\pm 1.9}$	46.4$_{\pm 2.4}$	44.9$_{\pm 2.0}$	**47.3**$_{\pm 1.2}$
Russian	49.0$_{\pm 1.7}$	50.2$_{\pm 2.0}$	49.1$_{\pm 1.8}$	**50.6**$_{\pm 1.8}$
Spanish	50.9$_{\pm 0.9}$	51.7$_{\pm 1.6}$	51.4$_{\pm 2.0}$	**52.0**$_{\pm 1.5}$
Thai	**18.7**$_{\pm 0.8}$	17.7$_{\pm 1.4}$	18.3$_{\pm 0.8}$	18.6$_{\pm 1.1}$
Turkish	31.0$_{\pm 0.5}$	32.8$_{\pm 1.2}$	32.1$_{\pm 0.5}$	**33.3**$_{\pm 1.1}$
Vietnamese	38.0$_{\pm 0.5}$	41.3$_{\pm 2.7}$	38.4$_{\pm 1.3}$	**41.5**$_{\pm 2.9}$
Average	37.8$_{\pm 0.6}$	38.9$_{\pm 1.1}$	38.4$_{\pm 0.9}$	**39.7**$_{\pm 1.1}$

Table 10: XQuAD average F1-score across 5 seeds using distilMBERT by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment			ALIGNFREEZE with front-freezing			ALIGNFREEZE with back-freezing		
	-	FA	AA	BD	FA	AA	BD	FA	AA	BD
Afrikaans	87.0$_{\pm 0.4}$	**88.4**$_{\pm 0.3}$	88.2$_{\pm 0.2}$	88.2$_{\pm 0.3}$	87.3$_{\pm 0.4}$	87.7$_{\pm 0.3}$	87.7$_{\pm 0.3}$	88.0$_{\pm 0.6}$	88.0$_{\pm 0.2}$	87.5$_{\pm 0.5}$
Arabic	51.0$_{\pm 0.5}$	63.7$_{\pm 1.6}$	63.9$_{\pm 1.0}$	**65.1**$_{\pm 1.4}$	63.6$_{\pm 0.9}$	63.1$_{\pm 1.4}$	63.7$_{\pm 1.2}$	63.1$_{\pm 0.9}$	63.4$_{\pm 1.3}$	64.1$_{\pm 1.2}$
Bulgarian	86.3$_{\pm 0.8}$	87.9$_{\pm 0.7}$	88.1$_{\pm 0.3}$	**88.1**$_{\pm 0.5}$	87.8$_{\pm 0.6}$	87.8$_{\pm 0.6}$	87.8$_{\pm 0.4}$	87.5$_{\pm 0.6}$	87.8$_{\pm 0.7}$	87.9$_{\pm 0.3}$
Catalan	86.7$_{\pm 0.3}$	88.2$_{\pm 0.3}$	88.6$_{\pm 0.3}$	**89.0**$_{\pm 0.3}$	87.9$_{\pm 0.3}$	88.0$_{\pm 0.4}$	88.1$_{\pm 0.4}$	88.1$_{\pm 0.4}$	88.3$_{\pm 0.3}$	88.6$_{\pm 0.1}$
Chinese	65.7$_{\pm 1.0}$	67.9$_{\pm 1.3}$	67.4$_{\pm 0.1}$	69.0$_{\pm 0.5}$	67.6$_{\pm 1.1}$	66.8$_{\pm 0.1}$	69.0$_{\pm 0.8}$	68.4$_{\pm 1.1}$	68.2$_{\pm 0.6}$	**69.7**$_{\pm 0.6}$
Czech	84.2$_{\pm 0.9}$	85.9$_{\pm 1.1}$	85.9$_{\pm 0.5}$	**86.7**$_{\pm 0.5}$	86.1$_{\pm 0.8}$	86.0$_{\pm 0.8}$	86.4$_{\pm 0.4}$	85.6$_{\pm 0.8}$	85.7$_{\pm 0.9}$	86.5$_{\pm 0.5}$
Danish	89.3$_{\pm 0.3}$	89.3$_{\pm 0.1}$	**89.4**$_{\pm 0.2}$	**89.4**$_{\pm 0.2}$	**89.4**$_{\pm 0.2}$	89.3$_{\pm 0.3}$	**89.4**$_{\pm 0.2}$	89.0$_{\pm 0.2}$	89.1$_{\pm 0.2}$	89.2$_{\pm 0.2}$
Finnish	85.9$_{\pm 0.6}$	86.9$_{\pm 0.4}$	86.9$_{\pm 0.3}$	**87.1**$_{\pm 0.5}$	86.9$_{\pm 0.4}$	86.9$_{\pm 0.3}$	**87.1**$_{\pm 0.5}$	86.4$_{\pm 0.4}$	86.5$_{\pm 0.3}$	87.0$_{\pm 0.2}$
French	85.7$_{\pm 0.4}$	86.7$_{\pm 0.2}$	86.7$_{\pm 0.3}$	**86.9**$_{\pm 0.4}$	86.5$_{\pm 0.3}$	86.5$_{\pm 0.3}$	86.7$_{\pm 0.2}$	86.1$_{\pm 0.2}$	86.3$_{\pm 0.3}$	86.4$_{\pm 0.3}$
German	88.3$_{\pm 0.5}$	89.7$_{\pm 0.5}$	89.6$_{\pm 0.2}$	**89.9**$_{\pm 0.3}$	89.2$_{\pm 0.4}$	89.2$_{\pm 0.1}$	89.5$_{\pm 0.2}$	89.5$_{\pm 0.4}$	89.2$_{\pm 0.4}$	89.8$_{\pm 0.3}$
Greek	78.7$_{\pm 1.4}$	81.7$_{\pm 1.0}$	81.6$_{\pm 0.3}$	**82.4**$_{\pm 1.0}$	81.3$_{\pm 1.2}$	80.8$_{\pm 0.3}$	81.7$_{\pm 0.9}$	81.0$_{\pm 1.2}$	81.0$_{\pm 1.3}$	81.3$_{\pm 1.0}$
Hebrew	58.0$_{\pm 2.1}$	64.6$_{\pm 0.7}$	65.0$_{\pm 1.1}$	64.7$_{\pm 1.2}$	62.4$_{\pm 1.7}$	62.1$_{\pm 0.8}$	62.7$_{\pm 1.2}$	64.5$_{\pm 1.0}$	**65.2**$_{\pm 0.8}$	65.0$_{\pm 0.6}$
Hindi	67.7$_{\pm 0.7}$	70.1$_{\pm 2.1}$	69.6$_{\pm 1.2}$	70.0$_{\pm 3.2}$	**70.7**$_{\pm 1.8}$	69.3$_{\pm 2.0}$	69.6$_{\pm 2.5}$	67.2$_{\pm 2.4}$	68.6$_{\pm 2.9}$	69.9$_{\pm 2.6}$
Hungarian	82.2$_{\pm 0.5}$	82.6$_{\pm 0.4}$	82.9$_{\pm 0.3}$	**83.0**$_{\pm 0.5}$	82.5$_{\pm 0.4}$	82.4$_{\pm 0.5}$	82.9$_{\pm 0.3}$	82.1$_{\pm 0.4}$	82.0$_{\pm 0.4}$	82.6$_{\pm 0.3}$
Italian	84.3$_{\pm 0.5}$	85.6$_{\pm 0.3}$	85.5$_{\pm 0.6}$	**86.1**$_{\pm 0.4}$	85.4$_{\pm 0.3}$	85.0$_{\pm 0.3}$	85.3$_{\pm 0.2}$	85.4$_{\pm 0.3}$	85.6$_{\pm 0.3}$	85.8$_{\pm 0.3}$
Japanese	48.1$_{\pm 0.8}$	51.6$_{\pm 1.7}$	**55.0**$_{\pm 1.6}$	53.2$_{\pm 1.8}$	50.5$_{\pm 1.4}$	51.3$_{\pm 1.2}$	50.8$_{\pm 1.0}$	48.8$_{\pm 1.6}$	52.3$_{\pm 1.6}$	51.5$_{\pm 1.3}$
Korean	63.8$_{\pm 1.0}$	64.4$_{\pm 0.6}$	63.4$_{\pm 0.7}$	65.9$_{\pm 0.6}$	64.4$_{\pm 0.7}$	64.5$_{\pm 0.4}$	65.6$_{\pm 0.4}$	64.2$_{\pm 0.9}$	63.7$_{\pm 1.0}$	**66.3**$_{\pm 0.3}$
Latvian	81.3$_{\pm 0.5}$	82.8$_{\pm 0.5}$	**83.1**$_{\pm 0.6}$	82.6$_{\pm 0.6}$	82.4$_{\pm 0.3}$	82.8$_{\pm 0.3}$	82.5$_{\pm 0.2}$	82.5$_{\pm 0.4}$	82.9$_{\pm 0.6}$	82.4$_{\pm 0.4}$
Lithuanian	81.5$_{\pm 0.5}$	82.5$_{\pm 0.4}$	**83.0**$_{\pm 0.2}$	82.8$_{\pm 0.5}$	82.7$_{\pm 0.2}$	82.9$_{\pm 0.2}$	**83.0**$_{\pm 0.2}$	81.8$_{\pm 0.4}$	82.7$_{\pm 0.5}$	82.1$_{\pm 0.6}$
Norwegian	90.6$_{\pm 0.4}$	91.4$_{\pm 0.2}$	91.5$_{\pm 0.2}$	**91.5**$_{\pm 0.4}$	91.1$_{\pm 0.4}$	91.2$_{\pm 0.2}$	91.2$_{\pm 0.4}$	91.2$_{\pm 0.3}$	91.4$_{\pm 0.2}$	91.4$_{\pm 0.3}$
Persian	73.6$_{\pm 0.5}$	73.9$_{\pm 0.7}$	74.0$_{\pm 0.6}$	74.4$_{\pm 0.9}$	**74.9**$_{\pm 0.8}$	74.7$_{\pm 0.6}$	**74.9**$_{\pm 0.8}$	73.0$_{\pm 0.6}$	73.5$_{\pm 0.5}$	73.8$_{\pm 0.9}$
Polish	82.8$_{\pm 0.8}$	84.4$_{\pm 0.7}$	84.3$_{\pm 0.5}$	84.8$_{\pm 0.6}$	**84.9**$_{\pm 0.6}$	84.5$_{\pm 0.6}$	84.8$_{\pm 0.5}$	84.1$_{\pm 0.5}$	84.3$_{\pm 0.6}$	84.5$_{\pm 0.4}$
Portuguese	82.7$_{\pm 0.5}$	83.3$_{\pm 0.2}$	83.5$_{\pm 0.2}$	**83.9**$_{\pm 0.2}$	83.7$_{\pm 0.2}$	83.3$_{\pm 0.5}$	83.7$_{\pm 0.1}$	83.1$_{\pm 0.4}$	83.1$_{\pm 0.2}$	83.4$_{\pm 0.3}$
Romanian	83.4$_{\pm 0.7}$	85.4$_{\pm 0.4}$	85.4$_{\pm 0.3}$	**85.6**$_{\pm 0.5}$	85.3$_{\pm 0.5}$	85.1$_{\pm 0.5}$	85.3$_{\pm 0.3}$	85.2$_{\pm 0.3}$	85.3$_{\pm 0.6}$	85.5$_{\pm 0.4}$
Russian	81.4$_{\pm 1.3}$	84.1$_{\pm 0.5}$	83.9$_{\pm 0.4}$	**84.7**$_{\pm 0.5}$	83.8$_{\pm 0.7}$	83.8$_{\pm 0.8}$	83.9$_{\pm 0.4}$	83.5$_{\pm 0.6}$	83.5$_{\pm 0.7}$	84.4$_{\pm 0.5}$
Slovak	82.8$_{\pm 1.3}$	85.3$_{\pm 0.9}$	85.5$_{\pm 0.6}$	**86.6**$_{\pm 0.7}$	85.6$_{\pm 0.8}$	85.5$_{\pm 1.1}$	86.0$_{\pm 0.7}$	84.9$_{\pm 0.6}$	85.1$_{\pm 0.9}$	86.2$_{\pm 0.8}$
Slovenian	83.5$_{\pm 0.7}$	84.9$_{\pm 0.7}$	84.8$_{\pm 0.4}$	85.7$_{\pm 0.4}$	85.8$_{\pm 0.7}$	85.7$_{\pm 0.7}$	**85.9**$_{\pm 0.4}$	84.4$_{\pm 0.4}$	84.2$_{\pm 0.6}$	85.1$_{\pm 0.3}$
Spanish	85.1$_{\pm 0.2}$	85.8$_{\pm 0.3}$	85.9$_{\pm 0.2}$	**86.1**$_{\pm 0.3}$	85.9$_{\pm 0.3}$	85.6$_{\pm 0.2}$	85.9$_{\pm 0.3}$	85.5$_{\pm 0.3}$	85.8$_{\pm 0.2}$	85.7$_{\pm 0.2}$
Swedish	90.3$_{\pm 0.3}$	**91.4**$_{\pm 0.3}$	91.3$_{\pm 0.2}$	**91.4**$_{\pm 0.3}$	91.0$_{\pm 0.3}$	90.9$_{\pm 0.2}$	90.8$_{\pm 0.3}$	91.1$_{\pm 0.3}$	91.3$_{\pm 0.4}$	91.3$_{\pm 0.2}$
Tamil	58.1$_{\pm 0.9}$	60.2$_{\pm 1.1}$	61.0$_{\pm 0.7}$	60.9$_{\pm 0.7}$	59.2$_{\pm 1.1}$	59.8$_{\pm 0.7}$	60.7$_{\pm 0.5}$	59.1$_{\pm 1.0}$	58.9$_{\pm 0.5}$	**61.0**$_{\pm 0.9}$
Thai	52.0$_{\pm 1.3}$	60.9$_{\pm 0.7}$	61.2$_{\pm 0.6}$	**62.6**$_{\pm 0.5}$	58.1$_{\pm 1.5}$	59.7$_{\pm 1.1}$	60.8$_{\pm 0.5}$	59.7$_{\pm 0.4}$	60.7$_{\pm 0.3}$	62.2$_{\pm 0.7}$
Turkish	71.5$_{\pm 0.9}$	**72.3**$_{\pm 0.5}$	72.1$_{\pm 0.6}$	72.2$_{\pm 0.8}$	71.8$_{\pm 0.6}$	71.5$_{\pm 0.7}$	71.6$_{\pm 0.6}$	72.0$_{\pm 0.6}$	71.8$_{\pm 0.5}$	71.2$_{\pm 1.3}$
Ukrainian	82.0$_{\pm 1.2}$	84.8$_{\pm 0.8}$	84.9$_{\pm 0.3}$	**85.0**$_{\pm 0.5}$	84.5$_{\pm 0.7}$	84.4$_{\pm 0.7}$	84.3$_{\pm 0.5}$	84.5$_{\pm 0.6}$	84.6$_{\pm 0.6}$	84.8$_{\pm 0.6}$
Vietnamese	62.3$_{\pm 0.3}$	61.0$_{\pm 0.6}$	61.5$_{\pm 0.4}$	61.9$_{\pm 0.5}$	62.1$_{\pm 0.5}$	62.2$_{\pm 0.6}$	**62.4**$_{\pm 0.6}$	61.0$_{\pm 0.5}$	61.3$_{\pm 0.4}$	61.9$_{\pm 0.5}$
Average	77.0$_{\pm 0.5}$	79.1$_{\pm 0.3}$	79.2$_{\pm 0.2}$	**79.6**$_{\pm 0.4}$	78.9$_{\pm 0.4}$	78.8$_{\pm 0.3}$	79.2$_{\pm 0.2}$	78.6$_{\pm 0.3}$	78.9$_{\pm 0.3}$	79.3$_{\pm 0.3}$

