Calibrating Zero-shot Cross-lingual (Un-)structured Predictions

Zhengping Jiang, Anqi Liu, Benjamin Van Durme
Johns Hopkins University
{zjiang31, aliu74, vandurme}@jhu.edu

Abstract

We investigate model calibration in the setting of zero-shot cross-lingual transfer with large-scale pre-trained language models. The level of model calibration is an important metric for evaluating the trustworthiness of predictive models. There exists an essential need for model calibration when natural language models are deployed in critical tasks. We study different post-training calibration methods in structured and unstructured prediction tasks. We find that models trained with data from the source language become less calibrated when applied to the target language and that calibration errors increase with intrinsic task difficulty and relative sparsity of training data. Moreover, we observe a potential connection between the level of calibration error and an earlier proposed measure of the distance from English to other languages. Finally, our comparison demonstrates that among other methods Temperature Scaling (TS) generalizes well to distant languages, but TS fails to calibrate more complex confidence estimation in structured predictions compared to more expressive alternatives like Gaussian Process Calibration.

1 Introduction

While deep neural networks, especially large pre-trained language models, have driven striking improvements on various standard benchmarks (Wang et al., 2018, 2019), it is never a good practice to assume their predictions are accurate and should be taken blindly. In many cases, it is important to understand “what a model does not know” through its estimation of its uncertainty. For example, reliable model confidence is important in high-stakes domains (Begoli et al., 2019; Zhong et al., 2019), or when downstream tasks leverage confidence scores to mitigate error propagation (Chang et al., 2007). Moreover, accurate confidence can serve as a measure of the value of information in iterative data collection or human-in-the-loop learning (Zhang et al., 2019; Chaudhary et al., 2021).

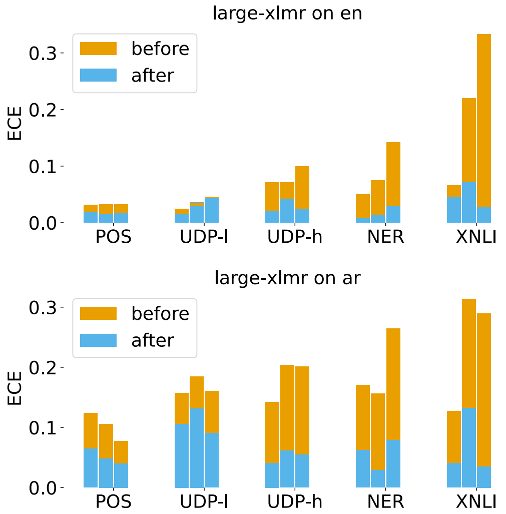

Figure 1: Averaged Expected Calibration Error (ECE) before and after temperature scaling on English (top) and Arabic (bottom) `xlm-roberta-large`; lower is better. Multiple bars for a task reference full-data, low-data, and very-low-data (from left to right) settings. Models appear less calibrated when transferred to other languages while temperature scaling remains effective.

Whether the model confidence is accurate is usually measured by how well it matches the observational data – through confidence calibration (Guo et al., 2017). Yet modern neural networks are criticized for being overconfident with their predictions, given their increased capacity to fit the training dataset (Guo et al., 2017). This problem is exacerbated by *domain-shift* (Ovadia et al., 2019) or zero-/few-shot transfer (Liu et al., 2018). An important task that is often concerned with such data shift is zero-shot cross-lingual transfer, which has been viewed as a natural extension to domain adaptation (Ruder et al., 2019; Xian et al., 2021).

Existing studies in natural language processing have mainly focused on zero-shot transfer accuracy alone (Wu and Dredze, 2019; K et al., 2020; Lauscher et al., 2020), without concern for the uncertainty measures of massive cross-lingual pre-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2648 - 2674
December 7-11, 2022 ©2022 Association for Computational Linguistics

training models (Devlin et al., 2019; Conneau et al., 2020a; Liu et al., 2020; Xue et al., 2021). On the other hand, large-scale uncertainty estimation and calibration work have mostly been conducted in the vision domain (Ovadia et al., 2019; Minderer et al., 2021). Large-scaled calibration studies put predominant importance on computer vision. In natural language processing, while model calibration has wide applications w.r.t tasks such as text classification (Jung et al., 2020; Kong et al., 2020), seq2seq generation (Ott et al., 2018; Dong et al., 2018; Wang et al., 2020), question answering (Ye and Durrett, 2022; Kamath et al., 2020) and zero-shot learning (Zhao et al., 2021), benchmarking results are not as comprehensive as in vision.

In this work, we evaluate how the calibration of large-scale multilingual models is affected by the zero-shot cross-lingual transfer, and whether we might mitigate calibration error with standard techniques relying solely on the source language[1]. We conduct our experiments on six standard cross-lingual transfer tasks across seven typologically diverse target languages, using English as the annotated source language. Our key findings include:

- NLP models become less calibrated under cross-lingual transfer.
- Task difficulty, data sparsity, and distance between source and target languages each impact model calibration, as shown in fig. 1.
- Post-training calibration methods using only the source language effectively mitigates miscalibration on target languages.
- Post-training calibration in structured prediction is more challenging and requires a more expressive calibration function family.

2 Background

2.1 Calibration in NLP Tasks

Why calibration in NLP tasks? Uncertainty quantification for neural networks and model calibration has received attention from various machine-learning-related fields, especially when machine learning is applied in the high-stakes decision-making (Gal and Ghahramani, 2016; Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Grathwohl et al., 2020; Thulasidasan et al., 2019). In particular, in NLP tasks, uncertainty plays an important role in AI-aided mental health diagnosis (Chandler et al., 2022) and human-in-the-loop active data curation (Yuan et al., 2022). Also, in language-model-based reasoning engines, the searching or filtering step often requires a faithful scoring rule (Dalvi et al., 2021; Weir and Van Durme, 2022; Creswell and Shanahan, 2022), which could be a potential application of a calibrated entailment classifier.

Calibration of large scale models Noticeably, Ovadia et al. (2019); Minderer et al. (2021) have produced large-scale benchmarks over a variety of tasks and existing calibration methods with mixed results. While empirically Ovadia et al. (2019) shows that the traditional post-training calibration methods such as temperature scaling do not always transfer under domain shift, results from Minderer et al. (2021) indicates that there is a correlation between in-domain and out-of-domain calibration error for models with large capacities like ViT (Dosovitskiy et al., 2021), and that model calibration decreases more slowly than accuracy.

In NLP, Desai and Durrett (2020) shows that pretrained transformer models achieve better calibration and that temperature scaling further reduces calibration error in-domain. Mohta and Raffel (2021) demonstrates that the benefit of the pretrained model diminishes as the domain shift increases. Our work extends these analyses to model calibration under zero-shot cross-lingual transfer.

Calibration of structured prediction Calibration of structured prediction models is relatively under-explored, due to the difficulty in defining the calibration setting (Kuleshov and Liang, 2015). Jagannatha and Yu (2020) proposed a general calibration scheme where the calibration is measured on the sequence level. Yet under challenging transfer conditions for difficult tasks, the top-k sequences do not contain enough positive events, and letting the event set of interest depends on model prediction making cross-method comparison difficult. In this paper, we investigate model calibration of structured prediction tasks as well as classification, given the high interest in tasks with a sequence tagging nature where one has to model inter-label dependencies in the multilingual community. We employ a slightly different setting with (Jagannatha and Yu, 2020) whence either tag-wise calibration is measured (Reich et al., 2020; Kranzlein et al., 2021), or a balanced set of a positive or negative set of spans is used to construct the event set of interest. In section 3.3 we discuss our formulation in detail, and show that it is compatible with the framework

[1] source code available at: `https://github.com/zipJiang/cross-lingual-calibration`

proposed by Kuleshov and Liang (2015).

2.2 Understanding Cross-Lingual Transfer

Since massive language model pretraining yielded promising zero-shot transfer results on cross-lingual datasets (Conneau et al., 2018), much effort has been put into understanding why these language models work and what is the limit of standard and direct zero-shot transfer paradigms (Wu and Dredze, 2019; Pires et al., 2019; Conneau et al., 2020b; Libovický et al., 2020; Chi et al., 2020; Hewitt and Manning, 2019; Yarmohammadi et al., 2021). While useful, these works tend to employ model performance as the sole metric; in this work, we investigate the reliability of confidence estimation.

A frequently discussed topic for cross-lingual transfer evaluation is how language-specific features are able to influence transfer performance. A common way to do this is to differentiate languages by language groups (Wu and Dredze, 2020; Chi et al., 2020). Other works rely on the numeric distance calculated from information depicting some specific aspect of language similarity (Lauscher et al., 2020; Pires et al., 2019). A line of research that tries to parameterize language relationships is typological embeddings (Littell et al., 2017; Malaviya et al., 2017; Cotterell and Eisner, 2017). Results from comprehensive transfer evaluation work also induce certain proximity between languages (Wu and Dredze, 2019; Han et al., 2019; Fan et al., 2021; Yu et al., 2021). We observe these various notions of distance result in similar orderings across languages. Therefore we follow previous work by loosely referring to this language-specific characteristic as "language similarity".[2]

3 Metrics and Methods

3.1 Measuring Model Calibration

Consider a classifier $\hat{\mathbf{p}} : \mathcal{X} \rightarrow \Delta^{k-1}$ that maps each instance $x \in \mathcal{X}$ to some class membership probability, $(\hat{\mathbf{p}}_i(x), \hat{\mathbf{p}}_2(x), \dots \hat{\mathbf{p}}_k(x))$. We describe $\hat{\mathbf{p}}$ as **calibrated**, or more specifically **confidence-calibrated** (Kull et al., 2019), if for any $c \in [0, 1]$:

$$\Pr(Y = \arg\max_i \hat{\mathbf{p}}_i(\mathbf{x}) | \max_i \hat{\mathbf{p}}_i(\mathbf{x}) = c) = c. \quad (1)$$

Directly calculating probability in eq. (1) with a finite number of examples is impossible. Several empirical approximations have been proposed (Guo et al., 2017). Here we adopt the Expected Calibration Error (Naeini et al., 2015) (ECE), which is the most prevailing statistic, and the Brier Score (Brier et al., 1950).

