On the Calibration of Massively Multilingual Language Models

Kabir Ahuja[1]   **Sunayana Sitaram[1]**   **Sandipan Dandapat[2]**   **Monojit Choudhury[2]**

[1] Microsoft Research, India
[2] Microsoft R&D, India

{t-kabirahuja,sadandap,sunayana.sitaram,monojitc}@microsoft.com

Abstract

Massively Multilingual Language Models (MMLMs) have recently gained popularity due to their surprising effectiveness in cross-lingual transfer. While there has been much work in evaluating these models for their performance on a variety of tasks and languages, little attention has been paid on how well *calibrated* these models are with respect to the confidence in their predictions. We first investigate the calibration of MMLMs in the zero-shot setting and observe a clear case of miscalibration in low-resource languages or those which are typologically diverse from English. Next, we empirically show that calibration methods like temperature scaling and label smoothing do reasonably well in improving calibration in the zero-shot scenario. We also find that few-shot examples in the language can further help reduce calibration errors, often substantially. Overall, our work contributes towards building more reliable multilingual models by highlighting the issue of their miscalibration, understanding what language and model-specific factors influence it, and pointing out the strategies to improve the same.

1 Introduction

Massively Multilingual Language Models (MMLMs) like mBERT (Devlin et al., 2019), XLMR (Conneau et al., 2020), mT5 (Xue et al., 2021) and mBART (Liu et al., 2020) have been surprisingly effective at zero-shot cross-lingual transfer i.e. when fine-tuned on an NLP task in one language, they often tend to generalize reasonably well in languages unseen during fine-tuning.

These models have been evaluated for their performance across a range of multilingual tasks (Pan et al., 2017; Nivre et al., 2018; Conneau et al., 2018) and numerous methods like adapters (Pfeiffer et al., 2020), sparse fine-tuning (Ansell et al., 2022) and few-shot learning (Lauscher et al., 2020) have been proposed to further improve performance of cross-lingual transfer.

Despite these developments, there has been little to no attention paid to the calibration of these models across languages i.e. how reliable the confidence predictions of these models are. As these models find their way more and more into the real word applications with safety implications, like Hate Speech Detection (Davidson et al., 2017; Deshpande et al., 2022) it becomes important to only take extreme actions for high confidence predictions by the model (Sarkar and KhudaBukhsh, 2021). Hence, calibrated confidences are desirable to have when deploying such systems in practice.

Guo et al. (2017) showed that modern neural networks used for Image Recognition (He et al., 2016) though perform much better than the ones introduced decades ago (Lecun et al., 1998), but are significantly worse calibrated and often over-estimate their confidence on incorrect predictions. For NLP tasks specifically, Desai and Durrett (2020) showed that classifiers trained using pre-trained transformer-based models (Devlin et al., 2019) are well calibrated both in-domain and out-of-domain settings compared to non-pre-trained model baselines (Chen et al., 2017). Notably, Ponti et al. (2021) highlights, since zero-shot cross-lingual transfer represents shifts in the data distribution the point estimates are likely to be miscalibrated, which forms the core setting of this work.

In light of this, our work has three main contributions. *First*, we investigate the calibration of two commonly used MMLMs: mBERT and XLM-R on four NLU tasks under zero-shot setting where the models are fine-tuned in English and calibration errors are computed on unseen languages. We find a clear increase in calibration errors compared to English as can be seen in Figures 1a and 1b, with calibration being significantly worse for Swahili compared to English.

Second, we look for factors that might affect the zero-shot calibration of MMLMs and find in most cases that the calibration error is strongly correlated
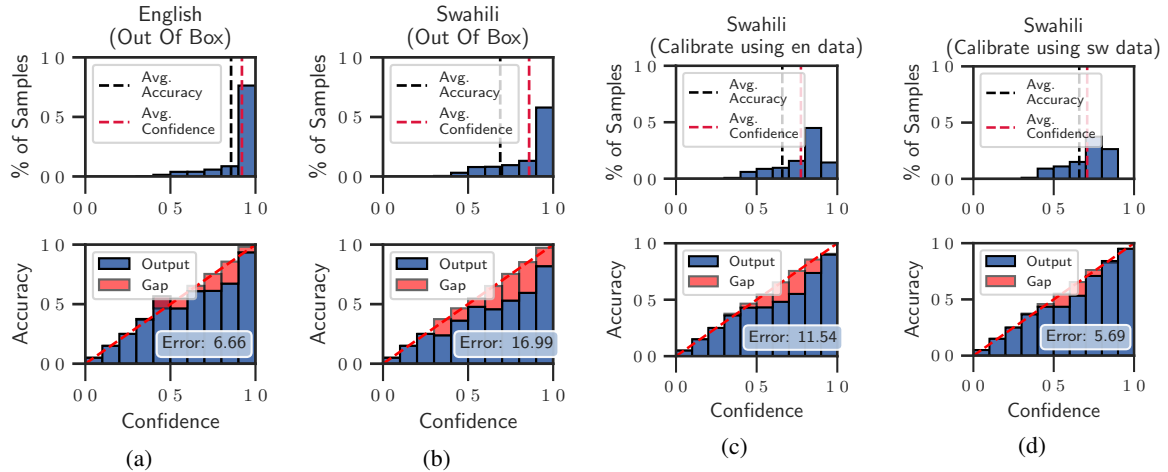
Figure 1: Reliability Diagrams for XLMR fine-tuned on XNLI using training data in English. 1a and 1b provides the calibration of XLMR fine-tuned without any calibration technique on English and Swahili. 1c and 1d are obtained by calibrating the model using TS + LS and Self-TS + LS techniques respectively as described in Section 2.1