Table 11: PoS tagging average accuracy results across 5 seeds using mBERT by freezing strategy, language, and aligner. Aligner names: FA - FastAlign, AA - AWESOME-align, BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Afrikaans	92.8$_{\pm 0.2}$	92.6$_{\pm 0.5}$	92.7$_{\pm 0.2}$	**92.8**$_{\pm 0.4}$
Arabic	67.1$_{\pm 0.9}$	68.9$_{\pm 1.5}$	68.9$_{\pm 1.8}$	**70.7**$_{\pm 2.0}$
Bulgarian	**90.7**$_{\pm 0.4}$	89.9$_{\pm 0.3}$	90.7$_{\pm 0.3}$	90.1$_{\pm 0.3}$
Catalan	92.8$_{\pm 0.2}$	**92.9**$_{\pm 0.1}$	92.8$_{\pm 0.1}$	92.8$_{\pm 0.1}$
Chinese	78.9$_{\pm 0.7}$	78.7$_{\pm 0.6}$	79.0$_{\pm 0.6}$	**79.6**$_{\pm 1.0}$
Czech	93.4$_{\pm 0.1}$	93.3$_{\pm 0.3}$	**93.6**$_{\pm 0.1}$	93.2$_{\pm 0.1}$
Danish	94.3$_{\pm 0.1}$	94.1$_{\pm 0.2}$	**94.3**$_{\pm 0.2}$	94.2$_{\pm 0.2}$
Finnish	**92.2**$_{\pm 0.3}$	91.7$_{\pm 0.4}$	92.0$_{\pm 0.2}$	91.8$_{\pm 0.3}$
French	88.6$_{\pm 0.7}$	88.1$_{\pm 0.3}$	88.7$_{\pm 1.1}$	**89.0**$_{\pm 0.9}$
German	**94.0**$_{\pm 0.1}$	93.3$_{\pm 0.3}$	93.9$_{\pm 0.1}$	93.5$_{\pm 0.2}$
Greek	**91.0**$_{\pm 0.3}$	90.5$_{\pm 0.4}$	90.7$_{\pm 0.4}$	90.7$_{\pm 0.5}$
Hebrew	**84.4**$_{\pm 0.2}$	83.8$_{\pm 0.4}$	**84.4**$_{\pm 0.2}$	83.7$_{\pm 0.4}$
Hindi	**82.7**$_{\pm 0.9}$	80.7$_{\pm 0.7}$	82.5$_{\pm 0.6}$	80.7$_{\pm 0.3}$
Hungarian	**91.7**$_{\pm 0.3}$	91.1$_{\pm 0.5}$	91.5$_{\pm 0.2}$	91.6$_{\pm 0.4}$
Italian	92.3$_{\pm 0.1}$	92.4$_{\pm 0.2}$	**92.5**$_{\pm 0.2}$	**92.5**$_{\pm 0.2}$
Japanese	69.2$_{\pm 1.5}$	**72.8**$_{\pm 0.7}$	72.0$_{\pm 0.3}$	72.5$_{\pm 0.6}$
Korean	84.3$_{\pm 0.5}$	84.0$_{\pm 0.6}$	**84.9**$_{\pm 0.7}$	83.8$_{\pm 0.8}$
Latvian	87.4$_{\pm 0.3}$	**87.7**$_{\pm 0.2}$	87.4$_{\pm 0.4}$	87.5$_{\pm 0.2}$
Lithuanian	**90.3**$_{\pm 0.2}$	89.7$_{\pm 0.4}$	89.8$_{\pm 0.5}$	89.8$_{\pm 0.3}$
Norwegian	91.3$_{\pm 0.2}$	90.8$_{\pm 0.6}$	**91.5**$_{\pm 0.5}$	91.1$_{\pm 0.4}$
Persian	70.9$_{\pm 1.3}$	71.2$_{\pm 1.1}$	70.8$_{\pm 1.8}$	**73.6**$_{\pm 0.5}$
Polish	92.2$_{\pm 0.1}$	92.0$_{\pm 0.3}$	**92.3**$_{\pm 0.1}$	92.1$_{\pm 0.2}$
Portuguese	**89.2**$_{\pm 0.4}$	88.4$_{\pm 0.5}$	89.1$_{\pm 0.6}$	88.4$_{\pm 0.4}$
Romanian	**88.3**$_{\pm 0.9}$	86.2$_{\pm 1.3}$	88.1$_{\pm 1.1}$	85.4$_{\pm 2.7}$
Russian	85.0$_{\pm 0.8}$	84.8$_{\pm 0.8}$	**85.5**$_{\pm 0.6}$	84.8$_{\pm 0.5}$
Slovak	**92.0**$_{\pm 0.2}$	91.7$_{\pm 0.3}$	91.8$_{\pm 0.3}$	91.8$_{\pm 0.3}$
Slovenian	92.3$_{\pm 0.4}$	92.3$_{\pm 0.2}$	92.4$_{\pm 0.2}$	**92.5**$_{\pm 0.3}$
Spanish	86.3$_{\pm 1.0}$	83.2$_{\pm 1.1}$	85.8$_{\pm 1.4}$	**86.3**$_{\pm 1.3}$
Swedish	88.8$_{\pm 1.7}$	86.8$_{\pm 0.9}$	**89.1**$_{\pm 0.7}$	88.7$_{\pm 0.6}$
Tamil	**80.1**$_{\pm 0.8}$	78.2$_{\pm 0.7}$	79.5$_{\pm 0.9}$	77.3$_{\pm 0.8}$
Thai	33.7$_{\pm 13.5}$	**69.6**$_{\pm 0.7}$	64.8$_{\pm 7.0}$	64.0$_{\pm 12.9}$
Turkish	**90.1**$_{\pm 0.7}$	89.4$_{\pm 0.7}$	89.4$_{\pm 0.5}$	89.5$_{\pm 0.5}$
Ukrainian	**89.4**$_{\pm 0.3}$	88.7$_{\pm 1.0}$	89.3$_{\pm 0.4}$	88.8$_{\pm 0.5}$
Vietnamese	86.8$_{\pm 0.4}$	87.2$_{\pm 0.6}$	86.7$_{\pm 0.5}$	**87.8**$_{\pm 0.5}$
Average	85.7$_{\pm 0.3}$	86.4$_{\pm 0.3}$	**86.7**$_{\pm 0.2}$	86.5$_{\pm 0.6}$

Table 12: NER accuracy results across 5 seeds using mBERT by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Arabic	64.6$_{\pm 0.5}$	65.0$_{\pm 0.6}$	**65.6**$_{\pm 0.2}$	65.0$_{\pm 0.8}$
Bulgarian	68.0$_{\pm 0.8}$	69.1$_{\pm 0.6}$	**69.3**$_{\pm 0.2}$	69.1$_{\pm 0.7}$
Chinese	68.9$_{\pm 0.6}$	69.5$_{\pm 0.7}$	69.2$_{\pm 0.4}$	**69.9**$_{\pm 0.6}$
French	72.8$_{\pm 0.6}$	73.6$_{\pm 0.3}$	**74.2**$_{\pm 0.3}$	73.7$_{\pm 0.5}$
German	70.1$_{\pm 0.5}$	70.3$_{\pm 0.6}$	**71.0**$_{\pm 0.3}$	70.9$_{\pm 0.6}$
Greek	66.6$_{\pm 0.7}$	67.5$_{\pm 0.6}$	**67.6**$_{\pm 0.6}$	67.4$_{\pm 0.8}$
Hindi	59.7$_{\pm 1.1}$	60.9$_{\pm 1.0}$	**61.0**$_{\pm 0.5}$	61.0$_{\pm 0.3}$
Spanish	73.4$_{\pm 0.4}$	73.9$_{\pm 0.3}$	**74.8**$_{\pm 0.3}$	74.2$_{\pm 0.3}$
Thai	53.3$_{\pm 2.3}$	**57.4**$_{\pm 0.8}$	56.8$_{\pm 0.3}$	56.1$_{\pm 0.8}$
Turkish	61.4$_{\pm 0.5}$	63.5$_{\pm 0.6}$	63.2$_{\pm 0.4}$	**63.8**$_{\pm 0.3}$
Vietnamese	69.0$_{\pm 0.5}$	70.3$_{\pm 0.2}$	**70.9**$_{\pm 0.3}$	70.8$_{\pm 0.1}$
Average	66.3$_{\pm 0.6}$	67.4$_{\pm 0.4}$	**67.7**$_{\pm 0.2}$	67.5$_{\pm 0.3}$

Table 13: XNLI average accuracy results across 5 seeds using mBERT by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Arabic	55.5$_{\pm 1.2}$	54.6$_{\pm 0.6}$	55.0$_{\pm 1.0}$	**55.6**$_{\pm 1.3}$
Chinese	**53.1**$_{\pm 0.9}$	52.4$_{\pm 1.2}$	53.0$_{\pm 0.8}$	52.8$_{\pm 0.7}$
German	67.7$_{\pm 0.3}$	67.7$_{\pm 0.9}$	67.7$_{\pm 0.7}$	**68.2**$_{\pm 0.2}$
Greek	53.0$_{\pm 0.9}$	**53.5**$_{\pm 0.5}$	53.3$_{\pm 0.9}$	53.0$_{\pm 0.3}$
Hindi	**49.1**$_{\pm 0.7}$	47.4$_{\pm 1.1}$	48.4$_{\pm 1.6}$	48.0$_{\pm 1.3}$
Romanian	66.1$_{\pm 0.6}$	66.8$_{\pm 0.2}$	66.6$_{\pm 0.6}$	**66.9**$_{\pm 0.2}$
Russian	**66.3**$_{\pm 0.8}$	64.9$_{\pm 0.2}$	65.1$_{\pm 0.6}$	65.6$_{\pm 0.4}$
Spanish	69.1$_{\pm 0.7}$	68.8$_{\pm 0.4}$	68.9$_{\pm 0.8}$	**70.0**$_{\pm 0.6}$
Thai	35.5$_{\pm 0.9}$	**35.6**$_{\pm 1.6}$	34.6$_{\pm 1.0}$	35.1$_{\pm 1.3}$
Turkish	47.2$_{\pm 1.4}$	46.8$_{\pm 1.4}$	**47.6**$_{\pm 0.8}$	46.7$_{\pm 1.4}$
Vietnamese	63.7$_{\pm 0.6}$	62.9$_{\pm 0.9}$	63.7$_{\pm 0.6}$	**63.7**$_{\pm 1.0}$
Average	56.9$_{\pm 0.3}$	56.5$_{\pm 0.5}$	56.7$_{\pm 0.5}$	**56.9**$_{\pm 0.5}$

Table 14: XQuAD average F1-score results across 5 seeds using mBERT by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment			ALIGNFREEZE with front-freezing			ALIGNFREEZE with back-freezing		
	-	FA	AA	BD	FA	AA	BD	FA	AA	BD
Afrikaans	88.4$_{\pm 0.3}$	88.6$_{\pm 0.1}$	88.7$_{\pm 0.1}$	**88.8**$_{\pm 0.1}$	88.6$_{\pm 0.2}$	88.6$_{\pm 0.2}$	**88.8**$_{\pm 0.1}$	88.6$_{\pm 0.2}$	88.7$_{\pm 0.2}$	88.4$_{\pm 0.1}$
Arabic	63.2$_{\pm 0.8}$	65.5$_{\pm 0.9}$	65.3$_{\pm 1.1}$	**67.6**$_{\pm 1.1}$	65.5$_{\pm 0.5}$	65.3$_{\pm 1.0}$	67.0$_{\pm 0.7}$	63.9$_{\pm 0.8}$	64.3$_{\pm 0.9}$	66.3$_{\pm 0.5}$
Bulgarian	89.3$_{\pm 0.5}$	89.1$_{\pm 0.3}$	89.4$_{\pm 0.2}$	89.1$_{\pm 0.2}$	89.5$_{\pm 0.1}$	**89.9**$_{\pm 0.3}$	89.5$_{\pm 0.2}$	88.9$_{\pm 0.3}$	88.9$_{\pm 0.4}$	88.9$_{\pm 0.4}$
Catalan	89.4$_{\pm 0.5}$	89.2$_{\pm 0.2}$	89.5$_{\pm 0.3}$	89.4$_{\pm 0.4}$	89.6$_{\pm 0.5}$	**89.8**$_{\pm 0.3}$	89.5$_{\pm 0.8}$	89.2$_{\pm 0.1}$	89.4$_{\pm 0.2}$	89.3$_{\pm 0.2}$
Chinese	71.4$_{\pm 0.4}$	70.5$_{\pm 0.4}$	70.7$_{\pm 0.8}$	**72.2**$_{\pm 0.7}$	71.4$_{\pm 0.3}$	71.1$_{\pm 0.6}$	72.0$_{\pm 0.8}$	70.4$_{\pm 0.9}$	70.6$_{\pm 0.7}$	71.5$_{\pm 0.7}$
Czech	86.6$_{\pm 0.7}$	87.0$_{\pm 0.4}$	87.1$_{\pm 0.3}$	87.3$_{\pm 0.3}$	87.3$_{\pm 0.2}$	87.4$_{\pm 0.2}$	**87.8**$_{\pm 0.3}$	86.6$_{\pm 0.7}$	86.7$_{\pm 0.6}$	86.8$_{\pm 0.7}$
Danish	90.2$_{\pm 0.3}$	89.9$_{\pm 0.1}$	90.0$_{\pm 0.0}$	90.2$_{\pm 0.1}$	90.0$_{\pm 0.1}$	90.0$_{\pm 0.2}$	**90.4**$_{\pm 0.1}$	89.7$_{\pm 0.2}$	89.7$_{\pm 0.1}$	89.8$_{\pm 0.1}$
Finnish	88.3$_{\pm 0.5}$	88.1$_{\pm 0.0}$	88.2$_{\pm 0.1}$	88.3$_{\pm 0.2}$	88.3$_{\pm 0.2}$	88.5$_{\pm 0.1}$	**88.7**$_{\pm 0.2}$	87.6$_{\pm 0.2}$	87.7$_{\pm 0.2}$	88.0$_{\pm 0.2}$
French	87.1$_{\pm 0.2}$	87.5$_{\pm 0.1}$	87.7$_{\pm 0.1}$	87.7$_{\pm 0.1}$	87.6$_{\pm 0.1}$	87.6$_{\pm 0.3}$	**87.9**$_{\pm 0.1}$	87.2$_{\pm 0.2}$	87.4$_{\pm 0.1}$	87.4$_{\pm 0.2}$
German	89.0$_{\pm 0.4}$	89.9$_{\pm 0.3}$	**90.1**$_{\pm 0.3}$	89.9$_{\pm 0.3}$	89.9$_{\pm 0.3}$	89.9$_{\pm 0.2}$	89.9$_{\pm 0.3}$	89.7$_{\pm 0.2}$	89.8$_{\pm 0.3}$	89.7$_{\pm 0.4}$
Greek	84.8$_{\pm 0.9}$	85.0$_{\pm 0.4}$	84.7$_{\pm 0.4}$	**85.6**$_{\pm 0.5}$	85.1$_{\pm 0.2}$	85.1$_{\pm 0.5}$	**85.6**$_{\pm 0.3}$	85.0$_{\pm 0.6}$	84.7$_{\pm 0.4}$	85.0$_{\pm 0.9}$
Hebrew	67.7$_{\pm 1.5}$	67.2$_{\pm 0.4}$	67.6$_{\pm 0.8}$	66.7$_{\pm 0.9}$	68.4$_{\pm 0.4}$	68.5$_{\pm 0.2}$	**68.6**$_{\pm 0.5}$	67.0$_{\pm 0.7}$	67.5$_{\pm 0.7}$	66.2$_{\pm 1.4}$
Hindi	71.2$_{\pm 1.7}$	72.0$_{\pm 1.3}$	72.2$_{\pm 0.6}$	72.9$_{\pm 0.8}$	74.5$_{\pm 2.2}$	74.7$_{\pm 0.9}$	**75.2**$_{\pm 2.1}$	70.6$_{\pm 0.7}$	70.9$_{\pm 0.7}$	72.3$_{\pm 1.0}$
Hungarian	85.2$_{\pm 0.5}$	84.8$_{\pm 0.2}$	85.0$_{\pm 0.1}$	85.2$_{\pm 0.2}$	85.1$_{\pm 0.1}$	85.2$_{\pm 0.2}$	**85.3**$_{\pm 0.1}$	84.5$_{\pm 0.3}$	84.5$_{\pm 0.4}$	84.8$_{\pm 0.3}$
Italian	86.2$_{\pm 0.3}$	86.4$_{\pm 0.1}$	**86.7**$_{\pm 0.1}$	**86.7**$_{\pm 0.1}$	86.6$_{\pm 0.1}$	**86.7**$_{\pm 0.2}$	**86.7**$_{\pm 0.2}$	86.2$_{\pm 0.1}$	86.3$_{\pm 0.1}$	86.5$_{\pm 0.2}$
Japanese	56.5$_{\pm 2.4}$	54.9$_{\pm 1.5}$	56.2$_{\pm 0.9}$	**59.6**$_{\pm 0.5}$	56.3$_{\pm 1.2}$	56.5$_{\pm 1.0}$	58.5$_{\pm 0.9}$	54.0$_{\pm 1.9}$	54.9$_{\pm 1.0}$	58.6$_{\pm 1.0}$
Korean	66.3$_{\pm 0.8}$	64.5$_{\pm 0.7}$	64.7$_{\pm 0.5}$	65.9$_{\pm 0.7}$	66.0$_{\pm 0.5}$	66.3$_{\pm 0.3}$	**66.8**$_{\pm 0.3}$	64.4$_{\pm 0.7}$	64.2$_{\pm 0.6}$	66.1$_{\pm 0.5}$
Latvian	86.0$_{\pm 0.4}$	85.8$_{\pm 0.1}$	86.0$_{\pm 0.2}$	86.0$_{\pm 0.2}$	86.1$_{\pm 0.1}$	86.1$_{\pm 0.2}$	**86.2**$_{\pm 0.2}$	85.2$_{\pm 0.2}$	85.7$_{\pm 0.1}$	85.3$_{\pm 0.1}$
Lithuanian	86.3$_{\pm 0.4}$	86.2$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.5$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.4$_{\pm 0.2}$	**86.6**$_{\pm 0.1}$	85.9$_{\pm 0.3}$	86.2$_{\pm 0.3}$	86.0$_{\pm 0.1}$
Norwegian	91.9$_{\pm 0.2}$	91.9$_{\pm 0.1}$	**92.0**$_{\pm 0.2}$	**92.0**$_{\pm 0.1}$	91.9$_{\pm 0.1}$	91.9$_{\pm 0.1}$	**92.0**$_{\pm 0.2}$	91.9$_{\pm 0.1}$	**92.0**$_{\pm 0.1}$	**92.0**$_{\pm 0.1}$
Persian	**77.1**$_{\pm 0.7}$	75.2$_{\pm 0.6}$	75.8$_{\pm 0.7}$	75.3$_{\pm 0.6}$	76.9$_{\pm 0.7}$	76.7$_{\pm 0.5}$	77.0$_{\pm 0.3}$	74.5$_{\pm 0.6}$	74.9$_{\pm 0.4}$	74.9$_{\pm 0.4}$
Polish	84.8$_{\pm 0.8}$	85.4$_{\pm 0.5}$	85.6$_{\pm 0.5}$	85.2$_{\pm 0.4}$	85.9$_{\pm 0.3}$	86.0$_{\pm 0.3}$	**86.1**$_{\pm 0.2}$	84.5$_{\pm 0.6}$	84.8$_{\pm 0.6}$	84.8$_{\pm 0.6}$
Portuguese	84.1$_{\pm 0.2}$	84.1$_{\pm 0.2}$	84.1$_{\pm 0.1}$	84.3$_{\pm 0.1}$	84.2$_{\pm 0.1}$	84.2$_{\pm 0.1}$	**84.4**$_{\pm 0.1}$	83.9$_{\pm 0.2}$	84.0$_{\pm 0.2}$	84.2$_{\pm 0.1}$
Romanian	86.9$_{\pm 0.3}$	86.8$_{\pm 0.5}$	87.1$_{\pm 0.4}$	86.7$_{\pm 0.4}$	87.6$_{\pm 0.5}$	**87.8**$_{\pm 0.5}$	87.5$_{\pm 0.2}$	86.7$_{\pm 0.4}$	86.7$_{\pm 0.3}$	86.7$_{\pm 0.5}$
Russian	87.3$_{\pm 0.4}$	86.5$_{\pm 0.4}$	86.8$_{\pm 0.2}$	87.2$_{\pm 0.5}$	86.9$_{\pm 0.4}$	87.1$_{\pm 0.4}$	**87.6**$_{\pm 0.3}$	86.4$_{\pm 0.5}$	86.5$_{\pm 0.3}$	86.6$_{\pm 0.2}$
Slovak	86.3$_{\pm 0.8}$	85.9$_{\pm 0.5}$	86.2$_{\pm 0.4}$	86.5$_{\pm 0.4}$	86.4$_{\pm 0.4}$	86.8$_{\pm 0.5}$	**87.6**$_{\pm 0.5}$	85.5$_{\pm 0.6}$	85.6$_{\pm 0.5}$	85.9$_{\pm 0.6}$
Slovenian	86.6$_{\pm 0.4}$	86.2$_{\pm 0.3}$	86.5$_{\pm 0.3}$	86.3$_{\pm 0.3}$	87.0$_{\pm 0.2}$	**87.3**$_{\pm 0.2}$	86.9$_{\pm 0.3}$	85.9$_{\pm 0.5}$	85.7$_{\pm 0.7}$	86.0$_{\pm 0.7}$
Spanish	86.7$_{\pm 0.4}$	86.7$_{\pm 0.2}$	86.9$_{\pm 0.2}$	86.9$_{\pm 0.3}$	87.0$_{\pm 0.3}$	87.1$_{\pm 0.2}$	**87.2**$_{\pm 0.1}$	86.5$_{\pm 0.2}$	86.7$_{\pm 0.2}$	86.8$_{\pm 0.3}$
Swedish	91.6$_{\pm 0.3}$	91.7$_{\pm 0.1}$	91.9$_{\pm 0.1}$	91.8$_{\pm 0.2}$	91.7$_{\pm 0.1}$	91.8$_{\pm 0.1}$	**92.0**$_{\pm 0.2}$	91.5$_{\pm 0.2}$	91.6$_{\pm 0.2}$	91.5$_{\pm 0.1}$
Tamil	61.4$_{\pm 0.6}$	63.0$_{\pm 0.5}$	63.4$_{\pm 0.3}$	**65.5**$_{\pm 0.5}$	62.3$_{\pm 0.3}$	62.3$_{\pm 0.4}$	63.8$_{\pm 0.6}$	62.0$_{\pm 1.2}$	62.6$_{\pm 1.3}$	63.8$_{\pm 0.5}$
Thai	69.0$_{\pm 0.4}$	67.2$_{\pm 0.1}$	68.2$_{\pm 0.3}$	68.7$_{\pm 0.2}$	68.8$_{\pm 0.5}$	69.1$_{\pm 0.4}$	**69.4**$_{\pm 0.6}$	67.1$_{\pm 0.5}$	67.7$_{\pm 0.3}$	68.6$_{\pm 0.3}$
Turkish	72.7$_{\pm 0.8}$	72.6$_{\pm 0.5}$	73.0$_{\pm 0.5}$	73.0$_{\pm 0.4}$	72.5$_{\pm 0.4}$	72.4$_{\pm 0.3}$	**73.1**$_{\pm 0.4}$	72.3$_{\pm 0.3}$	72.3$_{\pm 0.3}$	72.7$_{\pm 0.3}$
Ukrainian	86.2$_{\pm 0.3}$	85.9$_{\pm 0.6}$	86.1$_{\pm 0.3}$	86.4$_{\pm 0.4}$	86.0$_{\pm 0.3}$	86.2$_{\pm 0.4}$	**86.5**$_{\pm 0.3}$	85.7$_{\pm 0.3}$	85.7$_{\pm 0.3}$	85.9$_{\pm 0.4}$
Vietnamese	64.7$_{\pm 0.6}$	64.3$_{\pm 0.4}$	64.3$_{\pm 0.2}$	64.6$_{\pm 0.2}$	65.2$_{\pm 0.4}$	65.3$_{\pm 0.4}$	**65.4**$_{\pm 0.3}$	63.7$_{\pm 0.2}$	64.0$_{\pm 0.2}$	64.5$_{\pm 0.2}$
Average	80.9$_{\pm 0.1}$	80.8$_{\pm 0.2}$	81.0$_{\pm 0.2}$	81.3$_{\pm 0.1}$	81.2$_{\pm 0.1}$	81.3$_{\pm 0.2}$	**81.7**$_{\pm 0.2}$	80.4$_{\pm 0.1}$	80.6$_{\pm 0.1}$	80.9$_{\pm 0.2}$