For N predictions, ECE approximates eq. (1) by splitting $[0, 1]$ into M equal length bins $\{B_1, B_2, \dots, B_M\}$, and calculates a weighted average of absolute difference between within-bin accuracy and within-bin average confidence:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|.$$

The ECE score is sensitive to the choice of binning schemes, and a model can trivially achieve a perfect ECE score by returning the marginal class probability. As a result, several works have proposed alternatives to ECE to mitigate such problems (Nixon et al., 2019). Kull et al. (2019) proposes the classwise-ECE, where the ECE is calculated and averaged across all class labels. Kumar et al. (2019) shows that it is always possible to construct a poorly calibrated prediction even when ECE $= 0$. Despite these shortcomings, we still use the ECE as our primary statistics for evaluating calibration error for two reasons. First, we observe only a little variance when gradually reducing the number of bins from 100 to 10. Second, some of our experiments require the classification among an indefinite number of labels, making classwise statistics inapplicable.

3.2 Post-training Calibration

For each task, we tune the calibrator parameter with a development (dev) set that is different from the model-selection dev on **English** only. We study four carefully-selected post-training calibration methods (listed below) on typical zero-shot cross-lingual transfer tasks. Firstly, the methods should be intuitively extendable to an indefinite number of classes that suit our tasks like dependency head predictions. Secondly, they have relatively fewer hyperparameters to tune and thus tend to be largely accuracy-preserving (Gruber and Buettner, 2022). Specifically, for methods that are only applicable to binary classifications (e.g., histogram binning and beta calibration), we follow previous practice by Wenger et al. (2020) and Patel

[2] Each proposed similarity metric is based on statistics about certain aspects of languages, they are not necessarily serving as a measurement of universal language distance.

et al. (2021) to use a one-vs-rest extension to multi-class classification over the outputs of multi-class classifiers. All the methods share the same class-wise binning strategy. We do not renormalize the scaled probability following previous work as it is reported to mitigate the accuracy degradation Patel et al. (2021).

Temperature Scaling (Guo et al., 2017) Given a logits vector $\mathbf{z} = (\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_k) \in \mathcal{R}^K$, temperature scaling produces a normalized class membership probability vector $(\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_k)$ by a single scalar parameter $T > 0$:

$$\mathbf{q}_i = \frac{\exp(\mathbf{z}_i/T)}{\sum_{i=1}^{K} \exp(\mathbf{z}_i/T)}.$$

Temperature scaling has been proven effective in other scenarios (Ovadia et al., 2019; Desai and Durrett, 2020) and has the property of not changing model prediction orders. This makes the post-training calibration orthogonal to the overall model performance.

Histogram Binning (Zadrozny and Elkan, 2002) divides all uncalibrated predictions $\hat{\mathbf{p}}_y(\mathbf{x})$ into M mutually exclusive bins $\{B_1, \ldots, B_M\}$ and assigns calibrated probabilities $\mathbf{q}_y^m(\mathbf{x})$ that minimizes the bin-wise square loss:

$$\mathcal{L}(\mathbf{q}) = \sum_{m=1}^{M} \sum_{x \in \mathcal{X}} \mathbf{1}\big[x \in B_m\big] (\mathbf{q}_y^m(x) - y)^2.$$

Notice that evaluating against ECE instead of class-wise metrics enables us to jointly calibrate all one-vs-rest probabilities induced from multi-class classifiers without renormalization.

Beta Calibration (Kull et al., 2017) is a calibration function family defined based on the likelihood ratio between two Beta distributions. In the one-vs-rest case the calibration map can be re-parameterized into a bivariate logistic regression with $\ln \hat{\mathbf{p}}_y(\mathbf{x})$ and $-\ln(1 - \hat{\mathbf{p}}_y(\mathbf{x}))$ to predict a binary label $\mathbf{1}[\hat{y} = y]$.

GPcalib (Wenger et al., 2020) fits a one-dimensional Gaussian process to the latent function $g : \mathbb{R} \to \mathbb{R}$ that transforms raw logits. Given uncalibrated logits vector $\mathbf{z}$, the model output probability $\mathbf{q}_i$ is then given by:

$$\mathbf{q}_i = \frac{\exp(g(\mathbf{z}_i))}{\sum_{j=1}^{K} \exp(g(\mathbf{z}_j))}.$$

When the dataset is large, Wenger et al. (2020) proposes to use inducing point methods (Hensman et al., 2015) for scalability. Since the GPcalib framework uses the same function to transfer all components of $\mathbf{z}$, it is straightforward to batch the latent process along a dimension with an indefinite number of classes.

3.3 Calibration for Structured Prediction

For structure prediction tasks, a natural question is whether explicitly modeling inter-label dependencies can help with model calibration. A similar comparison has been hinted by Jagannatha and Yu (2020) and Reich et al. (2020), but no experiments has been proposed. However, the label space is exponentially large when we consider predictions over a complete sequence. It is thus difficult to define a calibration objective.

In this work, we follow previous efforts and define a set of "Events of Interests" $\mathcal{I}(x)$ (Kuleshov and Liang, 2015; Jagannatha and Yu, 2020). Given the complete label space $\mathcal{Y}$ of a structured prediction task, an event $E \in \mathcal{I}(x)$ is a subset $E \subset \mathcal{Y}$, whose probability we would like to calibrate. For sequence labeling tasks, a natural choice for $\mathcal{I}(x)$ is the model prediction at each position. This falls back to calibrating a multi-class classifier at each sequence position for a standard masked language model with a classification head. But we need to perform the constrained forward-backward (Culotta and McCallum, 2004) marginalization for a conditional random field (Lafferty et al., 2001) based model. A more interesting case will be named entity recognition (NER), where extracting an entity span often consists of multiple tag-level predictions. Jagannatha and Yu (2020) proposes to define each $E \in \mathcal{I}(x)$ as a set of tag sequences $\{y_1, \ldots, y_N\}$ that contains a single span from top-k $p(y|x)$ decoding. This does not suit our purpose as it is not convenient to compare calibration performance between models under that setup. For example, the model with very high precision and confidence would be considered more calibrated than its counterparts with more diverse candidates.

To remedy this problem, we define $\mathcal{I}(x)$ as a set of events where each event E corresponds to a set of sequences that extracts one of all possible span candidates $s \in \mathcal{S}$. This is equivalent to evaluating a model to perform binary classifications over whether a candidate is actually a valid span. Since the number of possible span candidates grows quadratically with the sequence length, we only consider spans with no more than a certain length l. Specifically, given a NER task with named entity type space $\mathcal{C}$ (e.g., "PER", "LOC", etc.), denote the corresponding tag space by $\mathcal{B}$ ("B-PER", "I-PER", "O", etc.). The probability of a span s with type $c \in \mathcal{C}$ and end points $1 \leq i < j \leq N = |x|$ being extracted under BIO sequence tagging can be written as:

$$\Pr(s, c|x) = \sum_{y \in \mathcal{Y}} \left\{ p(y|x) \prod_{k=i}^{j+1} \mathbf{1}\left[y_k \in s_k\right] \right\},$$

where $(s_i, \ldots, s_j, s_{j+1})$ is the tag subset sequence $(\{\text{B-}c\}, \ldots, \{\text{I-}c\}, \mathcal{B} \setminus \{\text{I-}c\})$. The classifier output can be directly multiplied to obtain the conditional probability when tags are independent. In the case of linear-chain CRF, we apply a constrained Forward-Backward (FB) algorithm.

4 Experiments

Tasks We consider six zeros-shot cross-lingual transfer tasks: part-of-speech tagging (POS), universal dependency parsing (UDP), named entity recognition (NER), cross-lingual natural language inference (XNLI), Automatic Content Extraction (ACE) and the Better Extraction from Text Towards Enhanced Retrieval (BETTER). These six tasks are of distinct formulations and have a reasonable spread over difficulty levels. For detailed data configuration and task descriptions, please refer to appendix A. Also, only plots relevant to the discussion are presented inline, please also refer to appendix A for complete experiment data.

Evaluation we evaluate the calibration before and after a post-training calibration step using ECE. To properly evaluate the ECE, we set `num_bins=100`. We choose this number to balance granularity with the amount of data, as we observe ECE tends to converge after the number of bins increases above a threshold. This binning scheme has been employed to evaluate calibration methods (Wenger et al., 2020; Minderer et al., 2021).

Base models We experiment with three common multilingual transformer encoders: `bert-base-multilingual-cased`, `xlm-roberta-base` and `xlm-roberta-large` .[3] We keep the token embedding weight fixed for all our experiments, and use `learning_rate = 1.2e-5` for pretrained transformer parameters, and `learning_rate = 1e-5` for the rest of models (except for very-low-data NLI, where we choose `learning_rate = 1e-4`).

Varying training size We evaluate our pipeline with three training-data-size configurations when available (that is, on POS, UDP NER and XNLI): *full-dataset*, where all the specified training data are used; *low-data*, where 1000 sentences are sampled for the sentence-level dataset, or 50 documents are sampled for the doc-level dataset; *very-low-data*, where 100 sentences or 10 documents are sampled respectively.