with pre-training data size, syntactic similarity, and sub-word overlap between the unseen language and English. This reveals that MMLMs are miscalibrated in the zero-shot setting for low-resource languages and languages that are typologically distant from English.

Finally, we show that model calibration across different languages can be substantially improved by utilizing standard calibration techniques like Temperature Scaling (Guo et al., 2017) and Label Smoothing (Pereyra et al., 2017) without collecting any data in the language (see Figure 1c). Using a few examples in a language (the few-shot setting), we see even more significant drops in the calibration errors as can be seen in Figure 1d.

To the best of our knowledge, ours is the first work to investigate and improve the calibration of MMLMs. We expect this study to be a significant contribution towards building reliable and linguistically fair multilingual models. To encourage future research in the area we make our code publicly available[1].

2 Calibration of Pre-trained MMLMs

Consider a classifier $h : \mathcal{X} \to [K]$ obtained by fine-tuning an MMLM for some task with training data in a pivot language p, where $[K]$ denotes the set of labels $\{1, 2, \cdots, K\}$. We assume h can predict confidence of each of the $[K]$ labels and is given by $h_k(x) \in [0, 1]$ for the kth label. h is said to be

[1] https://github.com/microsoft/MMLMCalibration

calibrated if the following equality holds:

$$p(y = k | h_k(x) = q) = q$$

In other words, for a perfectly calibrated classifier, if the predicted confidence for a label k on an input x is q, then with a probability q the input should actually be labelled k. Naturally, in practical settings the equality does not hold, and neural network based classifiers are often miscalibrated (Guo et al., 2017). One way of quantifying this notion of miscalibration is through the Expected Calibration Error (ECE) which is defined as the difference in expectation between the confidence in classifier's predictions and their accuracies (Refer Appendix A.1 for details). In our experiments we compute ECE on each language l's test data and denote their corresponding calibration errors as ECE(l).

2.1 Calibration Methods

We briefly review some commonly used methods for calibrating neural network based classifiers.

1. Temperature Scaling (TS and Self-TS)(Guo et al., 2017) is applied by scaling the output logits using a temperature parameter T before applying softmax i.e. :

$$h_k(x) = \frac{\exp o_k(x)/T}{\sum_{k' \in K} \exp o_{k'}(x)/T}$$

, where o_k denotes the logits corresponding to the kth class. T is a learnable parameter obtained post-training by maximizing the log-likelihood on the dev set while keeping other network parameters

fixed. We experiment with two settings for improving calibration on a target language: using dev data in English to perform temperature scaling (TS) and using the target language's dev data (Self-TS).

2. Label Smoothing (LS) (Pereyra et al., 2017) is a regularization technique that penalizes low entropy distributions by using *soft* labels that are obtained by assigning a fixed probability $q = 1 - \alpha$ to the true label ($0 < \alpha < 1$), and distributing the remaining probability mass uniformly across the remaining classes. Label smoothing has been empirically shown to be competitive with temperature scaling for calibration (Müller et al., 2019) especially in out of domain settings (Desai and Durrett, 2020)

3. Few-Shot Learning (FSL) We also investigate if fine-tuning the MMLM on a few examples in a target language in addition to the data in English, leads to any improvement in calibration as it does in the performance (Lauscher et al., 2020). Since these models are expected to be calibrated worse for out-of-domain data compared to in-domain data (Desai and Durrett, 2020), we try to improve calibration by reducing the domain shift through few-shot learning.

Apart from these, we also consider combinations of different calibration methods in our experiments, including Label Smoothing with Temperature Scaling (TS + LS or Self-TS + LS) and Few-Shot Learning with Label Smoothing (FSL + LS).

3 Experiments

We seek to answer the following research questions: a) How well calibrated are fine-tuned MMLMs in the zero-shot cross lingual setting? b) What linguistic and model-specific factors influence calibration errors across languages? c) Can we improve the calibration of fine-tuned models across languages?

3.1 Experimental Setup

Datasets We consider 4 multilingual classification datasets to study calibration of MMLMs which include: i) The Cross-Lingual NLI Corpus (XNLI) (Conneau et al., 2018), ii) Multilingual Dataset for Causal Commonsense Reasoning (XCOPA) (Ponti et al., 2020), iii) Multilingual Amazon Reviews Corpus (MARC) (Keung et al., 2020) and, iv) Cross-lingual Adversarial Dataset for Paraphrase Identification (PAWS-X) (Yang et al., 2019). Statistics of these datasets can be found in Table 5.