Table 15: PoS tagging average accuracy results across 5 seeds using XLM-R by freezing strategy, language, and aligner. Aligner names: FA - FastAlign, AA - AWESOME-align, BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Afrikaans	91.7$_{\pm 0.3}$	91.8$_{\pm 0.1}$	91.5$_{\pm 0.3}$	**91.9**$_{\pm 0.3}$
Arabic	75.5$_{\pm 0.8}$	**77.3**$_{\pm 1.1}$	76.4$_{\pm 1.3}$	76.8$_{\pm 1.4}$
Bulgarian	89.8$_{\pm 0.4}$	89.8$_{\pm 0.2}$	89.9$_{\pm 0.1}$	**89.9**$_{\pm 0.3}$
Catalan	**91.1**$_{\pm 0.1}$	90.9$_{\pm 0.2}$	91.0$_{\pm 0.1}$	90.8$_{\pm 0.1}$
Chinese	**79.0**$_{\pm 0.8}$	78.4$_{\pm 0.4}$	78.4$_{\pm 0.6}$	78.0$_{\pm 0.3}$
Czech	**92.5**$_{\pm 0.2}$	92.1$_{\pm 0.3}$	92.2$_{\pm 0.1}$	92.2$_{\pm 0.2}$
Danish	**93.6**$_{\pm 0.2}$	93.5$_{\pm 0.1}$	93.6$_{\pm 0.1}$	93.3$_{\pm 0.0}$
Finnish	**91.0**$_{\pm 0.1}$	90.5$_{\pm 0.1}$	90.7$_{\pm 0.2}$	90.4$_{\pm 0.1}$
French	86.7$_{\pm 0.4}$	87.3$_{\pm 0.3}$	86.8$_{\pm 0.4}$	**87.7**$_{\pm 0.9}$
German	**91.6**$_{\pm 0.2}$	90.9$_{\pm 0.2}$	91.3$_{\pm 0.1}$	91.1$_{\pm 0.1}$
Greek	**91.7**$_{\pm 0.4}$	91.7$_{\pm 0.1}$	91.2$_{\pm 0.4}$	91.5$_{\pm 0.2}$
Hebrew	81.6$_{\pm 0.3}$	81.7$_{\pm 0.4}$	81.4$_{\pm 0.3}$	**81.9**$_{\pm 0.3}$
Hindi	**82.2**$_{\pm 0.1}$	81.4$_{\pm 0.5}$	81.8$_{\pm 0.6}$	80.9$_{\pm 0.5}$
Hungarian	**92.0**$_{\pm 0.3}$	91.6$_{\pm 0.2}$	91.3$_{\pm 0.4}$	91.6$_{\pm 0.1}$
Italian	90.8$_{\pm 0.3}$	**90.9**$_{\pm 0.1}$	90.8$_{\pm 0.2}$	**90.9**$_{\pm 0.1}$
Japanese	70.2$_{\pm 1.6}$	70.7$_{\pm 1.3}$	**71.2**$_{\pm 1.4}$	70.3$_{\pm 1.0}$
Korean	**79.4**$_{\pm 0.9}$	78.6$_{\pm 0.3}$	79.4$_{\pm 0.7}$	77.8$_{\pm 0.9}$
Latvian	**88.5**$_{\pm 0.5}$	88.4$_{\pm 0.6}$	88.2$_{\pm 0.6}$	88.0$_{\pm 0.4}$
Lithuanian	**89.6**$_{\pm 0.2}$	89.6$_{\pm 0.1}$	89.4$_{\pm 0.4}$	89.5$_{\pm 0.2}$
Norwegian	91.9$_{\pm 0.5}$	**92.2**$_{\pm 0.2}$	92.0$_{\pm 0.3}$	92.0$_{\pm 0.2}$
Persian	73.9$_{\pm 1.1}$	**77.9**$_{\pm 1.0}$	75.8$_{\pm 0.7}$	76.2$_{\pm 1.5}$
Polish	**91.1**$_{\pm 0.2}$	90.9$_{\pm 0.1}$	90.9$_{\pm 0.1}$	90.8$_{\pm 0.1}$
Portuguese	87.6$_{\pm 0.8}$	**88.2**$_{\pm 0.4}$	88.0$_{\pm 0.4}$	87.8$_{\pm 0.2}$
Romanian	84.0$_{\pm 0.7}$	86.6$_{\pm 2.1}$	84.9$_{\pm 0.3}$	**86.8**$_{\pm 2.2}$
Russian	**84.8**$_{\pm 0.6}$	83.9$_{\pm 0.4}$	84.4$_{\pm 0.3}$	83.7$_{\pm 0.3}$
Slovak	90.5$_{\pm 0.3}$	90.4$_{\pm 0.5}$	90.1$_{\pm 0.4}$	**90.7**$_{\pm 0.4}$
Slovenian	**91.1**$_{\pm 0.3}$	91.1$_{\pm 0.1}$	90.5$_{\pm 0.3}$	91.1$_{\pm 0.2}$
Spanish	86.1$_{\pm 1.9}$	**88.5**$_{\pm 0.2}$	86.5$_{\pm 1.2}$	88.0$_{\pm 0.4}$
Swedish	89.6$_{\pm 0.9}$	**90.8**$_{\pm 0.5}$	90.4$_{\pm 0.7}$	90.2$_{\pm 0.6}$
Tamil	**81.3**$_{\pm 0.9}$	80.1$_{\pm 0.4}$	80.0$_{\pm 0.6}$	79.3$_{\pm 0.2}$
Thai	19.2$_{\pm 0.4}$	**26.5**$_{\pm 3.4}$	21.0$_{\pm 1.1}$	20.7$_{\pm 0.6}$
Turkish	90.7$_{\pm 0.6}$	**90.8**$_{\pm 0.2}$	90.1$_{\pm 0.4}$	90.8$_{\pm 0.1}$
Ukrainian	90.3$_{\pm 0.3}$	**90.3**$_{\pm 0.6}$	89.8$_{\pm 0.9}$	89.2$_{\pm 0.8}$
Vietnamese	84.6$_{\pm 0.4}$	86.2$_{\pm 0.6}$	84.3$_{\pm 0.7}$	**86.3**$_{\pm 0.3}$
Average	84.9$_{\pm 0.4}$	**85.3**$_{\pm 0.2}$	84.8$_{\pm 0.3}$	84.9$_{\pm 0.1}$

Table 16: NER accuracy results across 5 seeds using XLM-R Base by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Arabic	**70.8**$_{\pm 0.3}$	70.0$_{\pm 0.4}$	70.2$_{\pm 0.3}$	70.0$_{\pm 0.4}$
Bulgarian	**77.0**$_{\pm 0.2}$	75.5$_{\pm 0.3}$	76.5$_{\pm 0.3}$	75.8$_{\pm 0.4}$
Chinese	**72.5**$_{\pm 0.3}$	72.4$_{\pm 0.3}$	72.3$_{\pm 0.2}$	72.1$_{\pm 0.4}$
French	**77.1**$_{\pm 0.1}$	76.4$_{\pm 0.3}$	76.7$_{\pm 0.1}$	76.0$_{\pm 0.3}$
German	**75.7**$_{\pm 0.4}$	75.0$_{\pm 0.3}$	75.2$_{\pm 0.4}$	74.6$_{\pm 0.4}$
Greek	**75.2**$_{\pm 0.3}$	74.0$_{\pm 0.3}$	74.7$_{\pm 0.2}$	73.5$_{\pm 0.4}$
Hindi	**69.0**$_{\pm 0.3}$	68.7$_{\pm 0.5}$	68.9$_{\pm 0.5}$	68.0$_{\pm 0.8}$
Spanish	**78.3**$_{\pm 0.2}$	77.3$_{\pm 0.2}$	77.6$_{\pm 0.2}$	76.8$_{\pm 0.2}$
Thai	**71.1**$_{\pm 0.3}$	70.4$_{\pm 0.3}$	71.0$_{\pm 0.3}$	70.1$_{\pm 0.3}$
Turkish	**71.9**$_{\pm 0.5}$	71.6$_{\pm 0.4}$	71.7$_{\pm 0.3}$	71.2$_{\pm 0.4}$
Vietnamese	**73.8**$_{\pm 0.4}$	73.3$_{\pm 0.3}$	73.7$_{\pm 0.3}$	73.2$_{\pm 0.4}$
Average	**73.9**$_{\pm 0.2}$	73.2$_{\pm 0.2}$	73.6$_{\pm 0.2}$	72.9$_{\pm 0.3}$

Table 17: NLI average accuracy results across 5 seeds using XLM-R by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment	ALIGNFREEZE with front-freezing	ALIGNFREEZE with back-freezing
	-	BD	BD	BD
Arabic	**51.0**$_{\pm 1.0}$	50.8$_{\pm 0.4}$	49.9$_{\pm 0.7}$	50.8$_{\pm 0.3}$
Chinese	**47.5**$_{\pm 0.8}$	47.1$_{\pm 0.6}$	46.4$_{\pm 0.6}$	46.6$_{\pm 0.8}$
German	**65.7**$_{\pm 0.6}$	65.2$_{\pm 0.9}$	64.2$_{\pm 0.7}$	64.1$_{\pm 1.0}$
Greek	**61.6**$_{\pm 0.7}$	60.9$_{\pm 0.9}$	58.8$_{\pm 0.8}$	59.4$_{\pm 0.7}$
Hindi	**58.4**$_{\pm 1.0}$	57.7$_{\pm 0.7}$	56.2$_{\pm 0.7}$	56.4$_{\pm 0.8}$
Romanian	**69.4**$_{\pm 0.5}$	68.9$_{\pm 0.8}$	67.7$_{\pm 1.0}$	68.2$_{\pm 0.5}$
Russian	**66.3**$_{\pm 0.7}$	65.0$_{\pm 0.6}$	64.3$_{\pm 0.9}$	64.6$_{\pm 0.9}$
Spanish	**67.8**$_{\pm 1.0}$	67.7$_{\pm 0.9}$	67.1$_{\pm 0.1}$	67.3$_{\pm 1.0}$
Thai	**60.1**$_{\pm 0.9}$	57.9$_{\pm 2.1}$	56.9$_{\pm 1.2}$	57.6$_{\pm 0.4}$
Turkish	**60.4**$_{\pm 0.7}$	60.3$_{\pm 1.1}$	59.6$_{\pm 0.5}$	60.0$_{\pm 0.6}$
Vietnamese	**65.3**$_{\pm 0.2}$	65.2$_{\pm 0.4}$	64.5$_{\pm 0.7}$	64.4$_{\pm 0.9}$
Average	**61.2**$_{\pm 0.4}$	60.6$_{\pm 0.6}$	59.6$_{\pm 0.5}$	59.9$_{\pm 0.4}$

