# Statement-Tuning Enables Efficient Cross-lingual Generalization in Encoder-only Models

**Ahmed Elshabrawy**[1]   **Thanh-Nhi Nguyen** [*2, 3]   **Yeeun Kang** [*4]   **Lihan Feng** [*5]
**Annant Jain** [*6]   **Faadil A. Shaikh** [*]   **Jonibek Mansurov** [1]   **Mohamed Fazli Imam** [1]
**Jesus-German Ortiz-Barajas** [1]   **Rendi Chevi** [1]   **Alham Fikri Aji**[1]

[1] MBZUAI    [2] UIT, Vietnam    [3] VNU-HCM
[4] Yale University    [5] NYU Shanghai    [6] IIT Bombay

ahmed.elshabrawy@mbzuai.ac.ae

## Abstract

Large Language Models (LLMs) excel in zero-shot and few-shot tasks, but achieving similar performance with encoder-only models like BERT and RoBERTa has been challenging due to their architecture. However, encoders offer advantages such as lower computational and memory costs. Recent work adapts them for zero-shot generalization using Statement Tuning, which reformulates tasks into finite templates. We extend this approach to multilingual NLP, exploring whether encoders can achieve zero-shot cross-lingual generalization and serve as efficient alternatives to memory-intensive LLMs for low-resource languages. Our results show that state-of-the-art encoder models generalize well across languages, rivaling multilingual LLMs while being more efficient. We also analyze multilingual Statement Tuning dataset design, efficiency gains, and language-specific generalization, contributing to more inclusive and resource-efficient NLP models. We release our code and models[1].

## 1 Introduction

Large Language Models (LLMs) have shown great capabilities in zero-shot and few-shot settings (Radford et al., 2019; Brown et al., 2020; Artetxe et al., 2022). However, these capabilities are often more difficult to observe in encoder-only models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) due to their architectural design. These models are typically pretrained with a Masked Language Modeling (MLM) objective and are fine-tuned by adding task-specific layers to enable their usage on a downstream task. These task-specific layers block the extension of these models to new tasks in a few-shot or zero-shot manner.

Despite these difficulties, applying encoder models for zero-shot task generalization offers several
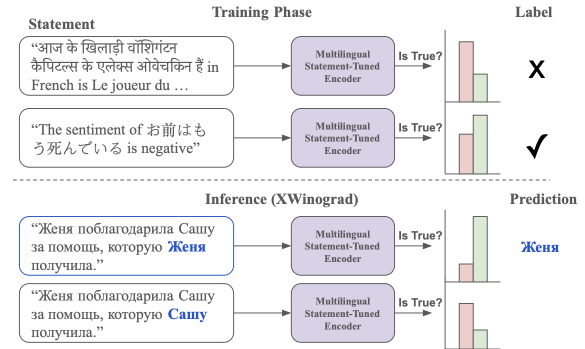
---



Figure 1: Encoder models trained with Multilingual Statement-Tuning can generalize across new task and unseen languages during finetuning.

advantages. First, encoder models are typically more lightweight than LLMs and therefore require less computational power and memory. Further, encoder models excel at generating contextual embeddings that better capture semantic meaning. For instance, recent work (Qorib et al., 2024) demonstrated that decoder-only LLMs perform worse on word meaning comprehension than encoder-only models. Encoder-only models are also more efficient in inference for tasks such as sequence labeling due to their architecture (Soltan et al., 2023).

To enable encoder model usage in zero-shot task generalization, Elshabrawy et al. (2025) introduced Statement-Tuning. This technique converts tasks into a set of statements with finite templates, training on an encoder-only model, RoBERTa, to discriminate between potential statements and derive final results. This method demonstrates the feasibility of using encoder models, typically specialized for specific tasks, to handle various unseen Natural Language Understanding (NLU) tasks, similar to zero-shot prompting in decoder models. It shows competitive performance compared to large language models (LLMs) with significantly fewer parameters and training data, highlighting its potential for zero-shot learning. However, the original

---

16226

approach focused only on English, raising questions about its applicability in multilingual settings and its ability to generalize to new tasks and languages.

In this work, we aim to explore the adaptation of Statement-Tuning for multilingual NLP tasks. Specifically, we investigate whether encoder models can achieve zero-shot cross-lingual generalization similar to decoder-only LLMs (Muennighoff et al., 2023). Given the multilingual setting, it is crucial to emphasize the importance of low-compute solutions. Speakers and users of low-resource languages often lack the computational resources necessary to utilize memory-intensive LLMs. Therefore, our method, which leverages efficient encoder models, is particularly important as it offers a more accessible and inclusive approach to zero-shot text classification in these contexts (Ruder, 2022).