Training setup We consider two commonly used MMLMs in our experiments i.e. Multilin-

Dataset	MMLM	ECE(en)	$\underset{l \in \mathcal{L}'}{\mathbb{E}}\left[\text{ECE}(l)\right]$	$\underset{l \in \mathcal{L}'}{\max}\text{ECE}(l)$
XNLI	XLM-R	7.32	13.34	19.07 (sw)
	mBERT	5.44	12.34	45.15 (th)
XCOPA	XLM-R	14.54	20.07	29.33 (sw)
	mBERT	23.4	23.51	29.02 (sw)
MARC	XLM-R	7.15	9.65	13.45 (zh)
	mBERT	9.38	11.11	17.33 (ja)
PAWS-X	XLM-R	1.93	4.28	5.88 (ja)
	mBERT	3.57	10.32	15.65 (ko)

Table 1: Calibration Errors across tasks for XLM-R and mBERT. $\mathcal{L}'$ in the fourth column denotes the set of supported languages in a task other than English. The language in parenthesis in the column 5 denotes the language with maximum calibration error.

Dataset	SIZE	SYN	SWO
XNLI	-0.8	**-0.88**	-0.85
XCOPA	**-0.85**	-0.73	-0.62
MARC	-0.41	**-0.46**	-0.27
PAWS-X	-0.48	**-0.93**	-0.92

Table 2: Pearson correlation coefficient of ECE with SIZE, SYN, and SWO features of different languages in the test set for XLMR.

gual BERT (mBERT) (Devlin et al., 2019), and XLM-RoBERTa (XLMR) (Conneau et al., 2020). mBERT is only available in the base variant with 12 layers and for XLMR we use the large variant with 24 layers. We use English training data to fine-tune the two MMLMs on all the tasks and evaluate the accuracies and ECEs on the test data for different languages. For the few-shot case we use the validation data in target languages to do continued fine-tuning (FSL) and temperature scaling (Self-TS). Refer to Section A.3 in the Appendix for more details.

3.2 Results

Out of Box Zero-Shot Calibration (OOB) We first investigate how well calibrated MMLMs are on the languages unseen during fine-tuning without applying any calibration techniques. As can be seen in the Table 1, the average calibration error on languages other than English (column 4) is almost always significantly worse than the errors on English test data (column 3) for both mBERT and XLMR across the 4 tasks. Along with the expected calibration errors across unseen languages we also report the worst case ECE (in column 5), where we see $2\times$ to $5\times$ increase in errors compared to English. The worst case calibration is commonly observed

Dataset	MMLM	Zero-Shot Calibration				Few-Shot Calibration			
		OOB	TS	LS	TS + LS	Self-TS	Self-TS + LS	FSL	FSL + LS
XNLI	XLM-R	13.34	6.74	6.93	**4.89**†	5.41	**4.05**‡	7.67	4.36
	mBERT	12.34	**6.29**†	10.42	6.70	4.77	4.69	3.14	**2.55**‡
XCOPA	XLM-R	20.07	15.95	5.47	**4.52**†	16.02	**4.06**‡	8.94	4.39
	mBERT	23.51	20.02	12.41	**6.77**†	20.09	6.89	3.75	**3.54**‡

Table 3: Calibration Errors ($\mathbb{E}_{l \in \mathcal{L}'}[\mathsf{ECE}(l)]$) for XLM-R and mBERT on using different methods for calibration. We categorize the methods into zero-shot i.e. the methods that do not use any target language data to calibrate and few-shot for the methods that require some examples in target language. Detailed results are in Table 9 of Appendix

for low resource languages like Swahili or the languages that are typologically diverse from English like Japanese. Consequently, the overall calibration is worse on tasks like XCOPA and XNLI compared to PAWS-X and MARC, as the former two have more diverse set of languages while the latter consists of high resource languages only.

Factors Affecting Calibration Next, we analyze which model-specific and typological features might influence the out-of-box calibration across languages. For a given task and MMLM, we compute Pearson correlation coefficients between the calibration errors and three factors studied extensively in zero shot cross lingual transfer literature which are:

i) **SIZE**: Logarithm of the pre-training data size (number of tokens) in a language i.e. how well a language is represented in the pre-training corpus of an MMLM (Lauscher et al., 2020; Wu and Dredze, 2020),

ii) **SYN**: We utilize the syntactic features provided by the URIEL project (Littell et al., 2017) to compute the syntactic similarity between the pivot and target language as done in Lin et al. (2019).

iii) **SWO**: Finally we consider the sub-word overlap between the pivot and target language as defined in Srinivasan et al. (2021). To compute SWO, first vocabularies V_p and V_t are identified for the pivot and target langauge respectively by tokenizing the wikipedias in the two languages and getting rid of the tokens that appear less than 10 times in the corpora. The subword overlap is then computed as :

$$\mathrm{SWO} = \frac{|V_p \cap V_t|}{|V_p \cup V_t|}$$

.