Table 18: XQuAD average F1-score results across 5 seeds using XLM-R Base by freezing strategy, language, and aligner. Aligner names: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment						ALIGNFREEZE with front-freezing					
	-	FA		AA		BD		FA		AA		BD	
	-	0%	50%	0%	50%	0%	50%	0%	50%	0%	50%	0%	50%
Arabic	59.2$_{\pm 0.3}$	61.4$_{\pm 0.5}$	61.9$_{\pm 0.6}$	**62.2**$_{\pm 0.4}$	61.6$_{\pm 0.3}$	60.2$_{\pm 0.6}$	59.6$_{\pm 1.3}$	60.2$_{\pm 0.7}$	60.7$_{\pm 0.4}$	60.9$_{\pm 0.4}$	61.0$_{\pm 0.4}$	60.6$_{\pm 0.5}$	60.0$_{\pm 0.1}$
Bulgarian	63.4$_{\pm 0.3}$	65.3$_{\pm 0.3}$	65.5$_{\pm 0.4}$	**65.8**$_{\pm 0.4}$	65.4$_{\pm 0.5}$	**65.8**$_{\pm 0.6}$	65.5$_{\pm 0.4}$	65.0$_{\pm 0.3}$	65.1$_{\pm 0.5}$	65.6$_{\pm 0.3}$	65.5$_{\pm 0.4}$	65.0$_{\pm 0.4}$	64.9$_{\pm 0.2}$
Chinese	63.9$_{\pm 0.9}$	65.1$_{\pm 0.6}$	65.4$_{\pm 0.6}$	64.9$_{\pm 0.4}$	64.6$_{\pm 0.3}$	64.3$_{\pm 0.1}$	63.9$_{\pm 0.6}$	**65.5**$_{\pm 0.4}$	**65.5**$_{\pm 0.7}$	65.4$_{\pm 0.3}$	65.2$_{\pm 0.3}$	65.4$_{\pm 0.5}$	65.0$_{\pm 0.5}$
French	70.1$_{\pm 0.7}$	69.3$_{\pm 0.3}$	69.5$_{\pm 0.2}$	70.0$_{\pm 0.4}$	69.6$_{\pm 0.6}$	69.0$_{\pm 0.4}$	69.9$_{\pm 0.4}$	69.7$_{\pm 0.5}$	69.9$_{\pm 0.3}$	70.1$_{\pm 0.1}$	**70.5**$_{\pm 0.5}$	70.2$_{\pm 0.3}$	70.0$_{\pm 0.4}$
German	65.7$_{\pm 0.3}$	66.9$_{\pm 0.5}$	66.8$_{\pm 0.4}$	67.2$_{\pm 0.7}$	67.1$_{\pm 0.4}$	66.9$_{\pm 0.6}$	66.6$_{\pm 0.4}$	**67.4**$_{\pm 0.3}$	67.1$_{\pm 0.5}$	67.1$_{\pm 0.5}$	**67.4**$_{\pm 0.6}$	66.9$_{\pm 0.4}$	66.7$_{\pm 0.7}$
Greek	60.8$_{\pm 0.5}$	62.8$_{\pm 0.9}$	62.5$_{\pm 0.7}$	**64.4**$_{\pm 0.4}$	63.7$_{\pm 0.6}$	63.9$_{\pm 0.3}$	63.7$_{\pm 0.3}$	63.0$_{\pm 0.4}$	63.2$_{\pm 0.3}$	63.7$_{\pm 0.4}$	63.7$_{\pm 0.3}$	63.5$_{\pm 0.6}$	63.6$_{\pm 0.3}$
Hindi	54.1$_{\pm 0.7}$	56.3$_{\pm 0.4}$	56.2$_{\pm 0.2}$	**57.4**$_{\pm 0.7}$	57.2$_{\pm 0.4}$	56.6$_{\pm 0.7}$	56.2$_{\pm 0.6}$	55.3$_{\pm 0.4}$	55.5$_{\pm 0.5}$	55.7$_{\pm 0.3}$	56.0$_{\pm 0.5}$	56.3$_{\pm 0.5}$	56.2$_{\pm 0.5}$
Russian	63.6$_{\pm 0.3}$	64.6$_{\pm 0.4}$	64.7$_{\pm 0.4}$	65.0$_{\pm 0.6}$	64.8$_{\pm 0.5}$	63.9$_{\pm 0.3}$	63.7$_{\pm 0.9}$	64.4$_{\pm 0.3}$	64.7$_{\pm 0.4}$	64.7$_{\pm 0.5}$	**65.2**$_{\pm 0.3}$	64.3$_{\pm 0.6}$	64.3$_{\pm 0.6}$
Spanish	70.0$_{\pm 0.4}$	69.9$_{\pm 0.5}$	70.0$_{\pm 0.2}$	70.5$_{\pm 0.3}$	70.2$_{\pm 0.3}$	70.0$_{\pm 0.5}$	**70.6**$_{\pm 0.4}$	69.9$_{\pm 0.6}$	70.0$_{\pm 0.3}$	70.1$_{\pm 0.3}$	70.1$_{\pm 0.3}$	**70.6**$_{\pm 0.2}$	70.4$_{\pm 0.6}$
Thai	36.1$_{\pm 0.5}$	47.0$_{\pm 2.0}$	46.0$_{\pm 1.2}$	49.1$_{\pm 2.0}$	49.8$_{\pm 1.1}$	**49.9**$_{\pm 1.5}$	49.2$_{\pm 1.5}$	44.2$_{\pm 1.8}$	43.8$_{\pm 1.2}$	44.6$_{\pm 1.2}$	45.3$_{\pm 1.7}$	43.7$_{\pm 2.2}$	43.6$_{\pm 1.4}$
Turkish	57.0$_{\pm 0.5}$	61.2$_{\pm 0.5}$	60.8$_{\pm 0.4}$	**62.3**$_{\pm 0.2}$	61.6$_{\pm 0.6}$	61.6$_{\pm 0.3}$	62.1$_{\pm 0.5}$	59.8$_{\pm 0.4}$	60.1$_{\pm 0.2}$	60.4$_{\pm 0.5}$	60.2$_{\pm 0.4}$	60.5$_{\pm 0.4}$	60.5$_{\pm 0.5}$
Vietnamese	57.6$_{\pm 2.6}$	65.6$_{\pm 0.3}$	66.1$_{\pm 0.3}$	**66.8**$_{\pm 0.4}$	66.0$_{\pm 0.8}$	65.5$_{\pm 0.5}$	65.3$_{\pm 0.2}$	65.4$_{\pm 0.4}$	65.2$_{\pm 0.3}$	65.7$_{\pm 0.4}$	65.3$_{\pm 0.6}$	66.2$_{\pm 0.6}$	65.8$_{\pm 0.3}$
Average	60.1$_{\pm 0.3}$	62.9$_{\pm 0.4}$	62.9$_{\pm 0.2}$	**63.8**$_{\pm 0.3}$	63.5$_{\pm 0.3}$	63.1$_{\pm 0.2}$	63.0$_{\pm 0.3}$	62.5$_{\pm 0.2}$	62.6$_{\pm 0.2}$	62.8$_{\pm 0.1}$	63.0$_{\pm 0.1}$	62.8$_{\pm 0.3}$	62.6$_{\pm 0.2}$

Table 19: NLI average accuracy results across 5 seeds using DistilMBERT by freezing strategy, language, aligner, and filtering threshold. Aligner names: FA - FastAlign, AA - AWESOME-align, BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	**vanilla realignment**						**ALIGNFREEZE with front-freezing**					
	-	0%	25%	37%	50%	62%	75%	0%	25%	37%	50%	62%	75%
Afrikaans	85.5$_{\pm 0.2}$	86.4$_{\pm 0.3}$	86.3$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.2$_{\pm 0.5}$	**86.6**$_{\pm 0.3}$	86.2$_{\pm 0.2}$	86.2$_{\pm 0.2}$	86.3$_{\pm 0.2}$	86.3$_{\pm 0.4}$	86.2$_{\pm 0.2}$	86.2$_{\pm 0.2}$
Arabic	51.7$_{\pm 1.7}$	63.9$_{\pm 0.5}$	63.6$_{\pm 0.6}$	63.9$_{\pm 0.9}$	63.6$_{\pm 0.8}$	64.1$_{\pm 0.9}$	63.6$_{\pm 0.8}$	63.3$_{\pm 0.5}$	63.5$_{\pm 0.7}$	63.9$_{\pm 0.2}$	63.8$_{\pm 0.7}$	**64.2**$_{\pm 0.5}$	64.0$_{\pm 0.4}$
Bulgarian	85.0$_{\pm 0.5}$	**87.4**$_{\pm 0.2}$	87.2$_{\pm 0.2}$	87.3$_{\pm 0.3}$	87.3$_{\pm 0.3}$	87.1$_{\pm 0.4}$	87.1$_{\pm 0.2}$	87.1$_{\pm 0.3}$	87.1$_{\pm 0.3}$	87.1$_{\pm 0.2}$	87.1$_{\pm 0.1}$	87.0$_{\pm 0.4}$	87.1$_{\pm 0.2}$
Catalan	86.6$_{\pm 0.4}$	**87.8**$_{\pm 0.2}$	**87.8**$_{\pm 0.2}$	87.7$_{\pm 0.1}$	**87.8**$_{\pm 0.2}$	87.7$_{\pm 0.1}$	87.7$_{\pm 0.2}$	87.6$_{\pm 0.3}$	87.7$_{\pm 0.2}$	87.7$_{\pm 0.3}$	**87.8**$_{\pm 0.1}$	87.7$_{\pm 0.1}$	87.6$_{\pm 0.1}$
Chinese	64.3$_{\pm 1.4}$	66.2$_{\pm 0.5}$	66.4$_{\pm 0.7}$	**66.7**$_{\pm 0.4}$	66.4$_{\pm 0.6}$	66.5$_{\pm 0.3}$	66.4$_{\pm 0.2}$	66.6$_{\pm 0.5}$	66.5$_{\pm 0.5}$	66.6$_{\pm 0.3}$	66.6$_{\pm 0.4}$	66.6$_{\pm 0.4}$	66.5$_{\pm 0.3}$
Czech	79.1$_{\pm 0.7}$	**84.6**$_{\pm 0.3}$	84.2$_{\pm 0.2}$	84.3$_{\pm 0.2}$	84.2$_{\pm 0.3}$	84.1$_{\pm 0.4}$	84.1$_{\pm 0.3}$	83.7$_{\pm 0.3}$	83.7$_{\pm 0.2}$	83.8$_{\pm 0.1}$	83.8$_{\pm 0.3}$	83.7$_{\pm 0.3}$	83.8$_{\pm 0.4}$
Danish	87.8$_{\pm 0.3}$	88.1$_{\pm 0.1}$	87.9$_{\pm 0.2}$	88.0$_{\pm 0.2}$	88.0$_{\pm 0.2}$	87.9$_{\pm 0.2}$	88.1$_{\pm 0.1}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.2}$	**88.6**$_{\pm 0.3}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.2}$
Finnish	82.3$_{\pm 0.8}$	84.5$_{\pm 0.4}$	84.3$_{\pm 0.4}$	84.5$_{\pm 0.4}$	84.3$_{\pm 0.4}$	84.5$_{\pm 0.4}$	84.5$_{\pm 0.3}$	84.7$_{\pm 0.4}$	**84.8**$_{\pm 0.2}$	**84.8**$_{\pm 0.2}$	84.7$_{\pm 0.2}$	**84.8**$_{\pm 0.3}$	84.7$_{\pm 0.1}$
French	85.4$_{\pm 0.2}$	**86.5**$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.4$_{\pm 0.1}$	86.3$_{\pm 0.3}$	86.3$_{\pm 0.3}$	**86.5**$_{\pm 0.3}$	**86.5**$_{\pm 0.2}$	**86.5**$_{\pm 0.2}$	**86.5**$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.4$_{\pm 0.3}$
German	87.4$_{\pm 0.4}$	88.6$_{\pm 0.1}$	88.7$_{\pm 0.2}$	**88.8**$_{\pm 0.2}$	88.6$_{\pm 0.3}$	88.7$_{\pm 0.1}$	**88.8**$_{\pm 0.2}$	88.2$_{\pm 0.2}$	88.2$_{\pm 0.2}$	88.3$_{\pm 0.3}$	88.2$_{\pm 0.3}$	88.2$_{\pm 0.1}$	88.2$_{\pm 0.2}$
Greek	74.9$_{\pm 1.2}$	**78.8**$_{\pm 0.8}$	78.3$_{\pm 0.7}$	78.5$_{\pm 0.9}$	78.5$_{\pm 0.8}$	78.4$_{\pm 0.6}$	78.3$_{\pm 0.6}$	77.7$_{\pm 0.6}$	77.9$_{\pm 0.8}$	77.7$_{\pm 0.6}$	77.8$_{\pm 0.7}$	77.7$_{\pm 0.4}$	77.8$_{\pm 0.4}$
Hebrew	62.3$_{\pm 0.9}$	64.3$_{\pm 0.6}$	64.1$_{\pm 1.0}$	64.2$_{\pm 0.3}$	64.2$_{\pm 0.6}$	63.9$_{\pm 0.6}$	63.9$_{\pm 0.8}$	64.7$_{\pm 0.9}$	**64.8**$_{\pm 0.7}$	64.6$_{\pm 0.6}$	64.6$_{\pm 0.5}$	64.5$_{\pm 0.4}$	64.4$_{\pm 0.6}$
Hindi	60.7$_{\pm 3.2}$	67.5$_{\pm 3.0}$	68.4$_{\pm 1.8}$	67.4$_{\pm 2.3}$	68.0$_{\pm 2.2}$	68.1$_{\pm 1.8}$	**68.7**$_{\pm 1.9}$	65.9$_{\pm 1.8}$	66.1$_{\pm 1.9}$	65.4$_{\pm 1.6}$	65.8$_{\pm 2.4}$	66.1$_{\pm 1.7}$	65.5$_{\pm 2.1}$
Hungarian	79.1$_{\pm 0.2}$	**81.3**$_{\pm 0.6}$	81.0$_{\pm 0.5}$	81.0$_{\pm 0.5}$	81.1$_{\pm 0.1}$	81.1$_{\pm 0.3}$	81.0$_{\pm 0.5}$	80.9$_{\pm 0.5}$	81.1$_{\pm 0.3}$	81.0$_{\pm 0.3}$	81.0$_{\pm 0.3}$	81.2$_{\pm 0.3}$	81.1$_{\pm 0.4}$
Italian	85.0$_{\pm 0.4}$	85.4$_{\pm 0.2}$	85.4$_{\pm 0.2}$	85.4$_{\pm 0.2}$	85.4$_{\pm 0.2}$	85.4$_{\pm 0.2}$	85.4$_{\pm 0.2}$	**85.7**$_{\pm 0.2}$	**85.7**$_{\pm 0.2}$	**85.7**$_{\pm 0.2}$	**85.7**$_{\pm 0.1}$	85.6$_{\pm 0.1}$	85.6$_{\pm 0.2}$
Japanese	47.8$_{\pm 2.1}$	51.4$_{\pm 0.9}$	52.6$_{\pm 0.8}$	52.8$_{\pm 1.3}$	52.4$_{\pm 1.3}$	**53.1**$_{\pm 1.5}$	53.0$_{\pm 1.5}$	49.8$_{\pm 0.5}$	50.3$_{\pm 0.9}$	50.4$_{\pm 0.9}$	50.0$_{\pm 1.1}$	50.7$_{\pm 0.6}$	50.3$_{\pm 0.8}$
Korean	55.4$_{\pm 2.7}$	58.8$_{\pm 1.1}$	59.4$_{\pm 1.2}$	59.9$_{\pm 0.6}$	59.3$_{\pm 0.6}$	59.8$_{\pm 0.3}$	59.3$_{\pm 0.9}$	59.6$_{\pm 1.5}$	59.8$_{\pm 1.7}$	**60.7**$_{\pm 0.8}$	60.1$_{\pm 0.8}$	60.4$_{\pm 0.6}$	60.1$_{\pm 1.3}$
Latvian	69.5$_{\pm 2.0}$	76.9$_{\pm 0.3}$	77.0$_{\pm 0.2}$	76.9$_{\pm 0.4}$	77.2$_{\pm 0.3}$	77.2$_{\pm 0.5}$	**77.3**$_{\pm 0.2}$	75.3$_{\pm 0.3}$	75.5$_{\pm 0.3}$	75.6$_{\pm 0.3}$	75.3$_{\pm 0.3}$	75.6$_{\pm 0.3}$	75.8$_{\pm 0.1}$
Lithuanian	71.6$_{\pm 1.8}$	76.6$_{\pm 0.6}$	77.2$_{\pm 0.3}$	77.0$_{\pm 0.5}$	77.2$_{\pm 0.5}$	77.0$_{\pm 0.5}$	**77.4**$_{\pm 0.3}$	76.3$_{\pm 0.4}$	76.4$_{\pm 0.4}$	76.5$_{\pm 0.2}$	76.3$_{\pm 0.4}$	76.3$_{\pm 0.4}$	76.5$_{\pm 0.2}$
Norwegian	88.7$_{\pm 0.4}$	90.2$_{\pm 0.2}$	90.2$_{\pm 0.2}$	90.2$_{\pm 0.3}$	90.2$_{\pm 0.2}$	**90.4**$_{\pm 0.2}$	90.3$_{\pm 0.2}$	89.5$_{\pm 0.4}$	89.6$_{\pm 0.3}$	89.6$_{\pm 0.2}$	89.6$_{\pm 0.2}$	89.7$_{\pm 0.3}$	89.7$_{\pm 0.2}$
Persian	72.6$_{\pm 0.7}$	72.2$_{\pm 0.7}$	71.6$_{\pm 0.8}$	72.1$_{\pm 0.8}$	71.8$_{\pm 0.5}$	72.0$_{\pm 0.9}$	71.7$_{\pm 0.4}$	**74.1**$_{\pm 0.3}$	73.8$_{\pm 0.2}$	73.9$_{\pm 0.3}$	73.7$_{\pm 0.4}$	74.0$_{\pm 0.1}$	73.7$_{\pm 0.3}$
Polish	79.7$_{\pm 0.3}$	**83.4**$_{\pm 0.3}$	83.1$_{\pm 0.2}$	83.3$_{\pm 0.5}$	83.1$_{\pm 0.2}$	83.3$_{\pm 0.3}$	83.0$_{\pm 0.2}$	83.3$_{\pm 0.4}$	83.3$_{\pm 0.2}$	83.3$_{\pm 0.3}$	83.2$_{\pm 0.2}$	83.2$_{\pm 0.3}$	83.2$_{\pm 0.1}$
Portuguese	83.0$_{\pm 0.3}$	83.5$_{\pm 0.1}$	83.4$_{\pm 0.1}$	83.4$_{\pm 0.1}$	83.5$_{\pm 0.2}$	83.4$_{\pm 0.1}$	83.3$_{\pm 0.2}$	83.5$_{\pm 0.2}$	**83.6**$_{\pm 0.1}$	**83.6**$_{\pm 0.1}$	**83.6**$_{\pm 0.1}$	**83.6**$_{\pm 0.1}$	83.5$_{\pm 0.1}$
Romanian	80.0$_{\pm 0.5}$	**83.5**$_{\pm 0.2}$	83.4$_{\pm 0.2}$	83.3$_{\pm 0.5}$	83.4$_{\pm 0.2}$	83.4$_{\pm 0.3}$	83.4$_{\pm 0.4}$	83.1$_{\pm 0.3}$	83.2$_{\pm 0.2}$	83.1$_{\pm 0.3}$	83.1$_{\pm 0.4}$	83.3$_{\pm 0.3}$	83.2$_{\pm 0.3}$
Russian	81.5$_{\pm 0.6}$	**84.0**$_{\pm 0.4}$	83.6$_{\pm 0.3}$	83.7$_{\pm 0.5}$	83.6$_{\pm 0.5}$	83.5$_{\pm 0.7}$	83.5$_{\pm 0.4}$	**84.0**$_{\pm 0.4}$	83.9$_{\pm 0.5}$	83.9$_{\pm 0.5}$	83.9$_{\pm 0.3}$	83.7$_{\pm 0.5}$	83.8$_{\pm 0.5}$
Slovak	78.2$_{\pm 0.8}$	**84.5**$_{\pm 0.3}$	84.0$_{\pm 0.2}$	84.1$_{\pm 0.1}$	84.2$_{\pm 0.2}$	83.9$_{\pm 0.4}$	83.9$_{\pm 0.4}$	83.7$_{\pm 0.6}$	83.6$_{\pm 0.4}$	83.8$_{\pm 0.2}$	83.7$_{\pm 0.3}$	83.6$_{\pm 0.4}$	83.5$_{\pm 0.3}$
Slovenian	79.6$_{\pm 0.5}$	**83.6**$_{\pm 0.3}$	83.3$_{\pm 0.4}$	83.3$_{\pm 0.3}$	83.2$_{\pm 0.4}$	83.1$_{\pm 0.4}$	83.1$_{\pm 0.4}$	83.2$_{\pm 0.5}$	83.3$_{\pm 0.3}$	83.2$_{\pm 0.1}$	83.3$_{\pm 0.2}$	83.1$_{\pm 0.4}$	83.1$_{\pm 0.3}$
Spanish	84.4$_{\pm 0.4}$	85.5$_{\pm 0.1}$	85.3$_{\pm 0.2}$	85.4$_{\pm 0.1}$	85.4$_{\pm 0.2}$	85.3$_{\pm 0.2}$	85.3$_{\pm 0.3}$	85.8$_{\pm 0.2}$	85.7$_{\pm 0.2}$	85.8$_{\pm 0.3}$	**85.9**$_{\pm 0.2}$	85.7$_{\pm 0.2}$	85.7$_{\pm 0.1}$
Swedish	89.2$_{\pm 0.4}$	90.0$_{\pm 0.2}$	89.9$_{\pm 0.2}$	**90.1**$_{\pm 0.2}$	90.0$_{\pm 0.3}$	**90.1**$_{\pm 0.2}$	**90.1**$_{\pm 0.2}$	89.8$_{\pm 0.1}$	89.9$_{\pm 0.1}$	89.9$_{\pm 0.1}$	89.8$_{\pm 0.1}$	89.8$_{\pm 0.1}$	89.8$_{\pm 0.2}$
Tamil	51.9$_{\pm 1.0}$	54.6$_{\pm 1.2}$	**55.3**$_{\pm 0.7}$	55.0$_{\pm 0.4}$	55.0$_{\pm 1.3}$	55.1$_{\pm 0.6}$	55.0$_{\pm 0.6}$	54.7$_{\pm 0.7}$	55.0$_{\pm 0.4}$	55.0$_{\pm 0.5}$	54.4$_{\pm 0.4}$	55.2$_{\pm 0.4}$	**55.3**$_{\pm 0.3}$
Thai	31.4$_{\pm 6.0}$	52.7$_{\pm 0.8}$	**53.4**$_{\pm 1.1}$	52.9$_{\pm 1.3}$	52.8$_{\pm 1.0}$	53.0$_{\pm 1.3}$	52.8$_{\pm 0.5}$	49.8$_{\pm 0.8}$	50.3$_{\pm 1.3}$	50.7$_{\pm 0.9}$	49.5$_{\pm 1.6}$	50.2$_{\pm 0.4}$	50.8$_{\pm 0.6}$
Turkish	70.0$_{\pm 0.7}$	71.0$_{\pm 0.4}$	70.9$_{\pm 0.2}$	70.9$_{\pm 0.5}$	70.7$_{\pm 0.5}$	70.8$_{\pm 0.6}$	70.8$_{\pm 0.4}$	**71.4**$_{\pm 0.3}$	71.2$_{\pm 0.2}$	71.2$_{\pm 0.3}$	71.0$_{\pm 0.3}$	71.2$_{\pm 0.2}$	71.2$_{\pm 0.3}$
Ukrainian	81.4$_{\pm 0.3}$	**84.9**$_{\pm 0.3}$	84.7$_{\pm 0.3}$	84.7$_{\pm 0.5}$	84.7$_{\pm 0.3}$	84.5$_{\pm 0.5}$	84.5$_{\pm 0.3}$	84.4$_{\pm 0.5}$	84.6$_{\pm 0.4}$	84.5$_{\pm 0.2}$	84.5$_{\pm 0.2}$	84.4$_{\pm 0.3}$	84.3$_{\pm 0.3}$
Vietnamese	57.5$_{\pm 0.8}$	56.4$_{\pm 0.4}$	57.0$_{\pm 0.6}$	56.7$_{\pm 0.4}$	56.7$_{\pm 0.4}$	57.1$_{\pm 0.5}$	57.1$_{\pm 0.2}$	58.9$_{\pm 0.4}$	59.0$_{\pm 0.4}$	59.2$_{\pm 0.4}$	59.0$_{\pm 0.3}$	59.2$_{\pm 0.5}$	**59.3**$_{\pm 0.6}$
Average	73.8$_{\pm 0.6}$	**77.2**$_{\pm 0.2}$	**77.2**$_{\pm 0.1}$	**77.2**$_{\pm 0.2}$	**77.2**$_{\pm 0.1}$	**77.2**$_{\pm 0.2}$	**77.2**$_{\pm 0.2}$	77.0$_{\pm 0.3}$	77.1$_{\pm 0.3}$	77.1$_{\pm 0.2}$	77.0$_{\pm 0.1}$	77.1$_{\pm 0.2}$	77.1$_{\pm 0.2}$