Training details We train our models on a single RTX 6000 GPU until convergence or a maximum number of epochs (256) is reached. We use the dev set for model selection and early stopping, and gradually scale our learning rate by .25 on a plateau. For all tasks, we apply the four calibration methods mentioned in section 3.2 as the post-training calibration step. We set `learning_rate = .1` and use a large batch size to tune the calibration module parameters. We also gradually scale the learning rate by .25 on a plateau. The learning rate for temperature scaling is determined via an Optuna (Akiba et al., 2019) trial with a searching range between [5e-2, .5] on subtasks. For each calibration method, we do 10 runs and do a significant test with classic bootstrap from the dataset to address the concern of randomness raised by Vaicenavicius et al. (2019).

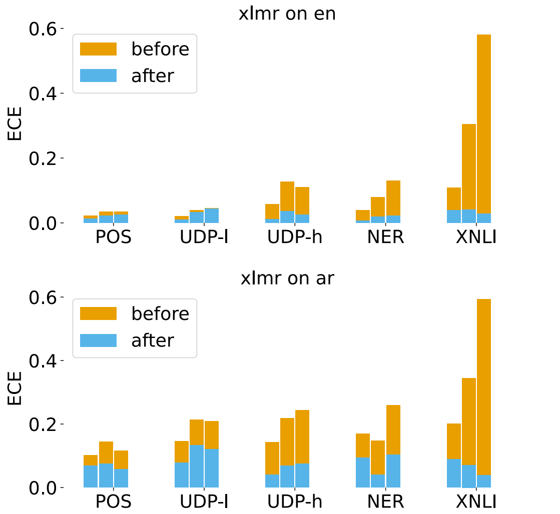

Figure 2: Averaged ECE before and after temperature scaling on English (top) and Arabic (bottom) for `xlm-roberta-base`; lower is better. Each bar in a group corresponds to a training data theme as in fig. 1.

[3] https://huggingface.co/models

4.1 Impact of Training Configurations

Impact of Data Size In most cases, training with more data helps calibration, especially when the difference in training data size is large (e.g., comparing the full-data setting and very-low-data setting. see fig. 2). However, we do not observe such a tendency when the task is simple enough and the model performance is reasonably high, like in POS. It indicates that the representation for the task has already been learned well during the pre-training phase, and the relevant information is easily recovered even with a small number of examples. Interestingly, the XNLI model trained under a very-low-data setting can be similarly or even better calibrated compared to the XNLI model trained under a full-data setting after post-training calibration, though the gap of accuracy for models trained with different data amounts is large (accuracy results are available in the appendix A). It indicates that a more accurate model is not always more calibrated.

Impact of Language Similarity Our result indicates that target language calibration errors are generally lower when the target language is similar to English as measured by human language learning distances (Chiswick and Miller, 2005) (see fig. 1, fig. 4, etc.). While the distance between languages is an intuitive concept among linguists in the abstract, there is no prevailing theory on how this should be quantitatively measured. We abstain from calculating direct correlations with scores proposed by Chiswick and Miller (2005), merely noting that further investigations into the relationship between language distance and domain shift are worth future consideration. This echoes the result from the previous research (Lauscher et al., 2020; Pires et al., 2019) showing that commonly perceived language difference influences the difficulty of zero-shot transfer. However, post-training calibration often has a smaller effect on more similar target languages.

Impact of Pretrained Model Size Giving the similar trend observed for different calibration methods, here we only plot post-training calibration statistics for temperature scaling (See section 4.2 below). Comparing results shown in fig. 1 and fig. 2, we come to the conclusion that the larger pre-trained language model is usually more calibrated before and after the post-training calibration. Though both large and base models become less and less calibrated while gradually transferring to more and more distant languages, the calibration error increases more slowly than the calibration error of smaller models. This becomes more prominent when the training data is smaller or the target language is more distant (see fig. 3). We hypothesize this is probably because a larger language model learns better cross-lingual representations that allow a better zero-shot cross-lingual transfer with sufficient training data. This echos previous findings by Minderer et al. (2021), where they have shown that the calibration error increases more slowly for larger models.

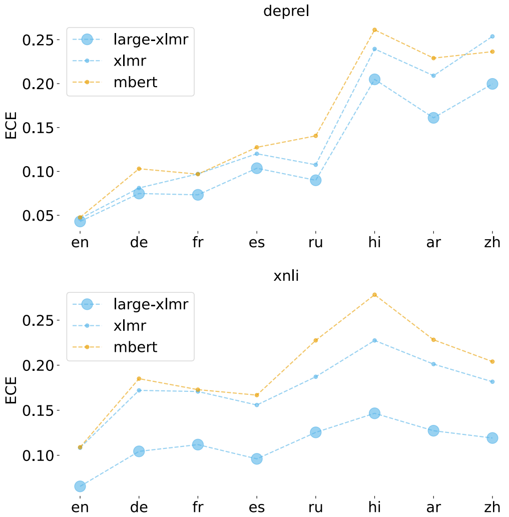

Figure 3: Calibration plot for different models when transferred to different languages on UDP (top) with *very-low-data*, and XNLI (bottom) with *full-data*. The result shows that a larger model generalizes better when the training data size is small or the task is difficult.

4.2 Comparing Calibration Methods

We do 10 runs of classic bootstrap from each dataset to evaluate all four calibration methods mentioned in section 3.2. All of the methods are able to significantly reduce the calibration error in terms of the ECE (see appendix A for complete statistics). fig. 4 demonstrates the effectiveness of different post-training calibration methods. In most cases, different calibration methods have similar performance. Models calibrated by any of the methods are still likely to be less and less calibrated when zero-shot transferred to more and more distant languages as described in section 4.1. In most cases, either temperature scaling or GPcalib is at or near the best, under all training data source settings. Histogram binning performs well in the source language, but it may decline the most in effectiveness in the test language. Moreover, when the model is zero-shot transferred to more distant languages, temperature scaling gains a small edge compared to other methods.

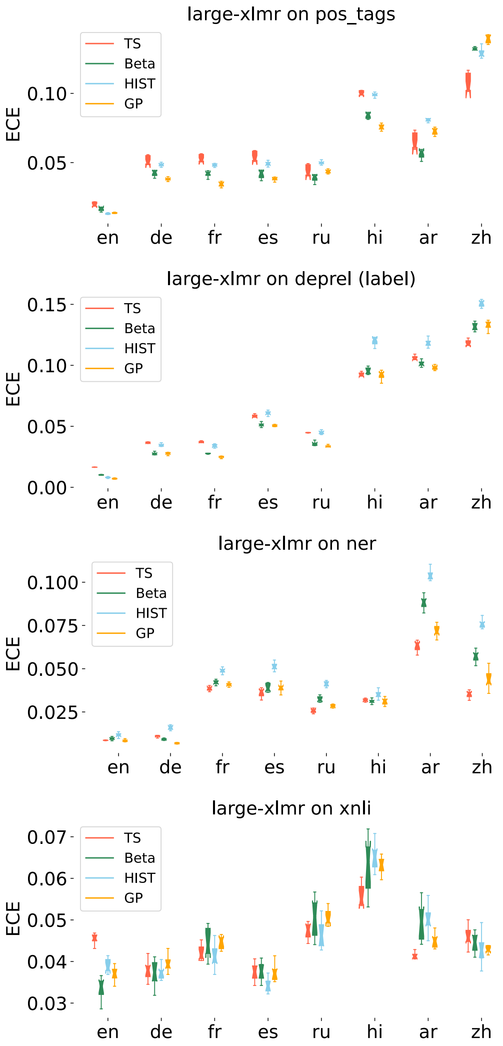

Figure 4: Box-and-whisker plots for different calibration methods for `xlm-roberta-large` when zero-shot transferred to different languages, sorted by language distance to English (Chiswick and Miller, 2005). Each calibration method is tested 10 times with bootstrapping.

Another observation is that the calibration effectiveness of methods is more variable on XNLI than other tasks, and the model calibration error after post-training calibration follows the language distance less strictly. This becomes more notable when examining smaller models and fewer training samples, as shown in fig. 5. This could be due to that XNLI requires more complex semantic knowledge (Lauscher et al., 2020) that is not directly accessible in the multilingual encoder, making the calibration less transferable to other languages.

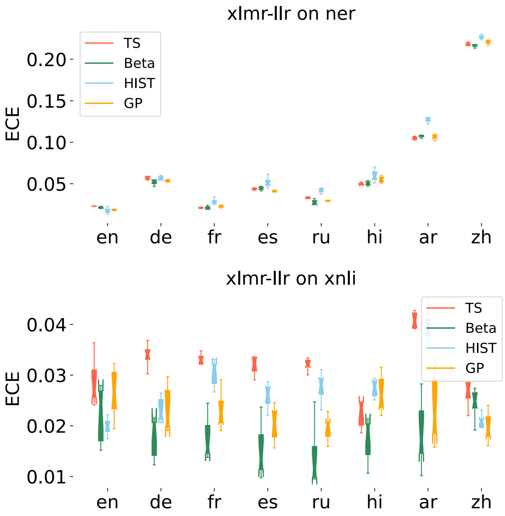

Figure 5: Box-and-whisker plots for different calibration methods for `xlm-roberta-base` on *very-low-data* setting when zero-shot transferred to different languages, sorted by language distance to English (Chiswick and Miller, 2005). Each calibration method is tested 10 times with bootstrapping.