In the case of XLMR, for all the tasks except MARC, we observe strong negative correlations between ECE and the three factors mentioned above (Table 2), meaning that lower the amount of pre-

training data present for a language or its relatedness with English, the higher is the calibration error. Out of the three, the correlations are more consistently (negatively) high with SYN. We observe similar correlations albeit to a slightly lower extent for mBERT as well (Table 6 in Appendix).

Improving Calibration Now that we have identified miscalibration as an issue in MMLMs and factors influencing the same, we seek to improve their calibration across languages. We utilize the calibration methods described in Section 2.1, and report the average calibration errors across the unseen languages in Table 3 for XNLI and XCOPA datasets [2]. Both zero-shot calibration methods (TS and LS) that only make use of English data to calibrate can be seen to obtain substantially lower calibration errors compared to out of box calibration (OOB) across all the tasks and MMLMs. Out of the two, temperature scaling often results in bigger drops in ECE compared to label smoothing with an exception of XCOPA dataset where label smoothing performs much better. In majority of the cases the errors can be reduced even further by considering the combination of the two techniques i.e. TS + LS. Temperature scaling by design does not affect the accuracy of uncalibrated models. In all our experiments, we observe that models trained with label smoothing also obtain accuracies very close to their uncalibrated counterparts. Refer to Appendix Table 8 for the exact accuracy numbers.

Next, we investigate if it is possible to reduce the calibration errors on a language even further if we are allowed to collect a few-samples in that language. We observe that when combined with label smoothing, both using few-shot data to do temperature scaling (Self-TS + LS) or fine-tuning (FSL + LS), often results in significant drops over the errors corresponding to the best performing

---
[2]Refer to Table 7 in Appendix for results on MARC and PAWS-X

zero-shot calibration method. We do not use more than 2500 examples in the target language in any of the experiments, and the number of examples can be as low as 100 for XCOPA. For XNLI and XCOPA datasets, we observe that Self-TS + LS performs better than FSL+LS for XLMR and the reverse is true for mBERT. One advantage of using FSL over TS is that it can often result in increase in accuracy in addition to the reducing the calibration errors. However, we do observe an exception to this phenomenon for XCOPA where fine-tuning on the 100 validation examples hurts the overall test accuracy of the models. Hence, our general recommendation is to use FSL + LS for calibrating the models if the amount of data that can be collected is not too low, otherwise it might be more appropriate to use Self-TS + LS as learning just one parameter (T) should be less prone to overfitting compared to the weights of the entire network.

4 Conclusion

In this work we showed that MMLMs like mBERT and XLMR are miscalibrated in a zero-shot cross lingual setting, with the calibration errors being even worse on low resource languages and languages that are typologically distant from the pivot language (often English). We then demonstrated the effectiveness of standard calibration techniques for improving calibration across languages both with and without collecting any new language-specific labelled data. We recommend that researchers and practitioners consider, measure and report the calibration of multilingual models while using them for scientific studies and building systems. In future work, we aim to bridge the gap between zero-shot and few-shot calibration methods by exploring unsupervised calibration methods under domain shift (Pampari and Ermon, 2020; Park et al., 2020) that utilizes unlabelled data in new domains to improve calibration. Investigating the cross lingual calibration of MMLMs for tasks other than sentence classification like Sequence Labelling (Pan et al., 2017; Nivre et al., 2018) and Question Answering (Artetxe et al., 2020) is another natural extension of our work.

Limitations

Our work focused on measuring and improving calibration of MMLMs across different languages and tasks. The languages that we considered in our experiments were the ones for which labelled test sets

were available in the 4 multilingual benchmarks that we considered. The number of languages in these benchmarks ranged from 6 in case of MARC to 15 in XNLI, covering mostly high resource languages [3] with Swahili being the lowest resource language studied. However, the MMLMs considered in this work supports around 100 languages many of which are arguably even lower resource compared to Swahili. Hence, how well the methods discussed in the paper work towards improving the calibration for such languages needs to be explored but is limited by the current state of multilingual benchmarking (Ahuja et al., 2022).

Additionally, investigating the state of calibration across the languages for the 4 tasks and 2 MMLMs for different hyper-parameters and random seeds required a reasonably large amount of GPU resources (we used NVIDIA V100 and P100 GPUs). However, the calibration methods that we describe in the paper can work with little (for temperature scaling and few-shot learning) to no (for label smoothing) additional compute over the standard model training.

Ethics Statement

Our work deals with calibration of confidence predictions of classifiers trained on top of pre-trained multilingual models. Having well calibrated predictions is imperative for building robust NLP systems especially when using them for security sensitive applications like Hate Speech Detection to flag social media accounts (Cuthbertson, 2021), decision making in law enforcement and fraud detection (Metz and Satariano, 2020), where extreme actions should only be taken when we are confident about the system's prediction. However, the predicted confidences mean essentially nothing if the model is miscalibrated, posing major risks in using such models. Through our work we highlight that the commonly used multilingual models are highly miscalibrated when used in a zero-shot setting for low resource and typologically diverse languages. Additionally, we manage to substantially improve the calibration of these models across languages, addressing this linguistic disparity and boosting the reliability of such models.