Table 20: PoS tagging average accuracy results across 5 seeds using DistilMBERT by freezing strategy, language, and filtering threshold. Aligner name: FA - FastAlign. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment						ALIGNFREEZE with front-freezing					
	-	0%	25%	37%	50%	62%	75%	0%	25%	37%	50%	62%	75%
Afrikaans	85.5$_{\pm 0.2}$	86.4$_{\pm 0.3}$	86.5$_{\pm 0.3}$	86.6$_{\pm 0.2}$	86.5$_{\pm 0.2}$	86.6$_{\pm 0.3}$	**86.7**$_{\pm 0.2}$	86.3$_{\pm 0.3}$	86.5$_{\pm 0.2}$	86.4$_{\pm 0.3}$	86.5$_{\pm 0.2}$	86.4$_{\pm 0.3}$	86.5$_{\pm 0.1}$
Arabic	51.7$_{\pm 1.7}$	63.6$_{\pm 0.3}$	63.3$_{\pm 0.2}$	63.7$_{\pm 0.6}$	63.3$_{\pm 0.5}$	63.7$_{\pm 0.7}$	63.6$_{\pm 0.7}$	63.0$_{\pm 0.5}$	63.4$_{\pm 0.4}$	**63.8**$_{\pm 0.7}$	63.7$_{\pm 0.5}$	63.7$_{\pm 0.6}$	**63.8**$_{\pm 0.5}$
Bulgarian	85.0$_{\pm 0.5}$	**87.6**$_{\pm 0.3}$	87.4$_{\pm 0.4}$	87.3$_{\pm 0.3}$	87.3$_{\pm 0.2}$	87.2$_{\pm 0.3}$	87.2$_{\pm 0.2}$	87.3$_{\pm 0.2}$	87.3$_{\pm 0.3}$	87.3$_{\pm 0.2}$	87.3$_{\pm 0.3}$	87.2$_{\pm 0.2}$	87.3$_{\pm 0.2}$
Catalan	86.6$_{\pm 0.4}$	**88.1**$_{\pm 0.2}$	88.0$_{\pm 0.2}$	87.8$_{\pm 0.1}$	87.9$_{\pm 0.1}$	87.9$_{\pm 0.1}$	87.8$_{\pm 0.2}$	87.8$_{\pm 0.2}$	87.8$_{\pm 0.2}$	87.6$_{\pm 0.1}$	87.6$_{\pm 0.1}$	87.7$_{\pm 0.2}$	87.7$_{\pm 0.2}$
Chinese	64.3$_{\pm 1.4}$	66.3$_{\pm 0.6}$	**66.5**$_{\pm 0.4}$	66.3$_{\pm 0.3}$	66.2$_{\pm 0.4}$	66.2$_{\pm 0.5}$	66.0$_{\pm 0.5}$	66.3$_{\pm 0.4}$	66.4$_{\pm 0.5}$	**66.5**$_{\pm 0.3}$	66.3$_{\pm 0.2}$	66.4$_{\pm 0.5}$	66.3$_{\pm 0.6}$
Czech	79.1$_{\pm 0.7}$	**84.7**$_{\pm 0.4}$	84.5$_{\pm 0.3}$	84.6$_{\pm 0.4}$	84.5$_{\pm 0.3}$	84.5$_{\pm 0.2}$	84.5$_{\pm 0.3}$	84.0$_{\pm 0.2}$	84.1$_{\pm 0.2}$	84.0$_{\pm 0.3}$	84.0$_{\pm 0.2}$	84.1$_{\pm 0.1}$	84.0$_{\pm 0.1}$
Danish	87.8$_{\pm 0.3}$	88.2$_{\pm 0.2}$	88.2$_{\pm 0.4}$	88.2$_{\pm 0.2}$	88.2$_{\pm 0.3}$	88.2$_{\pm 0.2}$	88.0$_{\pm 0.2}$	**88.7**$_{\pm 0.2}$	**88.7**$_{\pm 0.2}$	**88.7**$_{\pm 0.1}$	**88.7**$_{\pm 0.2}$	**88.7**$_{\pm 0.2}$	88.6$_{\pm 0.2}$
Finnish	82.3$_{\pm 0.8}$	84.1$_{\pm 0.4}$	84.3$_{\pm 0.2}$	84.2$_{\pm 0.4}$	84.1$_{\pm 0.5}$	84.3$_{\pm 0.4}$	83.8$_{\pm 0.5}$	84.7$_{\pm 0.2}$	**84.8**$_{\pm 0.2}$	84.7$_{\pm 0.2}$	**84.8**$_{\pm 0.2}$	84.7$_{\pm 0.2}$	**84.8**$_{\pm 0.2}$
French	85.4$_{\pm 0.2}$	**86.5**$_{\pm 0.2}$	**86.5**$_{\pm 0.1}$	86.4$_{\pm 0.1}$	86.4$_{\pm 0.1}$	**86.5**$_{\pm 0.2}$	86.3$_{\pm 0.2}$	**86.5**$_{\pm 0.2}$	**86.5**$_{\pm 0.2}$	**86.5**$_{\pm 0.1}$	**86.5**$_{\pm 0.2}$	86.4$_{\pm 0.2}$	86.4$_{\pm 0.2}$
German	87.4$_{\pm 0.4}$	88.5$_{\pm 0.1}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.1}$	88.6$_{\pm 0.1}$	**88.7**$_{\pm 0.2}$	88.2$_{\pm 0.1}$	88.2$_{\pm 0.3}$	88.2$_{\pm 0.2}$	88.3$_{\pm 0.2}$	88.2$_{\pm 0.2}$	88.3$_{\pm 0.2}$
Greek	74.9$_{\pm 1.2}$	**78.6**$_{\pm 0.7}$	**78.6**$_{\pm 0.7}$	78.5$_{\pm 0.1}$	78.2$_{\pm 0.8}$	78.5$_{\pm 0.8}$	78.4$_{\pm 0.5}$	78.1$_{\pm 0.5}$	77.9$_{\pm 0.5}$	78.0$_{\pm 0.4}$	78.1$_{\pm 0.4}$	77.9$_{\pm 0.6}$	78.2$_{\pm 0.5}$
Hebrew	62.3$_{\pm 0.9}$	64.0$_{\pm 1.0}$	64.0$_{\pm 0.6}$	64.3$_{\pm 0.7}$	63.6$_{\pm 0.4}$	64.5$_{\pm 0.9}$	64.0$_{\pm 0.5}$	64.8$_{\pm 0.6}$	64.8$_{\pm 0.6}$	64.8$_{\pm 0.5}$	64.7$_{\pm 0.4}$	**64.9**$_{\pm 0.4}$	64.6$_{\pm 1.0}$
Hindi	60.7$_{\pm 3.2}$	64.8$_{\pm 1.3}$	64.4$_{\pm 1.5}$	64.9$_{\pm 1.5}$	64.8$_{\pm 1.4}$	**65.2**$_{\pm 0.5}$	65.0$_{\pm 1.2}$	63.2$_{\pm 2.0}$	63.8$_{\pm 2.4}$	63.1$_{\pm 2.4}$	63.0$_{\pm 1.9}$	64.1$_{\pm 0.6}$	64.1$_{\pm 1.3}$
Hungarian	79.1$_{\pm 0.2}$	81.1$_{\pm 0.4}$	**81.5**$_{\pm 0.2}$	81.2$_{\pm 0.5}$	80.9$_{\pm 0.5}$	81.1$_{\pm 0.3}$	81.2$_{\pm 0.3}$	80.9$_{\pm 0.1}$	81.2$_{\pm 0.2}$	81.2$_{\pm 0.4}$	81.2$_{\pm 0.2}$	81.2$_{\pm 0.2}$	81.1$_{\pm 0.1}$
Italian	85.0$_{\pm 0.4}$	85.6$_{\pm 0.1}$	85.5$_{\pm 0.1}$	85.4$_{\pm 0.2}$	85.4$_{\pm 0.2}$	85.5$_{\pm 0.1}$	85.4$_{\pm 0.1}$	**85.7**$_{\pm 0.2}$	**85.7**$_{\pm 0.2}$	85.5$_{\pm 0.1}$	**85.7**$_{\pm 0.1}$	**85.7**$_{\pm 0.2}$	**85.7**$_{\pm 0.1}$
Japanese	47.8$_{\pm 2.1}$	53.0$_{\pm 1.5}$	52.9$_{\pm 1.1}$	**53.5**$_{\pm 1.1}$	52.9$_{\pm 1.6}$	53.3$_{\pm 1.2}$	**53.5**$_{\pm 1.3}$	49.8$_{\pm 1.5}$	49.8$_{\pm 1.0}$	49.8$_{\pm 1.0}$	49.0$_{\pm 1.0}$	49.9$_{\pm 0.7}$	49.9$_{\pm 1.1}$
Korean	55.4$_{\pm 2.7}$	59.9$_{\pm 1.9}$	60.2$_{\pm 1.3}$	60.7$_{\pm 1.0}$	59.8$_{\pm 1.0}$	60.6$_{\pm 1.4}$	59.7$_{\pm 1.0}$	60.2$_{\pm 1.4}$	**61.5**$_{\pm 0.7}$	61.2$_{\pm 1.0}$	60.2$_{\pm 1.1}$	61.1$_{\pm 1.3}$	61.4$_{\pm 0.9}$
Latvian	69.5$_{\pm 2.0}$	77.3$_{\pm 0.2}$	77.5$_{\pm 0.3}$	77.7$_{\pm 0.2}$	77.6$_{\pm 0.3}$	**77.8**$_{\pm 0.3}$	77.7$_{\pm 0.2}$	76.0$_{\pm 0.3}$	76.1$_{\pm 0.2}$	76.1$_{\pm 0.2}$	76.2$_{\pm 0.3}$	76.3$_{\pm 0.2}$	76.3$_{\pm 0.1}$
Lithuanian	71.6$_{\pm 1.8}$	78.0$_{\pm 0.4}$	78.1$_{\pm 0.2}$	78.0$_{\pm 0.3}$	**78.2**$_{\pm 0.3}$	78.1$_{\pm 0.5}$	77.8$_{\pm 0.2}$	77.0$_{\pm 0.5}$	77.2$_{\pm 0.3}$	77.0$_{\pm 0.3}$	77.1$_{\pm 0.3}$	77.1$_{\pm 0.2}$	77.3$_{\pm 0.2}$
Norwegian	88.7$_{\pm 0.4}$	90.3$_{\pm 0.2}$	90.3$_{\pm 0.2}$	90.3$_{\pm 0.2}$	90.3$_{\pm 0.2}$	**90.4**$_{\pm 0.2}$	90.3$_{\pm 0.2}$	89.5$_{\pm 0.3}$	89.7$_{\pm 0.3}$	89.6$_{\pm 0.2}$	89.7$_{\pm 0.2}$	89.7$_{\pm 0.2}$	89.7$_{\pm 0.2}$
Persian	72.6$_{\pm 0.7}$	71.9$_{\pm 0.4}$	71.7$_{\pm 0.9}$	72.0$_{\pm 0.7}$	71.5$_{\pm 0.5}$	71.7$_{\pm 0.4}$	71.2$_{\pm 0.4}$	73.3$_{\pm 0.3}$	73.5$_{\pm 0.5}$	**73.6**$_{\pm 0.6}$	73.4$_{\pm 0.3}$	73.4$_{\pm 0.4}$	73.3$_{\pm 0.2}$
Polish	79.7$_{\pm 0.3}$	83.6$_{\pm 0.2}$	**83.8**$_{\pm 0.3}$	83.6$_{\pm 0.3}$	83.7$_{\pm 0.2}$	83.7$_{\pm 0.1}$	83.7$_{\pm 0.3}$	83.5$_{\pm 0.2}$	83.6$_{\pm 0.2}$	83.6$_{\pm 0.2}$	83.6$_{\pm 0.3}$	83.5$_{\pm 0.2}$	83.5$_{\pm 0.3}$
Portuguese	83.0$_{\pm 0.3}$	83.4$_{\pm 0.1}$	**83.5**$_{\pm 0.1}$	83.4$_{\pm 0.2}$	**83.5**$_{\pm 0.1}$	83.4$_{\pm 0.1}$	83.4$_{\pm 0.2}$	**83.5**$_{\pm 0.1}$	**83.5**$_{\pm 0.2}$	**83.5**$_{\pm 0.1}$	83.4$_{\pm 0.1}$	**83.5**$_{\pm 0.1}$	**83.5**$_{\pm 0.1}$
Romanian	80.0$_{\pm 0.5}$	**83.8**$_{\pm 0.3}$	83.6$_{\pm 0.3}$	83.7$_{\pm 0.2}$	83.6$_{\pm 0.3}$	83.7$_{\pm 0.1}$	83.6$_{\pm 0.2}$	83.4$_{\pm 0.2}$	83.4$_{\pm 0.2}$	83.3$_{\pm 0.3}$	83.4$_{\pm 0.2}$	83.4$_{\pm 0.3}$	83.4$_{\pm 0.3}$
Russian	81.5$_{\pm 0.6}$	83.8$_{\pm 0.5}$	83.6$_{\pm 0.4}$	83.4$_{\pm 0.5}$	83.4$_{\pm 0.4}$	83.4$_{\pm 0.4}$	83.1$_{\pm 0.5}$	**84.0**$_{\pm 0.5}$	83.9$_{\pm 0.6}$	83.9$_{\pm 0.3}$	83.8$_{\pm 0.5}$	83.9$_{\pm 0.6}$	83.7$_{\pm 0.6}$
Slovak	78.2$_{\pm 0.8}$	84.6$_{\pm 0.4}$	**84.7**$_{\pm 0.3}$	**84.7**$_{\pm 0.4}$	84.6$_{\pm 0.3}$	84.5$_{\pm 0.4}$	84.6$_{\pm 0.3}$	84.0$_{\pm 0.3}$	84.1$_{\pm 0.2}$	84.0$_{\pm 0.4}$	84.1$_{\pm 0.4}$	84.1$_{\pm 0.2}$	84.0$_{\pm 0.2}$
Slovenian	79.6$_{\pm 0.5}$	**83.8**$_{\pm 0.3}$	83.6$_{\pm 0.1}$	83.5$_{\pm 0.5}$	83.5$_{\pm 0.1}$	83.3$_{\pm 0.4}$	83.3$_{\pm 0.3}$	83.7$_{\pm 0.2}$	83.6$_{\pm 0.2}$	83.4$_{\pm 0.3}$	83.5$_{\pm 0.2}$	83.5$_{\pm 0.2}$	83.4$_{\pm 0.3}$
Spanish	84.4$_{\pm 0.4}$	85.6$_{\pm 0.1}$	85.6$_{\pm 0.1}$	85.5$_{\pm 0.2}$	85.6$_{\pm 0.2}$	85.6$_{\pm 0.1}$	85.4$_{\pm 0.1}$	**85.8**$_{\pm 0.2}$	85.7$_{\pm 0.1}$	85.7$_{\pm 0.2}$	85.7$_{\pm 0.1}$	85.7$_{\pm 0.1}$	85.6$_{\pm 0.2}$
Swedish	89.2$_{\pm 0.4}$	90.1$_{\pm 0.2}$	90.1$_{\pm 0.3}$	90.0$_{\pm 0.2}$	90.1$_{\pm 0.2}$	**90.2**$_{\pm 0.1}$	90.0$_{\pm 0.1}$	89.8$_{\pm 0.1}$	89.9$_{\pm 0.1}$	89.9$_{\pm 0.1}$	89.9$_{\pm 0.1}$	89.9$_{\pm 0.1}$	89.8$_{\pm 0.1}$
Tamil	51.9$_{\pm 1.0}$	55.5$_{\pm 0.7}$	56.0$_{\pm 0.5}$	55.7$_{\pm 0.5}$	56.0$_{\pm 0.5}$	56.2$_{\pm 0.8}$	**56.3**$_{\pm 0.6}$	55.4$_{\pm 0.1}$	55.6$_{\pm 0.3}$	55.4$_{\pm 0.6}$	55.3$_{\pm 0.5}$	55.4$_{\pm 0.4}$	55.6$_{\pm 0.3}$
Thai	31.4$_{\pm 6.0}$	52.9$_{\pm 1.4}$	54.6$_{\pm 1.3}$	**54.8**$_{\pm 1.3}$	53.5$_{\pm 1.2}$	53.1$_{\pm 1.9}$	53.5$_{\pm 1.5}$	51.3$_{\pm 0.9}$	51.8$_{\pm 0.5}$	51.9$_{\pm 0.8}$	51.5$_{\pm 1.5}$	51.8$_{\pm 1.4}$	51.4$_{\pm 1.2}$
Turkish	70.0$_{\pm 0.7}$	70.4$_{\pm 0.3}$	70.2$_{\pm 0.4}$	70.5$_{\pm 0.3}$	70.2$_{\pm 0.2}$	70.5$_{\pm 0.5}$	70.4$_{\pm 0.1}$	70.9$_{\pm 0.3}$	**71.0**$_{\pm 0.2}$	70.9$_{\pm 0.5}$	70.9$_{\pm 0.3}$	70.9$_{\pm 0.3}$	**71.0**$_{\pm 0.3}$
Ukrainian	81.4$_{\pm 0.3}$	**85.0**$_{\pm 0.4}$	84.7$_{\pm 0.2}$	84.9$_{\pm 0.3}$	84.8$_{\pm 0.2}$	84.8$_{\pm 0.3}$	84.6$_{\pm 0.4}$	84.6$_{\pm 0.2}$	84.6$_{\pm 0.3}$	84.7$_{\pm 0.2}$	84.5$_{\pm 0.3}$	84.6$_{\pm 0.3}$	84.5$_{\pm 0.3}$
Vietnamese	57.5$_{\pm 0.8}$	56.9$_{\pm 0.6}$	57.3$_{\pm 0.4}$	57.2$_{\pm 0.7}$	57.0$_{\pm 0.5}$	57.3$_{\pm 0.5}$	57.2$_{\pm 0.4}$	58.8$_{\pm 0.5}$	**59.4**$_{\pm 0.5}$	59.3$_{\pm 0.6}$	59.2$_{\pm 0.3}$	**59.4**$_{\pm 0.6}$	**59.4**$_{\pm 0.4}$
Average	73.8$_{\pm 0.6}$	77.3$_{\pm 0.2}$	77.3$_{\pm 0.1}$	**77.4**$_{\pm 0.2}$	77.2$_{\pm 0.2}$	**77.4**$_{\pm 0.2}$	77.2$_{\pm 0.2}$	77.1$_{\pm 0.2}$	77.2$_{\pm 0.2}$	77.1$_{\pm 0.2}$	77.1$_{\pm 0.2}$	77.2$_{\pm 0.2}$	77.2$_{\pm 0.2}$