4.3 Calibration for Structured Prediction

We also consider model calibration for two structured prediction tasks: POS tagging and NER. We follow the definition of $\mathcal{I}(x)$ in section 3.2. The WikiAnn dataset (Pan et al., 2017) is very suitable for our purposes as it contains many short sequences that avoid span number explosion. We further restrict the maximum span length $l = 5$ and the maximum sequence length $s = 32$ to reduce the search space. To prevent the model from reducing calibration error by scaling down the extraction probability of all spans, we further subsample negative samples by probability $p = .01$. Notice that this kind of subsampling can be viewed as an adjusted environment for robust calibration and should not affect a perfectly calibrated model (Wald et al., 2021). It also corresponds in practice to the use case of performing span filtering from a high-quality subset.

However, when applied to structured labels like in span extraction, temperature scaling could be less effective. Particularly in NER calibration, we observe that GPcalib achieves a significantly better calibration result when compared to temperature scaling (see fig. 6), while on POS we do not observe such a gap. It could be that the structure for label spans is more complex and usually involves multiple labeling predictions. Therefore, in order to calibrate these probability combinations, one

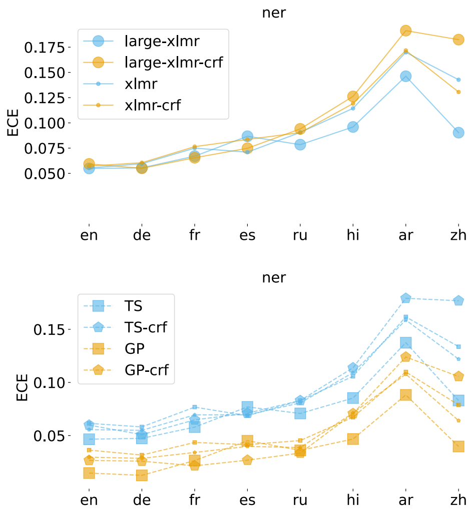

Figure 6: **Top**: Adding a CRF module doesn't helpful to model calibration either on the source language or on the target language, regardless of model size. **Bottom**: GPcalib is more effective in calibrating structured prediction regardless of the underlying model structure.

Task		F-1	t-ECE	f-ECE
mBERT				
ACE	raw	58.57	12.67	21.42
	cal.	-	10.73	21.89
BETTER	raw	35.26	17.85	32.87
	cal.	-	12.98	22.37
XLM-R				
ACE	raw	58.19	11.30	36.76
	cal.	-	11.45	32.56
BETTER	raw	36.24	14.60	39.83
	cal.	-	10.70	18.69

Table 1: Results for En-Ar transference (English). The **raw** row corresponds to out-of-the-box model and the **cal.** row shows the calibration error reduction by temperature scaling. **t-ECE** corresponds to the ECE of span typing, **f-ECE** corresponds to the ECE of span finding.

Task		F-1	t-ECE	f-ECE
mBERT				
ACE	raw	19.13	20.49	72.59
	cal.	-	12.68	71.76
BETTER	raw	18.45	23.68	58.37
	cal.	-	9.5	27.00
XLM-R				
ACE	raw	26.74	13.84	67.40
	cal.	-	13.36	62.40
BETTER	raw	23.68	21.05	57.26
	cal.	-	9.96	8.29

Table 2: Results for En-Ar transference (Arabic). The row & column interpretation is similar to table 1.

will need a more complex function family, which is not included in temperature scaling.

4.4 Evaluating on More Difficult Tasks

We further experiment with two more information extraction tasks, ACE and BETTER, where the training resource is more limited and the ontologies are more complex. For labeling problems, we follow the general setting in section 4.1. For tagging problems, we calibrate the label-wise probability for positive labels as discussed in section 3.3. In case of a linear chain CRF, we marginalize out all other positions to obtain the label-wise probability following Culotta and McCallum (2004) and Reich et al. (2020). For ACE and BETTER we do not evaluate under *low-data* and *very-low-data* setting given the relatively small size of their dev sets.

Impact of Task Type and Difficulty Our results align with the discovery of Lauscher et al. (2020), who showed that the transfer performance depends on a hypothetical "task level". Here we observe a larger ECE on ACE and BETTER as well as in "high level" semantic tasks like XNLI compared to "low level" sequence tagging tasks like POS, UDP, NER defined by Lauscher et al. (2020).

As shown in table 1 and table 2, in general, the structured prediction components (f-ECE) are less calibrated and remain so after temperature scaling, though for ACE there is some irregularity given the sparse event/argument span annotations on the English side on which our model has very high accuracy. To further show the impact of task difficulty on the model calibration, we observe that when trying to perform post-training calibration of ACE and BETTER models with temperature scaling, the scaling parameters tend to be larger, even reaching 38.45 for span-finding under a particular configuration while normally the scaling parameters are below 5 as shown in table 3 and table 4.

5 Conclusions

We explore the model calibration of large language models under the zero-shot cross-lingual transfer scenario. Our results show that the extent of miscalibration varies according to a number of aspects

Size	POS	UDP-label	UDP-head	NER	ACE-t	ACE-f	BETTER-t	BETTER-f
full	1.5	1.58	1.85	1.9	3.32	1.11	3.00	10.94
low	1.42	1.77	3.23	1.97	-	-	-	-
very-low	1.47	2.12	2.96	1.88	-	-	-	-

Table 3: Temperature scaling parameter for `bert-base-multilingual-cased`, from one run. In multilingual experiments, scaling parameters tend to be larger when the training data size is small, and scaling parameters for BETTER-{t, f} are larger as the task is more difficult.

Size	POS	UDP-label	UDP-head	NER	ACE-t	ACE-f	BETTER-t	BETTER-f
full	1.30	1.51	1.72	1.47	1.01	1.12	3.80	38.45
low	1.66	1.80	3.39	1.79	-	-	-	-
very-low	1.43	2.10	2.09	2.01	-	-	-	-

Table 4: Temperature scaling parameter for `xlm-roberta-base`, from one run. Scaling parameters exhibit similar trends as in the `bert-base-multilingual-cased` case.

of the training configuration. First, training with more data improves the cross-lingual calibration. Second, transferring from English to non-English intensifies miscalibration as the target language is farther from English. Also, larger models are likely to be less miscalibrated when the model is transferred to a different target language zero-shot. Moreover, our result shows that temperature scaling and Gaussian Process calibration methods are among the top-performing methods. While temperature scaling is easy to implement and generalizes well to distant languages, it's less effective when applied to complex structured probabilities. Finally, models are least calibrated on "high levels" tasks like XNLI and span extraction. Models are most calibrated on simple "low-level" tasks like POS.

Our results demonstrate that model confidence scores are useful metrics to understand the model behavior and language transfer pairs in cross-lingual tasks. We encourage users to calibrate their model before zero-shot deployment to produce more reliable confidence estimation and prevent over-confidence for downstream tasks.

6 Limitations

We discuss several limitations of the work. Regarding the research approach, our work focuses on the empirical investigation of model calibration under zero-shot cross-lingual transfer. We are interested in extensive theoretical explanations for the variability of model calibration under zero-shot cross-lingual transfer in the future. Also, regarding the results, even though we show promising calibration performances on pre-trained models, we are aware that calibration of LLMs is generally very hard, especially when the task is not a standard classification task and the validation data is not available. This is a promising research direction that should draw more attention by the community. It is especially relevant to the application of LLMs to safety-critical or high-stakes tasks. We leave further investigation of calibration for LLMs to future work.

Acknowledgement

We thank Yunmo Chen, Marc Marone, Guanghui Qin, Kate Sanders, and Elias Stengel-Eskin for the discussion on the experiment design and their valuable feedback on the writing.

This work was supported in part by IARPA BETTER (#2019-19051600005). The views and conclusions contained in this work are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, or endorsements of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23.

Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Chelsea Chandler, Peter W Foltz, and Brita Elvevåg. 2022. Improving the applicability of ai for psychiatric applications through human-in-the-loop methodologies. *Schizophrenia Bulletin*.

M Chang, Quang Do, and Dan Roth. 2007. Multilingual dependency parsing: A pipeline approach. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 292:55.

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Barry R Chiswick and Paul W Miller. 2005. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26(1):1–11.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*.

Aron Culotta and Andrew McCallum. 2004. Confidence estimation for information extraction. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 109–112.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2020. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*.

Sebastian Gruber and Florian Buettner. 2022. Better uncertainty calibration via proper scores for classification and beyond.

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

Ligong Han, Yang Zou, Ruijiang Gao, Lezi Wang, and Dimitris Metaxas. 2019. Unsupervised domain adaptation via calibrating uncertainties. In *CVPR Workshops*, volume 9.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. 2015. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhyuday Jagannatha and Hong Yu. 2020. Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092, Online. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf. 2020. Posterior calibrated training on sentence classification tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2723–2730, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Michael Kranzlein, Nelson F. Liu, and Nathan Schneider. 2021. Making heads and tails of models with marginal calibration for sparse tagsets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4919–4928, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Meelis Kull, Telmo Silva Filho, and Peter Flach. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pages 623–631. PMLR.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles.

Advances in neural information processing systems, 30.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An empirical study of pre-trained transformers for Arabic information extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4727–4734, Online. Association for Computational Linguistics.

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.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. 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.

Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. 2018. Generalized zero-shot learning with deep calibration network. *Advances in Neural Information Processing Systems*, 31.

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.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34.

Jay Mohta and Colin Raffel. 2021. The impact of domain shift on the calibration of fine-tuned models. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

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 the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional lstm-crf for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. 2021. Multi-class uncertainty calibration via mutual information maximization-based binning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J.

Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, 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.