---

[3]class 3 or above according to the hierarchy defined by Joshi et al. (2020)

Acknowledgements

References

Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022. Beyond static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 64–74, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of ACL 2020*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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 EMNLP 2018*, pages 2475–2485.

Anthony Cuthbertson. 2021. Ai mistakes 'black and white' chess chat for racism. *The Independent*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Neha Deshpande, Nicholas Farris, and Vidhur Kumar. 2022. Highly generalizable models for multilingual hate speech detection.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. *CoRR*, abs/2010.02573.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages

3125–3135, Florence, Italy. 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*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Cade Metz and Adam Satariano. 2020. An algorithm that grants freedom, or takes it away. *The New York Times*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. 2018. Universal dependencies 2.2.

Anusri Pampari and Stefano Ermon. 2020. Unsupervised calibration under covariate shift. *CoRR*, abs/2006.16405.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL 2017*, pages 1946–1958.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. 2020. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3219–3229. PMLR.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions.

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.

Edoardo M. Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the Association for Computational Linguistics*, 9:410–428.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Rupak Sarkar and Ashiqur R. KhudaBukhsh. 2021. Are chess discussions racist? an adversarial hate speech data set (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):15881–15882.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings*

of the 5th Workshop on Representation Learning for NLP, pages 120–130, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 Measuring Calibration

Reliability Diagrams are an effective way of visualizing the calibration of a classifier. To plot these, we first predict the (maximum) confidence for each example in the test set and group the data points into M equally sized bins based on the predicted confidence. For each bin, accuracy is computed (by considering label with maximum confidence as prediction) and plotted on the y axis with confidence being on the x-axis, as can be seen in Figure 1 (blue bars). For a perfectly calibrated classifier the confidence of a bin should match with the classifier's accuracy on that bin i.e. the accuracy vs confidence curve should lie on the line $y = x$ (red-dotted line in Figure 1). The gap between the two is plotted (as red bars in Figure 1) to represent the calibration error of the classifier.

Expected Calibration Error (ECE) is defined as expected value of the difference between the confidence and accuracy of the classifier's predictions. In practice ECE of a classifier is computed using a binning strategy as described in Guo et al. (2017) where the confidence predictions (corresponding to the maximum confidence class) on the n test examples are grouped into M uniform-sized bins, such that set of examples belonging to the mth bin are denoted by B_m. Accuracy ($\mathrm{acc}(B_m)$) and average confidence ($\mathrm{conf}(B_m)$) for each bin is computed and a weighted average of the differences between the two is taken to obtain ECE.

$$\mathsf{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$$

A.2 Datasets Description

1. XNLI[4] (Conneau et al., 2018) is a Natural Langauge Inference task where a premise and hypothesis are given and the task is to predict if the hypothesis is *entailed* in premise, *contradicts* the premise, or is *neutral* towards it, hence being a three way classification problem. MultiNLI (Williams et al., 2018) corpus which is available in English is used as training set, and dev and validation sets are obtained by manually translating crowd sourced English sentences for the task into 14 other languages (Arabic, Bulgarian, German, Greek, Spanish, French, Hindi, Russian, Swahili, Thai, Turkish, Urdu, Vietnamese and Chinese.)

2. XCOPA[5] (Ponti et al., 2020) is a multilingual benchmark for causal commonsense reasoning. It was obtained by extending the dev and test sets of Choice of Plausible Alternatives (COPA)[6] (Roemmele et al., 2011) dataset to 11 typologically diverse languages. The task here is, given a premise and two alternatives, predict which alternative has a causal relationship with the premise. The original COPA dataset has only 400 training examples in English, hence it is common to first train the model on Social-IQA (SIQA)[7] (Sap et al., 2019) dataset (which has around 33k training examples), and then fine-tune it on COPA, and that's the strategy we adopt in our experiments. SIQA is similar to COPA but the questions or premise are defined in such a way that the possible answers have social implications, and instead of two alternatives it has three. Out of the 11 supported languages we experiment with 9 languages, ignoring Quechua and Haitian Creole as both mBERT and XLM-R were not pre-trained on these languages, leaving us with Greek, Indonesian, Italian, Swahili, Tamil, Thai, Turkish, Vietnamese and Chinese.

3. MARC[8] (Keung et al., 2020) is the multilingual Amazon product reviews corpus. We are given title, body and category of the review and the task is to predict the corresponding rating from 1 to 5. The corpus contains train, dev and test sets in six high resource languages: English, German, Spanish, French, Japanese and Chinese. In our experiments we only fine-tune using title and body text and ignore the category information.

4. PAWS-X[9] (Yang et al., 2019) is an adversarial dataset for multilingual paraphrase detection. It was adapted from PAWS dataset (Zhang et al., 2019) by manually translating the dev and test sets in English to six high resource languages (French, German, Spanish, Chinese, Japanese and Korean). The task is, given a pair of sentences predict whether the two are paraphrases of each other. For training the original English PAWS dataset is used.