Table 21: PoS tagging average accuracy results across 5 seeds using DistilMBERT by freezing strategy, language, and filtering threshold. Aligner name: AA - AWESOME-align. The highest average accuracy value for each language is highlighted in bold.

	FT Only	vanilla realignment						ALIGNFREEZE with front-freezing					
	-	0%	25%	37%	50%	62%	75%	0%	25%	37%	50%	62%	75%
Afrikaans	85.5$_{\pm 0.2}$	85.6$_{\pm 0.4}$	86.1$_{\pm 0.3}$	85.8$_{\pm 0.4}$	85.7$_{\pm 0.4}$	85.7$_{\pm 0.4}$	85.6$_{\pm 0.2}$	86.1$_{\pm 0.3}$	**86.3**$_{\pm 0.3}$	86.0$_{\pm 0.2}$	86.2$_{\pm 0.1}$	**86.3**$_{\pm 0.2}$	86.1$_{\pm 0.3}$
Arabic	51.7$_{\pm 1.7}$	**66.6**$_{\pm 0.5}$	66.3$_{\pm 0.9}$	65.9$_{\pm 0.6}$	65.8$_{\pm 0.7}$	66.1$_{\pm 0.3}$	65.2$_{\pm 0.8}$	65.0$_{\pm 0.6}$	65.1$_{\pm 0.8}$	64.8$_{\pm 0.8}$	65.0$_{\pm 0.7}$	64.9$_{\pm 0.6}$	64.2$_{\pm 0.3}$
Bulgarian	85.0$_{\pm 0.5}$	**87.6**$_{\pm 0.4}$	**87.6**$_{\pm 0.4}$	87.4$_{\pm 0.3}$	87.5$_{\pm 0.2}$	87.4$_{\pm 0.1}$	87.2$_{\pm 0.3}$	87.2$_{\pm 0.3}$	87.2$_{\pm 0.3}$	87.2$_{\pm 0.3}$	87.2$_{\pm 0.4}$	87.1$_{\pm 0.3}$	87.0$_{\pm 0.3}$
Catalan	86.6$_{\pm 0.4}$	88.4$_{\pm 0.1}$	88.4$_{\pm 0.2}$	**88.5**$_{\pm 0.1}$	88.4$_{\pm 0.1}$	88.4$_{\pm 0.1}$	88.4$_{\pm 0.1}$	88.2$_{\pm 0.1}$	88.2$_{\pm 0.2}$	88.2$_{\pm 0.2}$	88.2$_{\pm 0.1}$	88.1$_{\pm 0.2}$	88.1$_{\pm 0.1}$
Chinese	64.3$_{\pm 1.4}$	**67.4**$_{\pm 0.7}$	66.9$_{\pm 0.6}$	**67.4**$_{\pm 0.6}$	67.0$_{\pm 0.5}$	67.3$_{\pm 0.9}$	66.9$_{\pm 0.8}$	67.3$_{\pm 0.6}$	67.1$_{\pm 0.4}$	67.3$_{\pm 0.5}$	67.2$_{\pm 0.5}$	67.2$_{\pm 0.7}$	67.1$_{\pm 0.7}$
Czech	79.1$_{\pm 0.7}$	85.3$_{\pm 0.5}$	**85.4**$_{\pm 0.3}$	85.2$_{\pm 0.4}$	85.2$_{\pm 0.4}$	85.2$_{\pm 0.4}$	84.9$_{\pm 0.2}$	84.3$_{\pm 0.3}$	84.3$_{\pm 0.3}$	84.2$_{\pm 0.5}$	84.1$_{\pm 0.3}$	84.0$_{\pm 0.3}$	83.8$_{\pm 0.2}$
Danish	87.8$_{\pm 0.3}$	88.3$_{\pm 0.2}$	88.2$_{\pm 0.1}$	88.1$_{\pm 0.2}$	88.2$_{\pm 0.2}$	88.0$_{\pm 0.2}$	88.0$_{\pm 0.2}$	**88.7**$_{\pm 0.2}$	**88.7**$_{\pm 0.1}$	88.6$_{\pm 0.3}$	**88.7**$_{\pm 0.2}$	88.6$_{\pm 0.3}$	88.5$_{\pm 0.2}$
Finnish	82.3$_{\pm 0.8}$	84.1$_{\pm 0.3}$	84.3$_{\pm 0.4}$	84.4$_{\pm 0.3}$	83.9$_{\pm 0.3}$	84.2$_{\pm 0.5}$	84.3$_{\pm 0.3}$	84.8$_{\pm 0.2}$	**84.9**$_{\pm 0.3}$	84.8$_{\pm 0.2}$	84.7$_{\pm 0.2}$	84.6$_{\pm 0.2}$	84.6$_{\pm 0.2}$
French	85.4$_{\pm 0.2}$	86.6$_{\pm 0.1}$	86.6$_{\pm 0.2}$	**86.7**$_{\pm 0.1}$	86.6$_{\pm 0.2}$	86.5$_{\pm 0.1}$	86.5$_{\pm 0.2}$	86.6$_{\pm 0.2}$	86.6$_{\pm 0.3}$	86.6$_{\pm 0.3}$	**86.7**$_{\pm 0.1}$	86.6$_{\pm 0.2}$	86.5$_{\pm 0.2}$
German	87.4$_{\pm 0.4}$	89.0$_{\pm 0.2}$	**89.1**$_{\pm 0.1}$	88.9$_{\pm 0.1}$	89.0$_{\pm 0.2}$	**89.1**$_{\pm 0.2}$	89.0$_{\pm 0.1}$	88.4$_{\pm 0.1}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.1}$	88.5$_{\pm 0.2}$	88.5$_{\pm 0.2}$
Greek	74.9$_{\pm 1.2}$	**80.1**$_{\pm 0.5}$	**80.1**$_{\pm 0.5}$	80.0$_{\pm 0.9}$	79.8$_{\pm 0.7}$	79.3$_{\pm 0.7}$	79.4$_{\pm 0.6}$	77.9$_{\pm 0.6}$	78.1$_{\pm 0.8}$	78.1$_{\pm 0.6}$	77.9$_{\pm 0.8}$	78.1$_{\pm 1.2}$	77.6$_{\pm 0.6}$
Hebrew	62.3$_{\pm 0.9}$	65.2$_{\pm 0.1}$	64.9$_{\pm 0.5}$	64.7$_{\pm 0.8}$	64.3$_{\pm 0.7}$	64.6$_{\pm 0.6}$	64.1$_{\pm 0.3}$	**65.6**$_{\pm 0.6}$	65.3$_{\pm 0.4}$	65.4$_{\pm 0.7}$	64.9$_{\pm 0.5}$	64.9$_{\pm 0.8}$	64.5$_{\pm 0.4}$
Hindi	60.7$_{\pm 3.2}$	65.9$_{\pm 3.3}$	65.9$_{\pm 2.4}$	65.9$_{\pm 2.4}$	65.5$_{\pm 2.5}$	65.7$_{\pm 3.2}$	**66.1**$_{\pm 3.0}$	63.8$_{\pm 2.2}$	63.8$_{\pm 2.5}$	64.0$_{\pm 2.4}$	64.5$_{\pm 1.9}$	63.9$_{\pm 2.8}$	63.7$_{\pm 2.4}$
Hungarian	79.1$_{\pm 0.2}$	81.9$_{\pm 0.3}$	**82.2**$_{\pm 0.8}$	82.0$_{\pm 0.4}$	81.8$_{\pm 0.4}$	81.8$_{\pm 0.3}$	81.6$_{\pm 0.5}$	81.4$_{\pm 0.1}$	81.5$_{\pm 0.4}$	81.5$_{\pm 0.3}$	81.4$_{\pm 0.2}$	81.3$_{\pm 0.2}$	81.2$_{\pm 0.3}$
Italian	85.0$_{\pm 0.4}$	85.9$_{\pm 0.1}$	85.9$_{\pm 0.1}$	85.9$_{\pm 0.2}$	85.9$_{\pm 0.1}$	85.8$_{\pm 0.2}$	85.8$_{\pm 0.1}$	86.0$_{\pm 0.2}$	**86.2**$_{\pm 0.2}$	86.1$_{\pm 0.2}$	86.0$_{\pm 0.2}$	86.0$_{\pm 0.2}$	85.8$_{\pm 0.2}$
Japanese	47.8$_{\pm 2.1}$	52.7$_{\pm 2.0}$	**52.8**$_{\pm 2.3}$	51.9$_{\pm 2.0}$	52.0$_{\pm 1.4}$	52.7$_{\pm 1.5}$	51.7$_{\pm 1.7}$	49.4$_{\pm 1.4}$	49.6$_{\pm 1.4}$	49.9$_{\pm 1.0}$	49.6$_{\pm 0.9}$	50.0$_{\pm 1.3}$	49.5$_{\pm 1.2}$
Korean	55.4$_{\pm 2.7}$	61.8$_{\pm 1.0}$	62.3$_{\pm 1.4}$	62.9$_{\pm 1.2}$	61.8$_{\pm 0.4}$	62.6$_{\pm 0.9}$	62.4$_{\pm 1.1}$	63.0$_{\pm 1.3}$	63.3$_{\pm 1.0}$	**63.8**$_{\pm 1.4}$	63.5$_{\pm 1.2}$	63.5$_{\pm 1.6}$	63.5$_{\pm 1.1}$
Latvian	69.5$_{\pm 2.0}$	**76.2**$_{\pm 0.6}$	76.1$_{\pm 0.6}$	75.9$_{\pm 0.4}$	76.0$_{\pm 0.6}$	75.7$_{\pm 0.5}$	75.9$_{\pm 0.1}$	75.3$_{\pm 0.1}$	75.3$_{\pm 0.2}$	75.2$_{\pm 0.2}$	75.2$_{\pm 0.4}$	75.2$_{\pm 0.4}$	74.9$_{\pm 0.3}$
Lithuanian	71.6$_{\pm 1.8}$	**76.3**$_{\pm 0.7}$	75.8$_{\pm 0.5}$	75.9$_{\pm 0.3}$	76.0$_{\pm 0.2}$	76.0$_{\pm 0.2}$	76.0$_{\pm 0.4}$	75.9$_{\pm 0.3}$	75.8$_{\pm 0.3}$	75.9$_{\pm 0.3}$	75.6$_{\pm 0.4}$	75.9$_{\pm 0.5}$	75.7$_{\pm 0.4}$
Norwegian	88.7$_{\pm 0.4}$	90.1$_{\pm 0.2}$	90.3$_{\pm 0.1}$	90.3$_{\pm 0.1}$	90.1$_{\pm 0.3}$	**90.4**$_{\pm 0.2}$	90.2$_{\pm 0.2}$	89.5$_{\pm 0.3}$	89.6$_{\pm 0.3}$	89.6$_{\pm 0.2}$	89.6$_{\pm 0.3}$	89.6$_{\pm 0.3}$	89.6$_{\pm 0.2}$
Persian	72.6$_{\pm 0.7}$	72.2$_{\pm 0.6}$	72.1$_{\pm 0.5}$	72.5$_{\pm 0.7}$	71.9$_{\pm 0.5}$	71.8$_{\pm 0.6}$	71.9$_{\pm 0.7}$	73.8$_{\pm 0.4}$	**73.9**$_{\pm 0.2}$	**73.9**$_{\pm 0.2}$	73.6$_{\pm 0.6}$	73.7$_{\pm 0.5}$	73.6$_{\pm 0.2}$
Polish	79.7$_{\pm 0.3}$	83.5$_{\pm 0.3}$	**83.6**$_{\pm 0.3}$	83.5$_{\pm 0.3}$	83.5$_{\pm 0.2}$	**83.6**$_{\pm 0.1}$	83.4$_{\pm 0.2}$	83.5$_{\pm 0.3}$	83.5$_{\pm 0.3}$	83.5$_{\pm 0.2}$	**83.6**$_{\pm 0.4}$	83.5$_{\pm 0.3}$	83.4$_{\pm 0.2}$
Portuguese	83.0$_{\pm 0.3}$	**84.1**$_{\pm 0.1}$	84.0$_{\pm 0.1}$	84.0$_{\pm 0.1}$	84.0$_{\pm 0.1}$	84.0$_{\pm 0.1}$	84.0$_{\pm 0.1}$	83.9$_{\pm 0.0}$	83.9$_{\pm 0.1}$	83.9$_{\pm 0.1}$	83.9$_{\pm 0.1}$	83.8$_{\pm 0.1}$	83.9$_{\pm 0.1}$
Romanian	80.0$_{\pm 0.5}$	83.4$_{\pm 0.5}$	83.4$_{\pm 0.4}$	83.4$_{\pm 0.4}$	**83.6**$_{\pm 0.3}$	83.4$_{\pm 0.2}$	83.5$_{\pm 0.4}$	83.0$_{\pm 0.4}$	83.1$_{\pm 0.5}$	83.0$_{\pm 0.3}$	83.2$_{\pm 0.4}$	83.2$_{\pm 0.3}$	83.0$_{\pm 0.4}$
Russian	81.5$_{\pm 0.6}$	**84.9**$_{\pm 0.3}$	84.8$_{\pm 0.5}$	84.8$_{\pm 0.5}$	84.7$_{\pm 0.2}$	84.6$_{\pm 0.1}$	84.2$_{\pm 0.5}$	84.2$_{\pm 0.4}$	84.2$_{\pm 0.5}$	84.2$_{\pm 0.5}$	84.0$_{\pm 0.3}$	84.0$_{\pm 0.3}$	83.8$_{\pm 0.6}$
Slovak	78.2$_{\pm 0.8}$	85.0$_{\pm 0.6}$	**85.4**$_{\pm 0.5}$	85.2$_{\pm 0.6}$	85.2$_{\pm 0.3}$	85.2$_{\pm 0.2}$	84.8$_{\pm 0.2}$	84.3$_{\pm 0.6}$	84.4$_{\pm 0.4}$	84.2$_{\pm 0.8}$	84.2$_{\pm 0.4}$	84.0$_{\pm 0.4}$	83.8$_{\pm 0.3}$
Slovenian	79.6$_{\pm 0.5}$	83.8$_{\pm 0.3}$	**83.9**$_{\pm 0.3}$	83.8$_{\pm 0.2}$	**83.9**$_{\pm 0.3}$	83.8$_{\pm 0.3}$	83.6$_{\pm 0.2}$	83.6$_{\pm 0.3}$	83.7$_{\pm 0.2}$	83.6$_{\pm 0.2}$	83.6$_{\pm 0.2}$	83.5$_{\pm 0.3}$	83.4$_{\pm 0.3}$
Spanish	84.4$_{\pm 0.4}$	85.7$_{\pm 0.2}$	85.8$_{\pm 0.3}$	85.7$_{\pm 0.1}$	85.7$_{\pm 0.2}$	85.6$_{\pm 0.2}$	85.7$_{\pm 0.2}$	85.7$_{\pm 0.2}$	**85.9**$_{\pm 0.2}$	85.8$_{\pm 0.2}$	85.8$_{\pm 0.1}$	85.8$_{\pm 0.2}$	85.8$_{\pm 0.3}$
Swedish	89.2$_{\pm 0.4}$	90.0$_{\pm 0.2}$	90.0$_{\pm 0.1}$	90.0$_{\pm 0.2}$	90.0$_{\pm 0.2}$	**90.1**$_{\pm 0.1}$	89.8$_{\pm 0.2}$	89.8$_{\pm 0.1}$	89.9$_{\pm 0.2}$	89.8$_{\pm 0.1}$	89.8$_{\pm 0.1}$	89.8$_{\pm 0.2}$	89.7$_{\pm 0.1}$