Steven Reich, David Mueller, and Nicholas Andrews. 2020. Ensemble Distillation for Structured Prediction: Calibrated, Accurate, Fast—Choose Three. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5583–5595, Online. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 71–79.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. 2019. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. 2021. On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, 34.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus LDC2006T06. *Web Download. Philadelphia: Linguistic Data Consortium.*

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Nathaniel Weir and Benjamin Van Durme. 2022. Dynamic generation of interpretable inference rules in a neuro-symbolic expert system. *arXiv preprint arXiv:2209.07662*.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. 2020. Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, pages 178–190. PMLR.

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. 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. 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.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th*

Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 149–159, Online. Association for Computational Linguistics.

Ruicheng Xian, Heng Ji, and Han Zhao. 2021. Learning invariant representations on multilingual language models for unsupervised cross-lingual transfer. In *International Conference on Learning Representations*.

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.

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.

Xi Ye and Greg Durrett. 2022. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6199–6212, Dublin, Ireland. Association for Computational Linguistics.

Dian Yu, Taiqi He, and Kenji Sagae. 2021. Language embeddings for typology and cross-lingual transfer learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7210–7225, Online. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics.

Daniel Zeman et al. 2021. Universal dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Yao Zhang et al. 2019. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical science*, 10(35):8154–8163.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Ruiqi Zhong, Yanda Chen, Desmond Patton, Charlotte Selous, and Kathy McKeown. 2019. Detecting and reducing bias in a high stakes domain. 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 4765–4775, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Detailed Task Descriptions

We evaluate the model calibration for the zero-shot cross-lingual transfer on a variety of classification and sequence-tagging tasks when they are used out-of-box and after post-training calibration. Our experiments largely follow the established settings by Yarmohammadi et al. (2021). For multi-lingual experiments, we consider Part-Of-Speech (POS) tagging, Universal Dependency Parsing (UDP), Named Entity Recognition (NER), and Natural Language Inference (NLI). For English-Arabic experiments, we additionally consider ACE[4] and BETTER[5]as they are only available to limited languages. We use English as the source language and seven target languages that are diverse in their typology (Clark et al., 2020; Pimentel et al., 2021). In cases where alternative English-side dev sets are available (NLI, POS, UDP) we directly use different dev sets for model selection and post-training calibration, otherwise, we split the dev set.

[4] https://www.ldc.upenn.edu/collaborations/past-projects/ace

[5] https://www.iarpa.gov/index.php/research-programs/better

Part-of-speech (POS) Tagging We use the Universal Dependencies (UD) Treebank (v2.9; Zeman et al, 2021).[6] The UD Treebank consists of data from a variety of sources, such that there may be potential domain mismatch across different treebanks (Sato et al., 2017). To overcome the domain discrepancy across different languages, we use the New Parallel UD (PUD) (Zeman et al., 2017) treebank test set in the UD Treebanks as the test set, which is available to all our target languages. Similar to NER, we generate the word representation by attention-weighting all subword token representations. We use a linear classifier to predict corresponding POS tags. We evaluate the performance by the accuracy of predicted POS tags.

Universal Dependency Parsing (UDP) We use the same set of treebanks as in appendix A.1 for the POS tagging task. To predict the dependency heads and dependency labels, we use a biaffine attention layer (Dozat and Manning, 2017). As in POS and NER, we generate word-level representations by attention-weighting the subword token representations. We evaluate the performance by labeled attachment score (LAS). For this task, we evaluate the model calibration for both the head prediction and the label prediction.

Named Entity Recognition (NER) We rely on WikiAnn (Pan et al., 2017) for named entity recognition. We use the Hugging Face Datasets version[7] which corresponds to the balanced train, dev, and test splits in Rahimi et al. (2019). Labels of the dataset consist of three types of named entities: PER, LOC, and ORG. We use an additional linear layer to predict word-level labels over the word representation, which is aggregated through an attention layer over the subword-level representation generated by the encoder. We evaluate the NER performance by the F1 score of the predicted entity.

Natural Language Inference (NLI) We evaluate the cross-lingual natural language inference performance with XNLI (Conneau et al., 2018). We train on the MultiNLI (Williams et al., 2018) training set. For a given instance, we concatenate the premise p and the hypothesis h as the joint input to our model. To predict the entailment label, we apply a linear classification head over the pooled sentence representation. We evaluate the model performance using the prediction accuracy.

ACE We use the English and Arabic subset of Automatic Content Extraction (ACE) 2005 (Walker et al., 2006) following Yarmohammadi et al. (2021). We evaluate our model on the trigger extraction and the argument extraction subtasks. We utilize the event extraction model of Xia et al. (2021), which consists of a BiLSTM-CRF BIO tagger (Panchendrarajan and Amaresan, 2018) and a type-classifier trained to predict child spans conditioned on parent spans and labels. This model structure yields a comparable performance to the state-of-the-art OneIE (Lin et al., 2020) on the trigger and argument identification. Here we use a shared model structure with other tasks in **BETTER** to facilitate the direct performance comparison. We use the same English split as in Lin et al. (2020), and for the Arabic split, we follow Lan et al. (2020).

BETTER The Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program aims to "develop enhanced methods for personalized, multilingual semantic extraction and retrieval from the text", given gold annotations only in English. Unlike in Yarmohammadi et al. (2021) which focused on "Abstract" event extraction, here we focus on the richer "Basic" task. Basic event extraction, structurally related to the FrameNet parsing, requires a model to identify a finer-grained set of event types than Abstract, along with their respective agent, patient, or event references. The documents come from the news-specific portion of Common Crawl. Performance on BETTER Basic is evaluated according to a program-defined "combined F1" metric, which is the product of "event match F1" and "argument match F1", calculated based on the best-effort alignment of predicted and reference event structures. We use the same model structure as in ACE. We run the model for multiple passes to produce level-wise predictions in parallel at the inference time.

A.2 Complete Multilingual Experiment Results

In this section, we present additional results for the multilingual experiment setting for all three encoders (`xlm-roberta-large`, `xlm-roberta-base` and `bert-base-multilingual-uncased`) and all training data size configurations (*full-data*, *low-data* and *very-low-data*). Results are shown in table 5 to table 11. As discussed in the main paper, the

[6] We train on the following English treebanks: English-Atis, English-EWT, English-GUM, English-LinES, English-ParTUT, and English-Pronouns.

[7] `https://huggingface.co/datasets/wikiann`

general trend is that models before and after post-training calibration become less calibrated with less training data. Also, as models are transferred to more distant languages the, the calibration error before and after post-training calibration usually becomes higher. Under most circumstances for classification tasks, the improvement is significant for any of the standard calibration methods when comparing with off-the-shelf models. In particular, temperature scaling (TS) performs on par with more expressive calibration function families. However as shown in table 14 and table 15, for more structured prediction tasks, more expressive calibration methods like GPcalib are preferred over TS, usually, with a significant margin.

Table 5, table 6 and table 7 show the results for `xlm-roberta-large` on three data settings. `xlm-roberta-large` in general achieves the best predictive performance as well as the best calibration in the zero-shot transfer, especially on distant languages. Table 8, table 9 and table 10 show the results for `xlm-roberta-base` on three data settings. They exhibit similar trends with results achieved by `xlm-roberta-large`. Similarly, table 11, table 12 and table 13 show the same general trends for `bert-base-multilingual-uncased`. These results indicate that the tendency of zero-shot cross-lingual calibration transfer is consistent across LMs with different backbones. In all these experiments, we observe significant improvements in ECE after post-training calibration, both on the source language and on target languages.

Table 14 and table 15 show the result for structured prediction tasks with and without a CRF prediction head for POS-tagging and NER. For POS-tagging, TS tends to be effective compared to GP-calib. But its effectiveness on a more complex task, NER, is less pronounced. For NER, the model has to make correct probability prediction on a sequence of tokens. In this kind of task, GPcalib often significantly out-performed TS.

Task	**Metric(%)**	**en**	**de**	**fr**	**es**	**ru**	**hi**	**ar**	**zh**
POS	*Acc*	96.47	91.60	90.98	91.17	91.21	82.29	84.59	74.36
	ECE	3.16	7.15	6.97	7.50	7.70	13.39	12.42	18.56
	TS	1.98	5.15	5.30	5.43	4.41	10.02	6.54	10.75
	Beta	1.66	4.24	4.18	4.16	3.90	8.40	5.64	13.18
	GPcalib	1.36	3.81	3.43	3.83	4.35	7.57	7.25	13.93
	HIST	1.32	4.85	4.81	4.92	4.99	9.89	8.09	12.97
UDP	*LAS*	88.10	84.45	82.34	79.43	78.73	50.76	65.51	48.38
	l-ECE	2.48	5.87	5.50	8.61	7.18	16.20	15.71	19.36
	l-TS	1.65	3.66	3.72	5.84	4.50	9.25	10.60	11.82
	l-Beta	1.02	2.76	2.78	5.12	3.59	9.55	10.15	13.17
	l-GPcalib	0.71	2.76	2.49	5.03	3.36	9.21	9.79	13.30
	l-HIST	0.81	3.46	3.40	6.09	4.49	12.04	11.84	15.08
	h-ECE	7.17	6.43	9.40	10.06	9.49	26.14	14.23	29.96
	h-TS	2.18	2.35	3.20	3.30	2.75	11.07	4.11	18.29
	h-Beta	2.03	1.90	2.95	3.28	2.41	12.82	4.41	19.19
	h-GPcalib	1.78	2.87	2.86	2.74	2.88	9.87	3.64	17.31
	h-HIST	2.12	1.97	4.13	4.43	3.93	15.70	6.93	21.20
NER	*F-1*	87.69	85.01	81.12	80.35	77.31	81.62	68.72	58.85
	ECE	5.04	4.16	9.17	10.52	8.76	8.86	17.04	13.33
	TS	0.86	1.12	3.85	3.60	2.57	3.18	6.29	3.51
	Beta	0.96	0.93	4.21	3.92	3.24	3.11	8.81	5.71
	GPcalib	0.85	0.74	4.08	3.88	2.86	3.15	7.22	4.38
	HIST	1.17	1.59	4.88	5.13	4.12	3.53	10.44	7.61
XNLI	*Acc*	87.86	81.80	82.44	83.51	79.12	75.71	77.94	78.36
	ECE	6.55	10.44	11.19	9.60	12.54	14.66	12.73	11.92
	TS	4.52	3.81	4.22	3.74	4.73	5.59	4.14	4.62
	Beta	3.36	3.87	4.62	3.86	5.18	6.47	5.11	4.56
	GPcalib	3.73	3.95	4.46	3.70	5.06	6.30	4.50	4.37
	HIST	3.89	3.67	4.18	3.42	4.66	6.47	5.02	4.27

Table 5: Experiment result with `xlm-roberta-large` on the *full-data* setting, shaded cells indicate significant improvements in calibration decided by a bootstrap from dataset and an independent t-test with $p < .05$.