The statistics of all these datasets are provided in Table 5

---

[4] https://github.com/facebookresearch/XNLI

[5] https://huggingface.co/datasets/xcopa
[6] https://huggingface.co/datasets/super_glue
[7] https://huggingface.co/datasets/social_i_qa
[8] https://registry.opendata.aws/amazon-reviews-ml/
[9] https://huggingface.co/datasets/paws-x

A.3 Detailed Experimental Setup

For fine-tuning the model we do a grid search over the learning rate([1e-5, 3e-5, 5e-6, 7e-6]) and the number of epochs ([1, 3, 4, 5]), run each setting for 3 random seeds (1, 11 and 22) and select the best hyper-parameter set corresponding to the average dev accuracy on English data [10]. The final set of hyperparameters used for each dataset and MMLM are provided in Table 4. Apart from these we use a batch size of 8 in all experiments. We use Adam optimizer (Kingma and Ba, 2015) to train all of our models and LBFGS to learn the temperature parameter while performing temperature scaling.

For computing ECE values for different tasks and languages we use $M = 10$ i.e. 10 buckets. For label smoothing we set the smoothing parameter $\alpha = 0.1$ in all experiments and initialize the temperature $T = 1.5$ while doing temperature scaling. For few-shot cases we use $\min(2500, |\mathcal{D}_{val}(\mathsf{t}, l)|)$ examples, where $|\mathcal{D}_{val}(\mathsf{t}, l)|$ denotes the number of dev examples available in task t for language l, and use that to perform continued fine-tuning or temperature scaling.

All the experiments were run on NVIDIA V100 and P100 GPUs with 32GB and 16GB memory respectively. We use pre-trained models available in Hugging Face's Transformers library (Wolf et al., 2020). For computing the calibration errors as well as plotting the reliability diagrams we use the open source tool Calibration Framework[11]. To encourage the research in this area we will make our code public.

---

[10]For XCOPA we did use dev data in all languages for model selection as we saw performance in English data to not necessarily correlate well with performance on other languages.

[11]https://github.com/fabiankueppers/calibration-framework

Dataset	MMLM	Learning Rate	Epochs	Few-Shot Learning Rate	Few-Shot Epochs
XNLI	XLM-R	7e-6	3	5e-06	2
	mBERT	3e-5	3	3e-5	1
XCOPA	XLM-R	5e-6 (for SIQA) & 5e-6 (for COPA)	4 (for S-IQA) & 10 (for COPA)	1e-5	1
	mBERT	1e-5 (for S-IQA) & 3e-5 (for COPA)	3 (for SIQA) & 10 (for COPA)	3e-5	10
MARC	XLM-R	5e-6	3	5e-6	1
	mBERT	5e-6	3	5e-6	1
PAWS-X	XLM-R	7e-6	3	3e-5	1
	mBERT	1e-5	3	3e-5	1

Table 4: Final list of hyperparameters used for reporting results.

Dataset	Number of Languages	Number of Labels	Training Size	Dev Size	Test Size
XNLI (sub-sampled)	15	3	40000	2500	2500
XCOPA	9	2	33410 (S-IQA) + 500 (COPA)	100	400
MARC (sub-sampled)	6	5	40000	2500	5000
PAWS-X	7	2	50000	2000	2000

Table 5: Dataset statistics for the 4 multilingual classification tasks that we study in our experiments. We use sub-sampled versions of XNLI and MARC and only use the first 40k examples in their training sets to reduce the compute overhead and making the scale of training data consistent across all the tasks. For completeness we also run some preliminary experiments with using the entire data for fine-tuning XNLI and present the calibration errors in Figure 2.
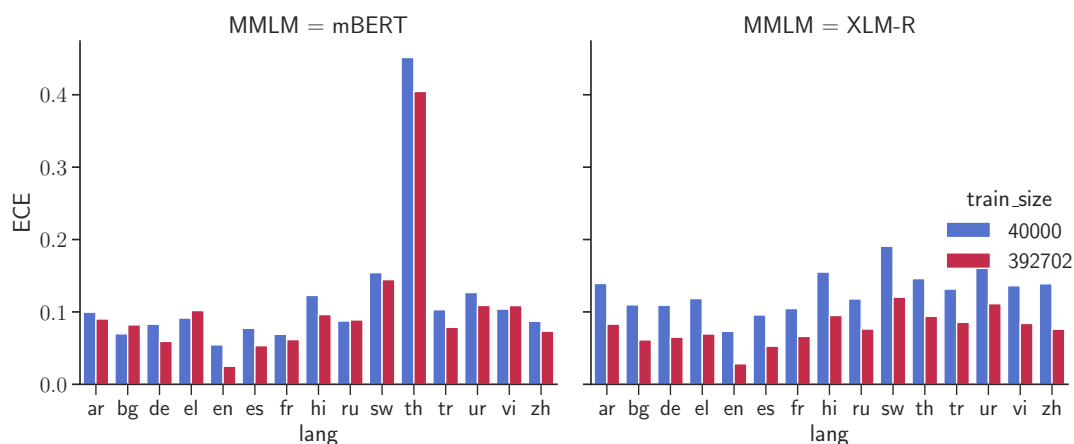

Figure 2: Out-of-Box Calibration Errors (ECE) for XLMR and mBERT trained on XNLI with 40k samples (sub-sampled) and 392k samples(full data). Even though the calibration is better with using the entire data for training, the observed patterns about the models being mis-calibrated on languages other than English, especially low resource languages like Swahili, Thai and Urdu still hold true.