Hence our contributions are as follows:

- We enable zero-shot generalization to unseen tasks and languages for encoder-only models, a capability typically limited to decoder-based models.

- Our benchmarking shows that state-of-the-art multilingual encoder-only models match LLMs in performance while being more efficient.

- We analyze multilingual Statement-Tuning dataset design, including language diversity and translated prompt templates.

- We investigate when and how multilingual Statement-Tuning generalizes effectively across languages.

- We compare Statement-Tuning inference speed and memory efficiency against generative models, showing significant advantages.

## 2 Related Work

**Zero-shot and Few-shot Approaches Using Encoder-Only Models** There have been several works exploring prompt-based approaches to enable few-shot and zero-shot generalization in Encoder-only Models. Finetuning on cloze templates or label discrimination (Schick and Schütze, 2021; Gao et al., 2021) effectively utilizes encoder models for few-shot learning. Cloze templates have also been shown to fare better than regular finetuning in cross-lingual few-shot transfer to unseen

languages from higher resourced languages (Zhao and Schütze, 2021; Tu et al., 2022; Ma et al., 2023, 2024).

However, to enable zero-shot task learning a reformulation of text classification tasks is necessary. Yin et al. (2019) introduce the reformulation of any zero-shot text classification in the form of entailment where statements can be formed from a series of choices and the correct choice is seen to be an entailment. Xu et al. (2023) use DeBERTa to show that the entailment formulation of zero-shot classification can be observed to be more effective than the generative approach employed by LLMs.

Finally, Elshabrawy et al. (2025) propose Statement-Tuning to show that through template-based data augmentation, much smaller RoBERTa models can be finetuned on limited data to match or even exceed the zero-shot NLU capabilities of several LLMs of up to 70B parameters on monolingual classification tasks. While Elshabrawy et al. (2025) focused only on English, we studied whether Statement-Tuning method is possible in other languages. Additionally, we explore the efficiency of the approach in more detail and offer insight on the effect of pretraining data on the performance of Statement-Tuning.

**Zero-shot Prompting and Multitask Tuning** While LLMs were shown to perform well on few-shot generalization (Brown et al., 2020), they showed less successful performance on zero-shot generalization. To tackle this issue, instruction tuning was proposed. Instruction tuning refers to fine-tuning language models on a collection of datasets described via instructions (Wei et al., 2022). Their model, Finetuned Language Net (FLAN), a decoder-only model of 137B parameters fine-tuned on more than 60 NLP datasets expressed via natural language instructions, proved effective in improving the zero-shot performance of models.

They also showed that increasing the number of tasks involved in instruction tuning improves unseen task generalization performance and asserted that the benefits of instruction tuning are emerging abilities of language models (i.e., they emerge with sufficient scale). Subsequent work by Sanh et al. (2022) explored instruction tuning with T5 encoder-decoder models and proposed the T0 models and datasets. They fine-tuned T5 models of 3B and 11B parameters, which were smaller than the FLAN model but still within the billions-of-parameters range. Their findings established that

with a more diverse prompt setup and an encoder-decoder model like T5, language models could achieve good performance with instruction tuning.

Chung et al. (2022) found that instruction tuning is effective across a variety of model classes, such as PaLM, T5, and U-PaLM, as well as different prompting setups including zero-shot, few-shot, and chain-of-thought. Their models, FLAN-T5, ranged from 80M to 11B parameters and showed better performance than prior T5 checkpoints. Meanwhile, Mishra et al. (2022); Wang et al. (2022); Honovich et al. (2023); Wang et al. (2023) also proposed large-scale natural language instruction datasets.

These methods fine-tune large models on constructed datasets with various task prompts, achieving strong zero-shot results. However, effective instruction-tuned models often require billions of parameters (Zhang et al., 2023b), limiting their application to smaller models. Ye et al. (2022) aim to distill this zero-shot ability in a smaller model like an LSTM through synthetic data creation using an LLM, but they create task-specific models rather than a single smaller model that is capable of generalizing.

Our work demonstrates achieving similar or superior generalization of LLMs using a single smaller MLM with less training data. Furthermore, our work expands on previous efforts by exploring encoder models, which contributes to parallel understanding when combined with works on decoder models (Wei et al., 2022) and encoder-decoder models (Sanh et al., 2022).

## 3 Method: Multilingual Statement-Tuning

In this section, we outline the steps involved in Statement-Tuning.

### 3.1 Multilingual Task Verbalization

First, using templates as shown in Figure 2, tasks are verbalized in natural language statements. We then train the statement discriminator to classify these statements as true or false.

```
"{{target_word}}" means the
same in "{{context_1}}" and
      "{{context_2}}"
```

Figure 2: Example of a statement template used during task verbalization for sentiment analysis.