Task	Metric(%)	en	de	fr	es	ru	hi	ar	zh
POS	*Acc*	96.33	91.93	85.74	90.42	90.96	79.78	85.10	67.22
	ECE	3.26	6.74	7.57	9.90	7.32	15.27	10.55	20.34
	TS	1.62	6.79	6.63	10.50	3.02	13.39	4.85	6.49
	Beta	1.45	6.42	5.00	8.24	2.78	10.97	4.42	8.19
	GPcalib	1.15	2.47	3.87	4.60	2.46	7.34	3.82	13.46
	HIST	0.52	6.40	6.23	7.99	3.62	11.33	6.61	10.53
UDP	*LAS*	88.27	78.38	77.15	74.92	71.00	45.77	58.25	44.35
	l-ECE	3.04	6.99	7.07	9.86	9.24	20.36	18.42	21.58
	l-TS	3.55	6.08	5.63	7.90	7.27	12.73	13.16	13.87
	l-Beta	1.57	2.94	2.71	4.98	4.27	10.94	10.60	12.22
	l-GPcalib	1.58	1.79	3.11	3.53	2.88	8.40	7.68	11.32
	l-HIST	1.90	4.34	5.88	7.47	7.24	18.53	16.37	20.39
	h-ECE	7.11	12.28	14.28	14.13	16.05	33.06	20.41	35.48
	h-TS	4.25	4.46	6.26	5.28	4.33	8.84	6.24	14.46
	h-Beta	3.95	2.39	3.15	3.07	3.23	15.15	5.38	19.62
	h-GPcalib	5.46	5.95	6.68	5.87	5.41	6.43	7.35	10.79
	h-HIST	3.65	1.66	3.87	3.37	4.37	16.52	7.09	21.03
NER	*F-1*	82.91	83.62	80.40	79.18	71.73	77.76	69.78	55.61
	ECE	7.51	4.56	8.76	11.21	10.66	11.62	15.63	14.76
	TS	1.41	3.03	2.20	3.39	2.56	3.79	2.91	3.97
	Beta	1.26	2.18	1.66	2.76	2.49	3.19	3.25	3.92
	GPcalib	0.83	1.84	1.34	2.94	2.10	3.35	3.34	4.36
	HIST	1.31	2.97	2.04	3.09	3.09	3.85	6.00	5.87
XNLI	*Acc*	76.79	70.86	71.98	73.25	68.84	65.23	66.83	67.60
	ECE	22.00	27.62	26.60	25.33	29.43	32.87	31.35	30.47
	TS	7.20	10.71	10.14	9.26	12.00	14.46	13.23	12.41
	Beta	5.71	8.66	8.07	7.29	9.49	11.83	11.08	10.17
	GPcalib	4.30	6.65	6.40	6.20	7.93	9.66	9.16	7.99
	HIST	1.51	6.68	5.64	4.66	8.63	12.00	10.24	9.32

Table 6: Experiment result with `xlm-roberta-large` under the *low-data* setting. The color scheme is the same as figures above.

Task	Metric(%)	en	de	fr	es	ru	hi	ar	zh
POS	*Acc*	95.45	91.46	84.00	89.96	90.39	79.60	84.55	64.80
	ECE	3.27	6.76	6.44	10.74	5.71	12.95	7.76	15.42
	TS	1.70	7.06	6.58	11.13	3.07	11.75	4.02	5.65
	Beta	1.48	6.08	4.81	8.74	2.87	9.03	3.20	6.67
	GPcalib	1.28	3.26	3.59	7.16	2.17	7.67	2.49	12.19
	HIST	0.92	5.78	5.18	7.84	3.59	9.15	3.76	8.83
UDP	*LAS*	77.62	66.92	65.18	63.55	61.19	36.31	47.84	34.03
	l-ECE	4.30	7.48	7.33	10.36	8.99	20.49	16.10	19.97
	l-TS	4.54	5.50	5.89	6.33	6.97	9.37	9.11	9.30
	l-Beta	2.28	2.69	2.34	4.54	3.59	11.87	8.07	11.77
	l-GPcalib	2.77	2.30	3.80	3.59	3.45	7.98	4.62	8.97
	l-HIST	2.82	3.44	4.53	5.87	5.28	15.88	10.97	15.69
	h-ECE	9.96	13.82	15.14	14.60	17.13	31.44	20.18	35.51
	h-TS	2.43	3.73	4.81	5.22	4.45	13.12	5.54	18.43
	h-Beta	4.40	3.31	3.82	4.41	2.61	13.47	4.14	17.81
	h-GPcalib	6.00	9.00	9.64	10.20	6.65	6.09	10.26	9.48
	h-HIST	4.34	2.83	4.16	4.75	4.01	15.63	6.52	19.84
NER	*F-1*	70.90	72.46	68.69	69.97	52.60	69.35	55.34	35.75
	ECE	14.17	8.51	15.69	17.32	20.54	18.81	26.45	32.04
	TS	2.90	4.33	2.96	3.55	6.88	4.69	7.95	14.16
	Beta	2.16	3.61	2.20	3.34	6.09	3.66	7.65	12.98
	GPcalib	1.33	3.51	2.10	3.64	5.06	4.48	6.78	12.30
	HIST	1.79	4.65	1.99	4.41	6.96	5.03	10.55	15.19
XNLI	*Acc*	40.54	38.08	40.32	39.38	35.99	39.04	38.92	37.98
	ECE	33.31	33.14	24.96	32.19	37.99	31.07	28.90	30.82
	TS	2.75	3.68	4.64	2.95	4.87	3.79	3.51	3.54
	Beta	2.19	4.76	3.02	3.39	6.90	3.54	2.88	3.70
	GPcalib	2.33	3.87	3.08	2.68	5.40	3.38	2.52	3.35
	HIST	3.99	5.28	5.02	4.20	7.03	5.53	3.98	5.18

Table 7: Experiment result with `xlm-roberta-large` under the *low-low data* setting. The color scheme is the same as figures above.

Task	Metric(%)	en	de	fr	es	ru	hi	ar	zh
POS	*Acc*	96.39	91.53	90.31	90.81	91.09	74.48	82.87	77.70
	ECE	2.27	5.52	5.03	5.64	6.18	15.82	10.31	14.20
	TS	1.42	3.72	3.16	4.02	4.42	13.38	7.00	10.41
	Beta	0.91	2.94	2.54	3.01	3.80	11.70	5.96	10.54
	GPcalib	1.02	3.42	3.45	3.92	4.53	13.42	7.75	12.45
	HIST	1.17	3.64	3.64	4.01	4.39	13.92	7.52	11.42
UDP	*LAS*	87.74	81.23	79.48	76.29	75.78	46.25	58.62	42.22
	l-ECE	2.03	5.99	5.25	8.09	6.54	15.11	14.69	17.02
	l-TS	1.07	3.25	3.02	4.64	3.16	6.74	7.97	7.75
	l-Beta	0.98	3.04	2.57	4.92	3.07	9.06	9.52	10.56
	l-GPcalib	0.73	3.23	2.65	4.79	2.95	8.77	9.31	10.92
	l-HIST	1.10	4.74	4.26	6.52	4.95	14.60	13.24	16.80
	h-ECE	5.70	5.92	8.74	9.37	9.35	22.56	14.26	24.42
	h-TS	1.19	2.72	2.87	2.56	2.64	8.75	4.15	8.55
	h-Beta	1.29	2.05	2.75	2.76	2.58	11.59	4.87	12.87
	h-GPcalib	1.29	2.90	2.81	2.59	2.71	8.26	4.41	6.82
	h-HIST	1.22	2.61	3.88	3.93	3.83	14.33	7.11	15.33
NER	*F-1*	86.99	79.84	78.38	78.56	68.02	70.11	58.42	40.23
	ECE	3.86	4.77	8.40	8.63	11.26	13.17	16.97	19.15
	TS	0.72	1.91	4.11	3.59	6.75	8.24	9.52	13.13
	Beta	0.68	1.69	3.59	3.10	6.62	7.59	9.81	13.71
	GPcalib	0.51	1.46	3.62	3.04	6.10	7.47	8.50	12.32
	HIST	1.53	2.50	4.31	4.35	7.33	8.70	11.69	14.72
XNLI	*Acc*	83.97	76.01	77.23	78.10	74.59	68.52	71.42	73.13
	ECE	10.83	17.20	17.08	15.58	18.71	22.74	20.12	18.16
	TS	3.98	7.52	7.82	6.15	8.60	11.74	9.06	7.38
	Beta	3.55	6.29	6.15	4.70	7.03	9.70	7.59	5.71
	GPcalib	3.59	6.38	6.35	5.08	7.44	10.23	7.66	5.93
	HIST	2.80	5.48	5.34	4.07	6.22	9.71	6.89	5.03

Table 8: Experiment result with `xlm-roberta-base` under the *full-data* setting. The color scheme is the same as figures above.