Tamil	51.9$_{\pm 1.0}$	**55.8**$_{\pm 0.7}$	55.6$_{\pm 0.7}$	54.7$_{\pm 0.7}$	55.1$_{\pm 0.6}$	55.6$_{\pm 1.0}$	55.3$_{\pm 0.7}$	54.7$_{\pm 0.9}$	54.5$_{\pm 1.1}$	54.8$_{\pm 0.6}$	54.5$_{\pm 0.9}$	55.1$_{\pm 0.7}$	55.2$_{\pm 0.6}$
Thai	31.4$_{\pm 6.0}$	**55.2**$_{\pm 0.7}$	55.0$_{\pm 0.6}$	54.2$_{\pm 0.9}$	54.3$_{\pm 0.6}$	54.5$_{\pm 0.9}$	52.0$_{\pm 1.3}$	51.7$_{\pm 0.6}$	51.7$_{\pm 0.4}$	51.3$_{\pm 1.0}$	51.6$_{\pm 1.0}$	51.7$_{\pm 0.6}$	50.5$_{\pm 1.3}$
Turkish	70.0$_{\pm 0.7}$	70.4$_{\pm 0.5}$	70.7$_{\pm 0.4}$	70.9$_{\pm 0.3}$	70.3$_{\pm 0.5}$	70.6$_{\pm 0.7}$	70.2$_{\pm 0.9}$	71.3$_{\pm 0.3}$	71.3$_{\pm 0.4}$	71.3$_{\pm 0.3}$	**71.4**$_{\pm 0.2}$	71.3$_{\pm 0.4}$	71.3$_{\pm 0.4}$
Ukrainian	81.4$_{\pm 0.3}$	85.0$_{\pm 0.2}$	85.0$_{\pm 0.3}$	**85.1**$_{\pm 0.3}$	85.0$_{\pm 0.1}$	84.9$_{\pm 0.2}$	84.8$_{\pm 0.4}$	84.4$_{\pm 0.3}$	84.4$_{\pm 0.3}$	84.5$_{\pm 0.4}$	84.3$_{\pm 0.2}$	84.2$_{\pm 0.3}$	84.1$_{\pm 0.3}$
Vietnamese	57.5$_{\pm 0.8}$	57.7$_{\pm 0.4}$	57.3$_{\pm 0.6}$	57.7$_{\pm 0.8}$	57.2$_{\pm 0.9}$	57.4$_{\pm 0.7}$	57.2$_{\pm 0.5}$	**59.6**$_{\pm 0.6}$	59.2$_{\pm 0.6}$	**59.6**$_{\pm 0.6}$	59.2$_{\pm 0.6}$	59.5$_{\pm 0.6}$	59.1$_{\pm 0.7}$
Average	73.8$_{\pm 0.6}$	**77.7**$_{\pm 0.3}$	**77.7**$_{\pm 0.3}$	77.6$_{\pm 0.2}$	77.5$_{\pm 0.2}$	77.6$_{\pm 0.1}$	77.3$_{\pm 0.3}$	77.3$_{\pm 0.2}$	77.3$_{\pm 0.2}$	77.3$_{\pm 0.2}$	77.3$_{\pm 0.2}$	77.3$_{\pm 0.2}$	77.1$_{\pm 0.2}$

Table 22: PoS tagging average accuracy results across 5 seeds using DistilMBERT by freezing strategy, language, and filtering threshold. Aligner name: BD - Bilingual Dictionary. The highest average accuracy value for each language is highlighted in bold.

	FT Only	Layer 0	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
Afrikaans	85.5$_{\pm 0.2}$	85.7$_{\pm 0.3}$	85.6$_{\pm 0.3}$	85.6$_{\pm 0.3}$	85.4$_{\pm 0.3}$	85.7$_{\pm 0.3}$	**85.7**$_{\pm 0.5}$	85.7$_{\pm 0.2}$
Arabic	51.7$_{\pm 1.5}$	51.3$_{\pm 1.0}$	60.4$_{\pm 0.8}$	62.2$_{\pm 0.7}$	**62.2**$_{\pm 1.4}$	62.0$_{\pm 0.6}$	55.7$_{\pm 1.5}$	50.3$_{\pm 1.2}$
Bulgarian	85.0$_{\pm 0.4}$	85.9$_{\pm 0.4}$	86.7$_{\pm 0.4}$	86.8$_{\pm 0.2}$	**86.9**$_{\pm 0.2}$	86.8$_{\pm 0.3}$	86.3$_{\pm 0.3}$	84.9$_{\pm 0.2}$
Catalan	86.6$_{\pm 0.4}$	87.3$_{\pm 0.4}$	87.6$_{\pm 0.4}$	87.6$_{\pm 0.3}$	87.9$_{\pm 0.2}$	**87.9**$_{\pm 0.3}$	87.6$_{\pm 0.2}$	86.7$_{\pm 0.4}$
Chinese	64.3$_{\pm 1.2}$	65.5$_{\pm 0.7}$	66.6$_{\pm 0.5}$	**66.7**$_{\pm 0.7}$	66.6$_{\pm 0.9}$	66.5$_{\pm 0.8}$	66.5$_{\pm 0.6}$	64.4$_{\pm 0.6}$
Czech	79.1$_{\pm 0.6}$	81.5$_{\pm 0.6}$	**84.2**$_{\pm 0.4}$	84.0$_{\pm 0.2}$	83.8$_{\pm 0.3}$	83.2$_{\pm 0.3}$	82.3$_{\pm 0.2}$	79.5$_{\pm 0.3}$
Danish	87.8$_{\pm 0.3}$	87.7$_{\pm 0.4}$	88.0$_{\pm 0.4}$	87.9$_{\pm 0.3}$	88.0$_{\pm 0.3}$	**88.4**$_{\pm 0.2}$	88.3$_{\pm 0.3}$	87.5$_{\pm 0.4}$
English	96.0$_{\pm 0.1}$	**96.1**$_{\pm 0.1}$	**96.1**$_{\pm 0.1}$	96.0$_{\pm 0.1}$	96.0$_{\pm 0.0}$	**96.1**$_{\pm 0.1}$	**96.1**$_{\pm 0.1}$	96.1$_{\pm 0.0}$
Finnish	82.3$_{\pm 0.7}$	83.3$_{\pm 0.4}$	83.9$_{\pm 0.3}$	84.0$_{\pm 0.2}$	84.1$_{\pm 0.4}$	**84.3**$_{\pm 0.3}$	83.6$_{\pm 0.3}$	82.0$_{\pm 0.5}$
French	85.4$_{\pm 0.2}$	85.6$_{\pm 0.4}$	86.0$_{\pm 0.3}$	86.2$_{\pm 0.3}$	86.3$_{\pm 0.2}$	**86.4**$_{\pm 0.2}$	86.2$_{\pm 0.3}$	85.4$_{\pm 0.3}$
German	87.4$_{\pm 0.3}$	87.9$_{\pm 0.4}$	88.1$_{\pm 0.2}$	**88.2**$_{\pm 0.3}$	88.0$_{\pm 0.3}$	88.0$_{\pm 0.2}$	87.7$_{\pm 0.3}$	87.5$_{\pm 0.4}$
Greek	74.9$_{\pm 1.1}$	76.6$_{\pm 1.2}$	**78.3**$_{\pm 1.0}$	78.2$_{\pm 0.7}$	77.9$_{\pm 0.7}$	77.4$_{\pm 0.6}$	77.1$_{\pm 0.4}$	75.1$_{\pm 1.1}$
Hebrew	62.3$_{\pm 0.8}$	62.0$_{\pm 1.0}$	64.1$_{\pm 0.5}$	64.2$_{\pm 0.8}$	63.1$_{\pm 0.8}$	**64.3**$_{\pm 0.6}$	63.2$_{\pm 0.5}$	61.1$_{\pm 1.3}$
Hindi	60.7$_{\pm 2.8}$	59.5$_{\pm 1.9}$	61.9$_{\pm 2.5}$	60.7$_{\pm 2.2}$	61.7$_{\pm 2.1}$	**62.0**$_{\pm 2.1}$	61.8$_{\pm 0.9}$	59.1$_{\pm 1.5}$
Hungarian	79.1$_{\pm 0.2}$	80.3$_{\pm 0.4}$	81.1$_{\pm 0.3}$	**81.5**$_{\pm 0.1}$	81.1$_{\pm 0.4}$	80.9$_{\pm 0.4}$	80.5$_{\pm 0.5}$	79.0$_{\pm 0.6}$
Italian	85.0$_{\pm 0.4}$	85.3$_{\pm 0.2}$	85.0$_{\pm 0.3}$	85.1$_{\pm 0.2}$	85.4$_{\pm 0.2}$	**85.7**$_{\pm 0.2}$	85.6$_{\pm 0.2}$	84.9$_{\pm 0.2}$
Japanese	47.8$_{\pm 1.9}$	47.3$_{\pm 1.8}$	**49.5**$_{\pm 2.1}$	49.5$_{\pm 1.8}$	48.3$_{\pm 1.8}$	48.4$_{\pm 1.6}$	47.6$_{\pm 1.0}$	46.6$_{\pm 1.8}$
Korean	55.4$_{\pm 2.4}$	59.9$_{\pm 1.0}$	**63.0**$_{\pm 0.8}$	62.0$_{\pm 1.4}$	60.5$_{\pm 2.1}$	60.3$_{\pm 2.2}$	59.6$_{\pm 2.6}$	55.1$_{\pm 1.5}$
Latvian	69.5$_{\pm 1.8}$	73.1$_{\pm 0.7}$	**74.5**$_{\pm 0.7}$	74.2$_{\pm 0.4}$	73.5$_{\pm 0.5}$	73.4$_{\pm 0.4}$	72.9$_{\pm 0.6}$	68.7$_{\pm 1.4}$
Lithuanian	71.6$_{\pm 1.6}$	73.3$_{\pm 0.5}$	**74.5**$_{\pm 0.7}$	74.5$_{\pm 0.5}$	74.4$_{\pm 0.5}$	74.5$_{\pm 0.6}$	73.7$_{\pm 0.7}$	71.1$_{\pm 1.0}$
Norwegian	88.7$_{\pm 0.4}$	88.8$_{\pm 0.1}$	**89.6**$_{\pm 0.3}$	89.2$_{\pm 0.3}$	88.9$_{\pm 0.4}$	88.9$_{\pm 0.4}$	88.6$_{\pm 0.3}$	88.3$_{\pm 0.3}$
Persian	72.6$_{\pm 0.7}$	72.2$_{\pm 0.6}$	72.7$_{\pm 0.1}$	73.3$_{\pm 0.2}$	73.3$_{\pm 0.4}$	73.8$_{\pm 0.3}$	**74.0**$_{\pm 0.5}$	71.8$_{\pm 0.9}$
Polish	79.7$_{\pm 0.3}$	80.8$_{\pm 0.3}$	82.1$_{\pm 0.2}$	82.2$_{\pm 0.2}$	82.6$_{\pm 0.3}$	**82.7**$_{\pm 0.4}$	81.8$_{\pm 0.3}$	79.7$_{\pm 0.4}$
Portuguese	83.0$_{\pm 0.2}$	83.1$_{\pm 0.3}$	83.2$_{\pm 0.3}$	83.3$_{\pm 0.2}$	**83.7**$_{\pm 0.3}$	83.6$_{\pm 0.3}$	83.4$_{\pm 0.2}$	83.0$_{\pm 0.3}$
Romanian	80.0$_{\pm 0.4}$	81.3$_{\pm 0.4}$	81.9$_{\pm 0.3}$	81.8$_{\pm 0.1}$	82.1$_{\pm 0.3}$	**82.2**$_{\pm 0.5}$	81.9$_{\pm 0.4}$	80.1$_{\pm 0.4}$
Russian	81.5$_{\pm 0.5}$	82.3$_{\pm 0.7}$	84.0$_{\pm 0.1}$	83.9$_{\pm 0.3}$	**84.1**$_{\pm 0.3}$	83.8$_{\pm 0.6}$	82.8$_{\pm 0.5}$	81.2$_{\pm 0.7}$
Slovak	78.2$_{\pm 0.7}$	81.4$_{\pm 0.7}$	**84.2**$_{\pm 0.3}$	84.0$_{\pm 0.2}$	83.8$_{\pm 0.6}$	83.6$_{\pm 0.2}$	82.6$_{\pm 0.4}$	78.9$_{\pm 0.7}$
Slovenian	79.6$_{\pm 0.4}$	81.2$_{\pm 0.6}$	82.9$_{\pm 0.2}$	83.4$_{\pm 0.3}$	**83.5**$_{\pm 0.3}$	83.2$_{\pm 0.3}$	82.2$_{\pm 0.3}$	80.1$_{\pm 0.6}$
Spanish	84.4$_{\pm 0.4}$	85.3$_{\pm 0.3}$	85.2$_{\pm 0.4}$	85.2$_{\pm 0.3}$	85.5$_{\pm 0.2}$	**85.8**$_{\pm 0.3}$	85.7$_{\pm 0.4}$	84.8$_{\pm 0.4}$
Swedish	89.2$_{\pm 0.3}$	89.1$_{\pm 0.4}$	**89.7**$_{\pm 0.2}$	89.5$_{\pm 0.3}$	89.2$_{\pm 0.2}$	89.4$_{\pm 0.2}$	89.4$_{\pm 0.2}$	88.5$_{\pm 0.4}$
Tamil	51.9$_{\pm 0.9}$	52.8$_{\pm 0.6}$	**54.8**$_{\pm 0.5}$	53.1$_{\pm 0.6}$	53.8$_{\pm 0.7}$	54.1$_{\pm 0.7}$	52.3$_{\pm 0.5}$	50.6$_{\pm 0.8}$
Thai	31.4$_{\pm 5.4}$	41.3$_{\pm 4.1}$	51.4$_{\pm 1.1}$	**51.8**$_{\pm 0.5}$	48.6$_{\pm 0.5}$	47.1$_{\pm 0.9}$	41.9$_{\pm 2.3}$	31.8$_{\pm 4.3}$
Turkish	70.0$_{\pm 0.7}$	70.2$_{\pm 0.4}$	70.4$_{\pm 0.2}$	69.9$_{\pm 0.5}$	69.9$_{\pm 0.6}$	70.8$_{\pm 0.3}$	**70.8**$_{\pm 0.6}$	69.7$_{\pm 0.5}$
Ukrainian	81.4$_{\pm 0.2}$	82.5$_{\pm 0.4}$	83.8$_{\pm 0.2}$	84.3$_{\pm 0.2}$	**84.3**$_{\pm 0.4}$	83.8$_{\pm 0.4}$	82.9$_{\pm 0.3}$	81.5$_{\pm 0.3}$
Vietnamese	57.5$_{\pm 0.7}$	57.9$_{\pm 0.4}$	56.7$_{\pm 0.8}$	56.6$_{\pm 1.0}$	57.1$_{\pm 0.6}$	**58.6**$_{\pm 0.8}$	58.5$_{\pm 0.4}$	57.3$_{\pm 0.3}$
Average	73.8$_{\pm 0.6}$	75.0$_{\pm 0.3}$	**76.5**$_{\pm 0.2}$	**76.5**$_{\pm 0.2}$	76.3$_{\pm 0.3}$	76.4$_{\pm 0.3}$	75.6$_{\pm 0.2}$	73.6$_{\pm 0.3}$