Dataset	SIZE	SYN	SWO
XNLI	-0.4	-0.38	**-0.41**
XNLI (wo th)	-0.86	**-0.88**	-0.78
XCOPA	-0.14	**-0.22**	0.07
MARC	-0.5	-0.69	**-0.86**
PAWS-X	-0.91	-0.91	**-0.98**

Table 6: Pearson correlation coefficient between the Expected Calibration Error (ECE) and SIZE, SYN, and SWO features of different languages in the test set for mBERT

(a)

(b)
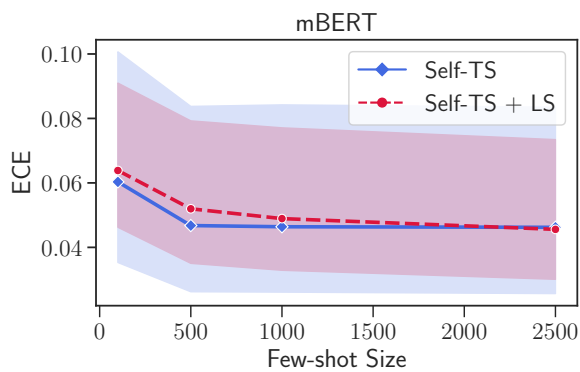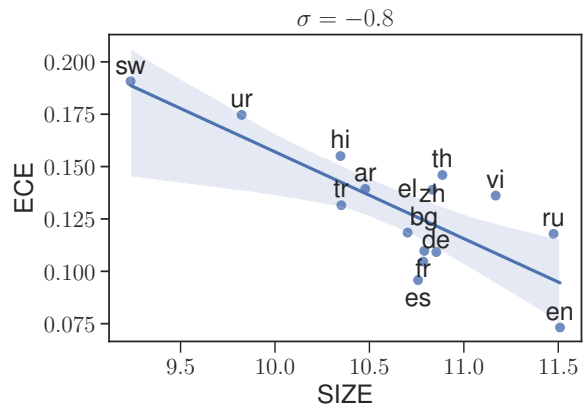
Figure 3: Variation in ECE as we use more and more data for calibrating using Self-TS method across languages for XNLI dataset. As can be seen 500 samples are sufficient to obtain low calibration errors.

(a)

(b)

(c)

Figure 4: Visualizing the correlations of ECE with SIZE, SYN and SWO for XLMR fine-tuned on XNLI.

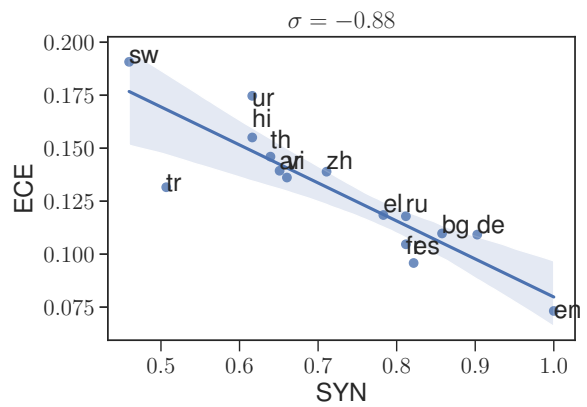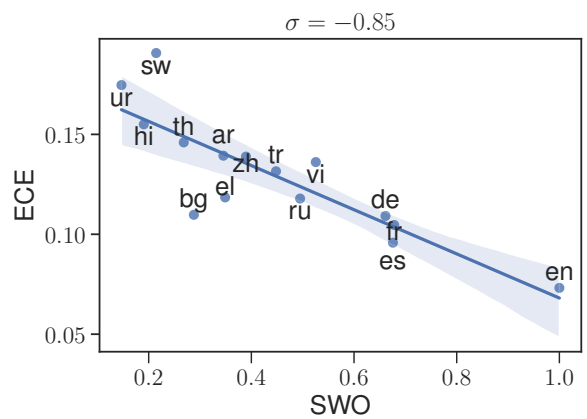| Dataset	MMLM	Zero-Shot Calibration				Few-Shot Calibration			
		OOB	TS	LS	TS + LS	Self-TS	Self-TS + LS	FSL	FSL + LS
MARC	XLM-R	9.65	**4.22**†	7.93	4.45	3.55	3.36	**2.51**‡	2.58
	mBERT	11.11	6.14	5.96	**4.46**†	4.62	4.71	**2.12**‡	3.01
PAWS-X	XLM-R	4.28	**2.29**†	3.37	4.44	-	-	-	-
	mBERT	10.33	5.64	5.58	**5.36**†	-	-	-	-

Table 7: Calibration Errors for MARC and PAWS-X tasks for XLM-R and mBERT on using different methods for calibration. For PAWS-X we only report numbers for Zero-Shot methods as the dev data and test data of the benchmark have sentences in common (even though the pairs are unique), hence we avoid using dev examples as few-shot in this case.