Task	Metric(%)	en	de	fr	es	ru	hi	ar	zh
POS	*Acc*	96.05	89.68	87.55	89.46	89.84	73.84	80.93	75.42
	ECE	3.49	8.55	9.77	8.98	8.41	19.74	14.50	15.88
	TS	2.27	4.07	5.29	5.09	3.76	11.09	7.64	5.88
	Beta	2.07	3.37	4.45	4.44	3.55	9.71	6.25	4.91
	GPcalib	1.20	2.29	4.51	3.65	2.82	11.04	5.83	8.16
	HIST	0.77	3.45	4.83	4.01	4.06	11.70	7.65	8.32
UDP	*LAS*	80.68	75.34	72.81	71.62	67.19	35.68	52.74	34.89
	l-ECE	3.88	7.71	8.61	11.18	10.22	21.92	21.45	25.98
	l-TS	3.31	5.20	5.77	7.71	7.01	12.25	13.49	16.21
	l-Beta	1.65	2.45	2.89	5.07	3.90	11.44	12.38	15.37
	l-GPcalib	1.65	1.50	2.75	3.29	2.67	6.80	8.47	11.04
	l-HIST	2.48	5.01	6.97	9.17	8.71	23.50	21.20	27.30
	h-ECE	12.69	13.46	15.26	14.95	18.29	40.68	21.83	39.58
	h-TS	3.71	6.86	7.51	6.94	5.33	11.58	7.01	10.52
	h-Beta	4.61	6.96	6.61	6.39	4.57	14.35	5.24	13.69
	h-GPcalib	3.99	7.39	8.25	7.06	5.99	10.13	7.96	8.10
	h-HIST	4.85	4.79	3.30	3.33	3.53	17.50	6.27	17.19
NER	*F-1*	80.42	76.31	77.43	78.28	66.52	69.54	69.92	39.04
	ECE	7.91	6.93	9.66	10.44	12.04	16.01	14.79	31.51
	TS	2.00	3.37	2.44	2.59	5.43	7.27	4.21	20.70
	Beta	1.53	2.45	1.98	1.91	4.47	6.59	3.83	20.48
	GPcalib	1.09	2.12	1.46	1.81	4.20	6.23	3.30	19.93
	HIST	1.56	3.39	2.13	2.51	4.79	6.94	4.99	21.57
XNLI	*Acc*	60.10	57.43	57.56	58.76	54.47	53.53	54.63	55.73
	ECE	30.40	32.16	32.57	31.58	35.92	35.57	34.45	33.57
	TS	4.17	5.10	5.25	4.71	7.72	7.10	7.10	6.02
	Beta	4.33	4.20	4.89	4.93	7.31	6.60	6.47	5.58
	GPcalib	4.30	4.41	4.62	4.08	7.56	6.45	6.18	5.44
	HIST	4.07	4.92	5.05	5.20	8.00	8.34	7.01	6.15

Table 9: Experiment result with `xlm-roberta-base` under the *low-data* setting. The color scheme is the same as figures above.

Task	Metric(%)	en	de	fr	es	ru	hi	ar	zh
POS	*Acc*	95.26	89.99	88.98	89.84	90.24	74.38	82.37	76.36
	ECE	3.55	7.24	7.28	7.48	6.62	14.29	11.64	11.84
	TS	2.68	3.23	4.81	3.92	3.30	9.20	5.85	6.54
	Beta	2.30	2.10	3.83	2.45	3.70	7.21	4.56	5.74
	GPcalib	1.73	1.63	2.65	2.43	2.38	6.58	3.40	4.69
	HIST	1.09	2.48	3.89	2.72	3.48	11.12	5.68	7.48
UDP	*LAS*	76.06	66.31	64.30	63.97	59.64	32.11	45.39	27.93
	l-ECE	4.62	8.11	9.71	12.01	10.76	23.95	20.90	25.38
	l-TS	4.55	5.09	5.43	7.35	6.99	12.04	12.22	13.69
	l-Beta	2.71	2.95	3.11	5.14	3.72	13.64	11.55	15.17
	l-GPcalib	2.80	2.06	3.00	3.75	2.96	8.99	7.67	10.31
	l-HIST	6.63	7.43	10.01	11.96	11.75	29.34	22.97	33.68
	h-ECE	10.98	15.14	15.20	14.68	19.24	37.07	24.39	36.01
	h-TS	2.61	3.06	5.17	5.19	5.29	17.11	7.62	14.46
	h-Beta	5.30	3.19	4.60	4.84	3.40	17.22	6.25	16.17
	h-GPcalib	7.59	8.62	11.49	10.37	7.08	7.84	6.98	4.61
	h-HIST	5.56	2.74	3.66	3.92	3.25	17.66	6.70	16.69
NER	*F-1*	70.09	69.02	67.81	67.07	55.29	64.94	53.13	30.55
	ECE	13.06	7.81	14.51	16.65	16.43	17.86	25.96	36.34
	TS	2.30	5.71	2.10	4.40	3.31	5.00	10.49	21.82
	Beta	2.14	5.18	2.11	4.45	2.77	5.06	10.80	21.70
	GPcalib	1.86	5.33	2.27	4.11	2.97	5.52	10.68	22.12
	HIST	1.85	5.67	2.93	5.13	4.26	5.99	12.80	22.74
XNLI	*Acc*	39.34	39.28	38.56	38.86	39.12	39.54	37.70	39.66
	ECE	58.11	57.91	58.69	58.17	58.09	56.93	59.33	57.34
	TS	2.92	3.40	3.23	3.20	3.22	2.54	4.00	2.76
	Beta	2.23	1.79	1.82	1.49	1.41	1.75	1.87	2.46
	GPcalib	2.66	2.31	2.24	2.04	1.95	2.63	2.34	1.99
	HIST	2.01	2.36	3.12	2.64	2.85	2.78	3.95	2.14

Table 10: Experiment result with `xlm-roberta-base` under the *very-low-data* setting. The color scheme is the same as figures above.

Task	Metric(%)	en	de	fr	es	ru	hi	ar	zh
POS	*Acc*	96.31	90.26	89.19	89.12	89.35	72.33	79.15	70.13
	ECE	2.86	7.47	7.45	7.89	8.02	18.31	14.78	21.84
	TS	1.88	4.57	3.86	4.46	4.59	12.26	9.11	14.91
	Beta	1.22	3.37	2.81	3.10	3.76	9.96	7.07	13.28
	GPcalib	0.99	3.05	2.51	2.83	3.65	9.64	7.08	13.22
	HIST	0.96	4.38	4.20	4.34	4.73	12.57	8.67	14.44
UDP	*LAS*	87.30	77.51	79.65	75.70	72.77	34.30	58.40	41.04
	l-ECE	2.37	7.16	5.54	8.61	7.66	19.22	14.97	18.92
	l-TS	1.31	3.46	2.67	4.74	3.59	9.52	7.41	11.02
	l-Beta	0.84	3.22	2.30	4.77	3.18	12.49	8.77	12.66
	l-GPcalib	0.78	3.41	2.23	4.76	3.45	11.64	8.65	12.26
	l-HIST	1.57	6.63	5.18	8.25	7.66	24.50	17.37	23.27
	h-ECE	6.25	7.34	9.10	10.10	11.46	31.07	14.63	29.57
	h-TS	1.56	3.34	2.57	2.38	2.28	12.51	3.46	13.68
	h-Beta	1.53	2.66	2.37	2.58	2.17	16.72	3.22	16.41
	h-GPcalib	1.57	3.76	2.61	2.30	2.07	11.24	3.51	12.56
	h-HIST	1.56	2.95	3.47	3.37	3.32	17.99	5.28	17.75
NER	*F-1*	87.71	85.16	79.88	80.88	71.68	75.19	57.67	56.46
	ECE	3.95	3.07	8.80	8.19	9.06	9.72	20.08	17.72
	TS	1.14	1.10	5.09	3.75	4.54	4.46	12.53	11.44
	Beta	0.93	0.82	4.60	3.47	4.24	4.41	12.23	11.06
	GPcalib	0.91	0.95	4.73	3.49	4.28	4.57	12.67	11.41
	HIST	1.21	1.35	5.08	4.28	4.79	5.01	13.01	12.23
XNLI	*Acc*	81.90	70.24	73.61	73.73	67.03	59.42	64.21	68.84
	ECE	10.90	18.51	17.29	16.68	22.75	27.82	22.82	20.40
	TS	3.20	7.53	7.05	6.18	11.36	15.74	10.66	9.02
	Beta	2.85	6.32	5.82	4.88	10.02	14.42	9.46	7.91
	GPcalib	3.46	6.01	5.86	4.80	9.87	14.39	9.32	7.72
	HIST	3.59	6.41	5.70	4.99	9.86	14.59	9.28	7.79

Table 11: Experiment result with `bert-base-multilingual-cased` under the *full-data* setting. The color scheme is the same as figures above.