Table 23: PoS tagging average accuracy results across 5 seeds using distilMBERT when performing realignment while freezing all layers but one (Aligner: bilingual dictionary)

	FT Only	Layer 0	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
Afrikaans	$85.5_{\pm 0.2}$	$85.7_{\pm 0.2}$	$85.8_{\pm 0.2}$	$\mathbf{85.9}_{\pm 0.2}$	$85.8_{\pm 0.3}$	$85.7_{\pm 0.2}$	$85.4_{\pm 0.2}$	$85.7_{\pm 0.2}$
Arabic	$51.7_{\pm 1.5}$	$\mathbf{66.7}_{\pm 0.4}$	$66.3_{\pm 0.3}$	$66.6_{\pm 0.2}$	$66.3_{\pm 0.2}$	$65.9_{\pm 0.4}$	$65.9_{\pm 0.5}$	$66.6_{\pm 0.4}$
Bulgarian	$85.0_{\pm 0.4}$	$87.6_{\pm 0.3}$	$87.5_{\pm 0.2}$	$87.6_{\pm 0.3}$	$87.6_{\pm 0.2}$	$\mathbf{87.7}_{\pm 0.2}$	$87.5_{\pm 0.3}$	$87.6_{\pm 0.3}$
Catalan	$86.6_{\pm 0.4}$	$88.4_{\pm 0.1}$	$88.4_{\pm 0.1}$	$88.3_{\pm 0.1}$	$\mathbf{88.4}_{\pm 0.2}$	$88.2_{\pm 0.1}$	$88.1_{\pm 0.2}$	$88.4_{\pm 0.1}$
Chinese	$64.3_{\pm 1.2}$	$67.4_{\pm 0.7}$	$67.3_{\pm 0.5}$	$67.4_{\pm 0.7}$	$67.2_{\pm 0.8}$	$67.2_{\pm 0.6}$	$66.7_{\pm 0.6}$	$\mathbf{67.5}_{\pm 0.7}$
Czech	$79.1_{\pm 0.6}$	$85.3_{\pm 0.4}$	$85.1_{\pm 0.4}$	$\mathbf{85.4}_{\pm 0.4}$	$\mathbf{85.4}_{\pm 0.4}$	$85.4_{\pm 0.3}$	$85.2_{\pm 0.3}$	$\mathbf{85.4}_{\pm 0.4}$
Danish	$87.8_{\pm 0.3}$	$88.3_{\pm 0.2}$	$88.3_{\pm 0.2}$	$\mathbf{88.4}_{\pm 0.2}$	$\mathbf{88.4}_{\pm 0.2}$	$88.1_{\pm 0.2}$	$88.2_{\pm 0.2}$	$88.3_{\pm 0.2}$
English	$\mathbf{96.0}_{\pm 0.1}$	$\mathbf{96.0}_{\pm 0.1}$	$\mathbf{96.0}_{\pm 0.1}$	$96.0_{\pm 0.0}$	$96.0_{\pm 0.0}$	$95.9_{\pm 0.0}$	$95.9_{\pm 0.1}$	$\mathbf{96.0}_{\pm 0.1}$
Finnish	$82.3_{\pm 0.7}$	$84.3_{\pm 0.2}$	$\mathbf{84.6}_{\pm 0.2}$	$84.4_{\pm 0.2}$	$84.4_{\pm 0.2}$	$84.2_{\pm 0.2}$	$84.3_{\pm 0.2}$	$84.2_{\pm 0.2}$
French	$85.4_{\pm 0.2}$	$\mathbf{86.6}_{\pm 0.2}$	$86.6_{\pm 0.1}$	$\mathbf{86.6}_{\pm 0.2}$	$\mathbf{86.6}_{\pm 0.2}$	$86.5_{\pm 0.2}$	$86.3_{\pm 0.1}$	$\mathbf{86.6}_{\pm 0.2}$
German	$87.4_{\pm 0.3}$	$88.9_{\pm 0.1}$	$88.9_{\pm 0.1}$	$88.9_{\pm 0.1}$	$89.0_{\pm 0.1}$	$88.9_{\pm 0.1}$	$88.9_{\pm 0.1}$	$\mathbf{89.1}_{\pm 0.1}$
Greek	$74.9_{\pm 1.1}$	$80.3_{\pm 0.4}$	$79.8_{\pm 0.3}$	$79.9_{\pm 0.4}$	$80.0_{\pm 0.2}$	$80.0_{\pm 0.1}$	$\mathbf{80.8}_{\pm 0.8}$	$80.2_{\pm 0.4}$
Hebrew	$62.3_{\pm 0.8}$	$65.0_{\pm 0.5}$	$64.9_{\pm 0.6}$	$65.0_{\pm 0.6}$	$65.2_{\pm 0.4}$	$64.5_{\pm 0.5}$	$\mathbf{65.6}_{\pm 0.6}$	$65.1_{\pm 0.4}$
Hindi	$60.7_{\pm 2.8}$	$66.1_{\pm 2.7}$	$65.2_{\pm 2.6}$	$66.0_{\pm 2.7}$	$66.0_{\pm 2.4}$	$65.1_{\pm 2.4}$	$\mathbf{67.4}_{\pm 3.1}$	$66.3_{\pm 2.6}$
Hungarian	$79.1_{\pm 0.2}$	$82.0_{\pm 0.4}$	$82.0_{\pm 0.3}$	$81.9_{\pm 0.3}$	$\mathbf{82.1}_{\pm 0.2}$	$81.9_{\pm 0.3}$	$81.9_{\pm 0.3}$	$82.0_{\pm 0.4}$
Italian	$85.0_{\pm 0.4}$	$85.9_{\pm 0.1}$	$85.9_{\pm 0.1}$	$85.9_{\pm 0.1}$	$\mathbf{85.9}_{\pm 0.2}$	$85.7_{\pm 0.0}$	$85.6_{\pm 0.2}$	$85.9_{\pm 0.1}$
Japanese	$47.8_{\pm 1.9}$	$52.7_{\pm 1.6}$	$52.2_{\pm 1.4}$	$52.1_{\pm 1.6}$	$52.4_{\pm 1.3}$	$51.5_{\pm 1.2}$	$\mathbf{53.5}_{\pm 2.0}$	$53.1_{\pm 1.6}$
Korean	$55.4_{\pm 2.4}$	$61.8_{\pm 0.9}$	$61.6_{\pm 0.8}$	$62.5_{\pm 0.4}$	$\mathbf{62.9}_{\pm 0.4}$	$62.4_{\pm 0.7}$	$62.4_{\pm 0.6}$	$62.2_{\pm 0.7}$
Latvian	$69.5_{\pm 1.8}$	$76.4_{\pm 0.2}$	$75.7_{\pm 0.2}$	$76.3_{\pm 0.2}$	$\mathbf{76.4}_{\pm 0.3}$	$76.2_{\pm 0.2}$	$76.3_{\pm 0.3}$	$76.4_{\pm 0.2}$
Lithuanian	$71.6_{\pm 1.6}$	$76.2_{\pm 0.3}$	$75.9_{\pm 0.3}$	$76.3_{\pm 0.3}$	$76.2_{\pm 0.3}$	$76.2_{\pm 0.3}$	$\mathbf{76.4}_{\pm 0.5}$	$76.3_{\pm 0.2}$
Norwegian	$88.7_{\pm 0.4}$	$90.1_{\pm 0.2}$	$89.9_{\pm 0.2}$	$90.0_{\pm 0.2}$	$90.1_{\pm 0.3}$	$90.1_{\pm 0.3}$	$90.1_{\pm 0.1}$	$\mathbf{90.2}_{\pm 0.2}$
Persian	$\mathbf{72.6}_{\pm 0.7}$	$72.1_{\pm 0.3}$	$72.6_{\pm 0.3}$	$72.2_{\pm 0.4}$	$72.4_{\pm 0.5}$	$72.1_{\pm 0.4}$	$72.1_{\pm 0.6}$	$72.1_{\pm 0.4}$
Polish	$79.7_{\pm 0.3}$	$83.6_{\pm 0.3}$	$83.5_{\pm 0.2}$	$\mathbf{83.7}_{\pm 0.2}$	$83.5_{\pm 0.3}$	$83.4_{\pm 0.2}$	$83.3_{\pm 0.3}$	$83.5_{\pm 0.2}$
Portuguese	$83.0_{\pm 0.2}$	$\mathbf{84.0}_{\pm 0.1}$	$\mathbf{84.0}_{\pm 0.1}$	$83.9_{\pm 0.1}$	$84.0_{\pm 0.0}$	$83.8_{\pm 0.1}$	$83.8_{\pm 0.2}$	$83.9_{\pm 0.1}$
Romanian	$80.0_{\pm 0.4}$	$83.4_{\pm 0.4}$	$83.2_{\pm 0.4}$	$83.4_{\pm 0.4}$	$83.4_{\pm 0.4}$	$83.4_{\pm 0.3}$	$83.2_{\pm 0.3}$	$\mathbf{83.5}_{\pm 0.4}$
Russian	$81.5_{\pm 0.5}$	$84.8_{\pm 0.4}$	$84.7_{\pm 0.4}$	$84.8_{\pm 0.3}$	$\mathbf{84.9}_{\pm 0.3}$	$84.7_{\pm 0.3}$	$84.8_{\pm 0.4}$	$84.8_{\pm 0.4}$
Slovak	$78.2_{\pm 0.7}$	$85.1_{\pm 0.5}$	$84.9_{\pm 0.5}$	$85.1_{\pm 0.4}$	$\mathbf{85.4}_{\pm 0.4}$	$85.1_{\pm 0.4}$	$84.7_{\pm 0.3}$	$85.1_{\pm 0.5}$
Slovenian	$79.6_{\pm 0.4}$	$83.9_{\pm 0.3}$	$83.9_{\pm 0.2}$	$\mathbf{84.0}_{\pm 0.2}$	$83.9_{\pm 0.2}$	$83.8_{\pm 0.2}$	$83.5_{\pm 0.3}$	$83.9_{\pm 0.3}$
Spanish	$84.4_{\pm 0.4}$	$85.7_{\pm 0.1}$	$85.7_{\pm 0.2}$	$85.7_{\pm 0.2}$	$\mathbf{85.7}_{\pm 0.3}$	$85.5_{\pm 0.2}$	$85.5_{\pm 0.2}$	$85.7_{\pm 0.2}$
Swedish	$89.2_{\pm 0.3}$	$\mathbf{90.1}_{\pm 0.3}$	$89.9_{\pm 0.2}$	$90.0_{\pm 0.2}$	$\mathbf{90.1}_{\pm 0.3}$	$90.0_{\pm 0.2}$	$90.0_{\pm 0.2}$	$90.1_{\pm 0.2}$
Tamil	$51.9_{\pm 0.9}$	$55.7_{\pm 0.8}$	$53.8_{\pm 0.6}$	$55.9_{\pm 0.5}$	$\mathbf{56.1}_{\pm 0.7}$	$54.6_{\pm 1.0}$	$55.4_{\pm 1.0}$	$55.5_{\pm 0.7}$
Thai	$31.4_{\pm 5.4}$	$54.9_{\pm 0.6}$	$54.1_{\pm 1.1}$	$54.8_{\pm 0.9}$	$54.8_{\pm 0.7}$	$55.1_{\pm 0.7}$	$\mathbf{55.2}_{\pm 0.8}$	$55.1_{\pm 0.6}$
Turkish	$70.0_{\pm 0.7}$	$70.5_{\pm 0.3}$	$70.4_{\pm 0.3}$	$70.7_{\pm 0.2}$	$\mathbf{70.8}_{\pm 0.4}$	$70.2_{\pm 0.4}$	$70.5_{\pm 0.3}$	$70.4_{\pm 0.3}$
Ukrainian	$81.4_{\pm 0.2}$	$85.0_{\pm 0.2}$	$85.0_{\pm 0.2}$	$85.0_{\pm 0.2}$	$85.0_{\pm 0.1}$	$\mathbf{85.1}_{\pm 0.1}$	$85.0_{\pm 0.2}$	$85.0_{\pm 0.2}$
Vietnamese	$57.5_{\pm 0.7}$	$57.5_{\pm 0.4}$	$57.8_{\pm 0.2}$	$57.7_{\pm 0.4}$	$\mathbf{57.8}_{\pm 0.4}$	$57.1_{\pm 0.5}$	$57.4_{\pm 0.4}$	$57.5_{\pm 0.3}$
Average	$73.8_{\pm 0.6}$	$77.7_{\pm 0.2}$	$77.5_{\pm 0.2}$	$77.7_{\pm 0.1}$	$\mathbf{77.8}_{\pm 0.2}$	$77.5_{\pm 0.1}$	$77.7_{\pm 0.2}$	$77.7_{\pm 0.2}$

Table 24: PoS tagging average accuracy results across 5 seeds using distilMBERT when performing realignment while freezing a single layer (Aligner: bilingual dictionary)