Dataset	MMLM	Zero-Shot Calibration				Few-Shot Calibration			
		OOB	TS	LS	TS + LS	Self-TS	Self-TS + LS	FSL	FSL + LS
XNLI	XLM-R	74.9	74.9	74.6	74.6	74.9	74.6	79.4	79.3
	mBERT	56.7	56.7	57.5	57.5	56.7	57.5	60.6	61.2
XCOPA	XLM-R	74.3	74.3	75.1	75.1	74.3	75.1	67.2	69.5
	mBERT	54.5	54.5	54.4	54.4	54.5	54.4	53.0	52.9
MARC	XLM-R	57.7	57.7	57.6	57.6	57.7	57.6	59.4	59
	mBERT	42.9	42.9	42.6	42.6	42.9	42.6	49.6	49.2
PAWS-X	XLM-R	76.1	76.1	75.6	75.6	-	-	-	-
	mBERT	80.4	80.4	81.1	81.1	-	-	-	-

Table 8: Accuracy ($\mathbb{E}_{l \in \mathcal{L}'}[\texttt{Accuracy}(l)]$) for XLM-R and mBERT on using different methods for calibration. Similar to Table 7, here again we skip few-shot calibration for PAWS-X due to the possible data leakage.

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	7.32	13.34	19.07
TS	2.02	6.74	11.81
LS	3.2	6.93	12.1
TS + LS	4.1	4.9	9.35
	Few-Shot Calibration		
Self-TS	2.02	5.41	9.7
Self-TS + LS	4.1	4.05	4.64
FSL	7.32	7.67	9.23
FSL + LS	3.2	4.37	5.73

(a) Detailed results on XNLI with XLMR

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	5.44	12.34	45.15
TS	2.51	6.29	37.25
LS	4.51	10.42	38.36
TS + LS	2.61	6.71	30.51
	Few-Shot Calibration		
Self-TS	2.51	4.77	29.82
Self-TS + LS	2.61	4.7	23.23
FSL	5.44	3.14	4.28
FSL + LS	2.61	2.55	4.26

(b) Detailed results on XNLI with mBERT

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	14.54	20.01	29.33
TS	12.31	15.95	24.04
LS	9.66	5.47	9.42
TS + LS	8.98	4.52	7.2
	Few-Shot Calibration		
Self-TS	12.31	16.02	24.02
Self-TS + LS	8.98	4.06	5.36
FSL	14.54	8.93	14.92
FSL + LS	9.66	4.4	5.74

(c) Detailed results on XCOPA with XLMR

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	23.4	23.51	29.02
TS	21.76	20.02	23.99
LS	15.85	12.41	15.87
TS + LS	11.28	6.77	8.93
	Few-Shot Calibration		
Self-TS	21.76	20.1	24.12
Self-TS + LS	11.28	6.89	9.01
FSL	23.4	3.75	10.5
FSL + LS	15.85	3.54	6.65

(d) Detailed results on XCOPA with mBERT

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	7.15	9.65	13.45
TS	2.75	4.22	5.8
LS	5.36	7.93	10.66
TS + LS	3.39	4.55	6.33
	Few-Shot Calibration		
Self-TS	2.75	3.56	4.1
Self-TS + LS	3.39	3.36	3.64
FSL	7.15	2.51	3.51
FSL + LS	5.36	2.59	3.28

(e) Detailed results on MARC with XLMR

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	9.38	11.1	17.3
TS	3.56	6.14	10.9
LS	5.19	5.96	12.9
TS + LS	3.70	4.47	10.1
	Few-Shot Calibration		
Self-TS	3.56	4.62	8.34
Self-TS + LS	3.70	4.71	7.50
FSL	9.38	2.12	3.45
FSL + LS	5.19	3.02	4.24

(f) Detailed results on MARC with mBERT

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	1.93	4.28	5.88
TS	0.81	2.29	3.39
LS	4.83	3.37	3.85
TS + LS	6.30	4.44	5.07

(g) Detailed results on PAWS-X with XLMR

Method	ECE(en)	$\mathbb{E}_{l\in\mathcal{L'}}[\text{ECE}(l)]$	$\max_{l\in\mathcal{L'}}\text{ECE}(l)$
	Zero-Shot Calibration		
OOB	3.57	10.3	15.6
TS	0.99	5.64	9.80
LS	3.35	5.57	8.94
TS + LS	5.27	5.36	7.40

(h) Detailed results on PAWS-X with mBERT

Table 9: Detailed results on improving calibration for the 4 datasets and 2 MMLMs considered in our experiments