Task	Metric(%)	en	de	fr	es	ru	hi	ar	zh
POS	*Acc*	95.52	89.55	88.22	88.57	87.43	69.59	78.09	69.44
	ECE	3.50	7.90	9.27	8.95	8.95	19.42	14.45	20.13
	TS	2.29	4.12	5.02	5.84	3.93	13.72	7.96	12.45
	Beta	1.85	2.44	3.58	4.32	2.54	10.04	5.32	10.77
	GPcalib	1.45	2.22	3.75	4.11	2.28	9.30	5.22	10.65
	HIST	0.59	2.93	4.45	4.23	4.00	12.13	6.42	12.32
UDP	*LAS*	81.61	71.18	72.07	70.90	65.23	25.92	50.87	36.01
	l-ECE	3.78	8.34	7.87	10.99	11.28	26.63	20.55	22.99
	l-TS	3.68	5.06	4.90	7.51	7.44	14.84	11.58	13.93
	l-Beta	2.14	2.18	3.57	4.23	3.40	14.56	9.68	12.02
	l-GPcalib	2.22	2.53	4.56	3.31	2.74	9.90	6.54	8.51
	l-HIST	3.43	6.56	7.09	9.21	10.17	31.69	21.06	26.81
	h-ECE	11.16	13.98	14.73	14.39	17.39	44.02	21.94	38.38
	h-TS	5.24	10.22	12.40	10.17	7.70	9.53	8.64	9.68
	h-Beta	6.55	9.54	9.86	8.74	6.49	15.95	6.06	12.28
	h-GPcalib	6.23	13.03	14.97	11.79	10.03	5.29	10.79	8.00
	h-HIST	6.53	7.01	6.17	5.39	4.13	18.28	6.68	15.03
NER	*F-1*	83.09	83.26	82.10	82.19	65.62	71.29	58.79	57.56
	ECE	7.69	4.61	8.69	8.95	13.52	14.18	22.22	18.80
	TS	2.66	4.02	2.97	3.71	5.30	5.34	9.48	8.93
	Beta	2.29	2.74	2.33	2.91	4.96	5.28	10.34	9.24
	GPcalib	2.03	2.42	2.34	3.03	4.46	4.83	9.50	8.93
	HIST	1.04	3.29	1.38	2.10	5.38	6.12	11.42	10.18
XNLI	*Acc*	59.36	55.91	56.05	55.83	54.65	52.85	54.59	54.23
	ECE	26.37	27.20	27.72	28.04	26.66	26.95	26.99	27.54
	TS	6.78	7.44	7.43	6.70	5.64	4.51	5.32	6.43
	Beta	5.80	6.11	6.28	5.75	4.68	3.93	4.18	5.52
	GPcalib	6.58	6.74	6.70	6.14	5.52	4.80	4.67	5.99
	HIST	5.23	5.64	5.42	4.55	5.80	5.53	4.96	5.45

Table 12: Experiment result with `bert-base-multilingual-cased` under the *low-data* setting. The color scheme is the same as figures above.

Task	**Metric(%)**	**en**	**de**	**fr**	**es**	**ru**	**hi**	**ar**	**zh**
POS	*Acc*	94.49	91.40	89.16	90.17	88.44	74.31	78.68	69.05
	ECE	4.13	5.70	7.65	6.93	7.54	16.03	13.07	20.31
	TS	3.16	2.83	3.07	4.02	2.18	11.75	5.98	10.90
	Beta	2.45	2.62	2.59	2.66	2.40	9.57	5.83	9.47
	GPcalib	1.87	3.23	2.09	2.37	2.66	7.53	3.91	8.25
	HIST	1.93	2.40	1.86	1.88	3.00	9.52	5.81	10.92
UDP	*LAS*	76.99	60.45	66.59	64.60	56.21	22.45	41.67	30.93
	l-ECE	4.76	10.30	9.68	12.74	14.05	26.13	22.90	23.64
	l-TS	4.63	5.25	5.55	7.62	8.19	8.74	11.02	10.25
	l-Beta	2.62	3.14	2.94	5.38	5.23	14.45	12.05	12.70
	l-GPcalib	3.02	2.12	2.79	3.75	3.29	9.76	8.84	9.14
	l-HIST	6.87	10.60	9.62	12.11	15.44	35.66	27.01	30.90
	h-ECE	11.80	19.07	15.30	15.18	19.72	43.00	25.54	38.62
	h-TS	8.14	7.76	11.40	11.17	7.50	9.21	6.51	8.86
	h-Beta	5.73	2.86	5.22	5.50	2.36	20.42	5.23	17.14
	h-GPcalib	7.16	8.03	10.94	10.77	8.28	7.26	7.13	7.29
	h-HIST	5.61	2.40	3.82	3.81	2.93	21.62	6.31	18.20
NER	*F-1*	72.56	73.71	71.47	70.56	50.96	62.88	54.51	41.96
	ECE	11.00	5.23	12.18	15.33	17.79	18.86	25.29	33.03
	TS	2.63	7.07	2.48	3.58	5.90	4.65	9.29	19.58
	Beta	2.57	6.36	2.50	3.73	5.65	5.06	10.24	19.53
	GPcalib	2.36	5.98	2.72	4.18	6.07	5.97	10.51	19.53
	HIST	1.53	6.02	3.16	4.21	6.36	5.63	11.61	20.43
XNLI	*Acc*	45.51	43.81	44.85	45.53	44.87	41.58	43.93	45.79
	ECE	45.87	45.91	45.04	44.66	44.37	47.20	45.54	43.91
	TS	5.40	4.92	4.11	6.01	4.18	5.32	4.88	4.86
	Beta	2.90	2.36	2.88	3.31	2.75	2.43	2.35	2.77
	GPcalib	4.72	3.56	4.31	3.88	3.71	4.29	3.82	3.42
	HIST	3.45	4.38	4.12	3.67	3.85	7.34	4.89	4.64

Table 13: Experiment result with `bert-base-multilingual-cased` under the *very-low-data* setting. The color scheme is the same as figures above.

Source		**en**	**de**	**fr**	**es**	**ru**	**hi**	**ar**	**zh**
full									
	ori	3.07	6.81	7.52	6.63	5.80	13.06	10.36	34.08
	TS	1.58	4.25	3.96	3.80	3.00	8.44	5.89	25.41
	GPcalib	1.17	3.79	4.54	3.48	2.47	7.49	4.84	27.89
-crf	ori	3.32	7.38	8.06	7.10	6.72	14.14	12.02	28.79
	TS	2.11	5.19	4.94	4.80	4.36	9.75	7.59	17.71
	GPcalib	1.63	4.09	3.47	3.84	3.36	9.43	6.83	9.74
low-data									
	ori	3.29	6.89	8.82	6.65	6.78	14.37	10.66	25.84
	TS	1.79	2.88	3.37	3.18	2.64	8.21	5.83	12.14
	GPcalib	1.01	2.03	3.78	1.75	1.52	5.82	3.00	15.83
-crf	ori	3.31	6.64	8.45	6.37	6.68	14.04	11.59	39.26
	TS	1.66	2.75	3.46	3.26	2.11	8.02	4.46	27.48
	GPcalib	1.42	5.39	7.25	6.44	2.72	10.25	6.04	21.25
very-low-data									
	ori	3.65	5.88	8.50	4.93	6.07	11.91	10.94	35.32
	TS	1.94	2.56	4.80	1.82	2.73	7.82	5.53	28.19
	GPcalib	1.66	2.06	4.59	1.46	2.04	7.41	5.88	30.15
-crf	ori	4.23	7.17	9.22	6.30	7.42	13.45	13.34	43.21
	TS	2.11	2.66	4.26	1.88	2.88	7.63	5.66	36.80
	GPcalib	1.57	2.33	2.85	1.58	2.59	7.76	3.91	34.24

Table 14: structured prediction experiments: POS, comparing different calibration methods with statistical significance tests. Blue shaded cells indicate significantly better performance in calibration by GPcalib vs TS decided by a bootstrap from the dataset and an independent t-test with $p < .05$. Brown shaded cells indicate significantly worse performance in calibration by GPcalib vs TS decided by the same criteria.

Source		en	de	fr	es	ru	hi	ar	zh
full									
	ori	5.49	5.55	6.70	8.68	7.84	9.59	14.62	9.03
	TS	4.66	4.74	5.80	7.70	7.08	8.53	13.75	8.30
	GPcalib	1.47	1.26	2.65	4.52	3.61	4.68	8.83	3.97
-crf	ori	5.94	5.50	6.55	7.49	9.41	12.62	19.14	18.25
	TS	5.99	5.09	6.48	7.11	8.30	11.38	17.92	17.69
	GPcalib	2.67	2.59	2.16	2.69	3.36	7.07	12.40	10.57
low-data									
	ori	9.14	6.99	8.28	8.01	12.57	14.78	18.48	20.49
	TS	7.85	6.48	8.13	7.28	11.55	14.42	17.53	20.21
	GPcalib	3.20	3.04	2.91	3.82	4.76	7.63	9.91	10.37
-crf	ori	8.59	7.26	8.01	7.58	12.28	13.18	18.56	20.48
	TS	7.55	6.45	7.31	6.77	11.17	12.68	17.60	20.26
	GPcalib	2.40	1.90	2.46	2.89	3.84	6.73	10.35	12.00
very-low-data									
	ori	14.67	11.09	14.41	15.96	14.72	19.43	20.58	18.25
	TS	10.14	7.08	9.65	10.69	8.88	13.26	12.65	12.82
	GPcalib	2.82	3.63	3.27	4.58	3.57	7.17	6.09	5.18
-crf	ori	15.44	13.63	14.50	15.63	21.42	21.08	24.23	20.34
	TS	6.71	4.29	7.63	8.20	6.10	7.61	7.62	11.07
	GPcalib	2.24	3.71	2.90	4.08	5.93	6.56	8.13	5.89

Table 15: structured prediction experiments: NER, comparing different calibration methods with statistical significance tests. Color scheme same as the one in the previous table.