As Elshabrawy et al. (2025) propose, any discriminative task with a finite set of targets can be verbalized into a finite set of natural language statements, one for each label. Similar to prompting, each task has its own statement templates (outlined in Appendix A). The truth label for training purposes on each statement depends on whether the statement contains the correct target label or not.

### 3.2 Statement Fine-Tuning Setup

To create the training dataset for statement fine-tuning, we exhaustively generate statements across 9 NLU tasks using many varied statement templates per dataset. For a detailed breakdown of the datasets used and what tasks they cover, refer to Appendix B.

The rule for task selection generally follows the structure in Elshabrawy et al. (2025), except for adding the machine-translation task. For each task, we randomly choose 1500 rows of training data for each language, with a balance of labels (controlling for positive and negative examples i.e. 750 examples for each label). This ensures the encoder models have sufficient data to train and explore their potential to be the generalizers. For selected low-resource languages, the total amount of data may be less than 1500; in that case, we choose all of the specific data for training. Data from the machine-translation task is added to enhance the generality of low-resource language in the tasks lacking data and models' cross-lingual ability. The rest of the tasks are selected either because they are often addressed by using LLMs or because we hypothesize they may enhance models' language understanding.

Our compilation of multilingual datasets amounts to 25 languages, both high- and low-resource. We include the full list of languages and additional language-specific information in the Appendix C.

We explore the different number of languages including in the dataset and the language of the statement templates. We fine-tune different multilingual encoder-only models, mBERT (Devlin et al., 2018), mDeBERTa (He et al., 2021), XLM-R base and large (Conneau et al., 2020) with a binary sequence classification head to predict the truth value of the statements. By fine-tuning the model across diverse tasks, languages, and templates, we hypothesize that the model should be able to generalize across unseen templates, unseen tasks, as

| Model | Parameters | XCOPA | XNLI | XStoryCloze | XWinoGrad |
|---|---|---|---|---|---|
| Qwen2 | 72B | 67.84 | 42.10 | 66.70 | 84.02 |
| Llama3.1 | 70B | 62.24 | 41.68 | 68.32 | 82.69 |
| Gemma 2 | 9B | 66.29 | 46.50 | 67.41 | 83.93 |
| Llama3.1 | 8B | 60.29 | 44.39 | 61.60 | 80.49 |
| Aya 23 | 8B | 54.60 | 42.44 | 60.36 | 69.36 |
| Aya 23 | 35B | 57.24 | 44.09 | 63.65 | 72.69 |
| Gemma 2 | 27B | 68.65 | 45.41 | 69.76 | 85.26 |
| Gemma 2 | 2B | 53.15 | 34.08 | 50.76 | 59.27 |
| Qwen2 | 1.5B | 53.44 | 34.73 | 51.87 | 66.94 |
| Qwen2 | 500M | 53.13 | 33.58 | 50.05 | 58.08 |
| mBERT(base) | 110M | 52.47 | 34.51 | 48.30 | 50.68 |
| XLMR-base | 250M | 56.69 | 35.33 | 60.71 | 51.34 |
| XLMR-large | 560M | 64.36 | 45.76 | 78.78 | 54.26 |
| **mDeBERTa (Best)** | 276M | $65.52_{(1.64)}$ | $47.84_{(1.65)}$ | $73.53_{(1.25)}$ | $54.75_{(1.24)}$ |

Table 1: **Accuracy of the multilingual decoder and encoder models finetuned on the same data mixture.** on XCOPA, XNLI, XStoryCloze, and XWinoGrad tasks. Results in grey highlight performances that are below the best-performing encoder model, mDeBERTa (276M). Additionally, we report the average standard deviation across languages over 3 training runs only for mDeBERTa to quantify the random deviation due to Statement-Tuning training.

well as unseen languages, as long as it can be transferred into a True/False statement, and the "unseen" languages are at least seen during the pre-training stage. Additionally, for mDeBERTa trained with an 11-language Statement-Tuning dataset, we report the average and standard deviation over 3 different training runs, to account for randomness and report it as such where appropriate. We were unable to perform this over all ablations due to the scale of the experimentation and limited computational and time resources.

## 3.3 Zero-Shot Inference

To perform inference on statement-tuned models, we convert testing tasks into declarative statements. Specifically, we generate a statement for each possible label and predict its probability of being true. The final label for a given task is the statement with the highest probability. To ensure robustness across different phrasings, we experiment with various templates for each task during both training and evaluation.

## 4 Experimental Setup

### 4.1 Models

We experiment with 4 multilingual encoder models of different sizes and multilingual capabilities, namely mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mDeBERTa-v3 (He et al., 2021), and XLM-V (Liang et al., 2023). We report
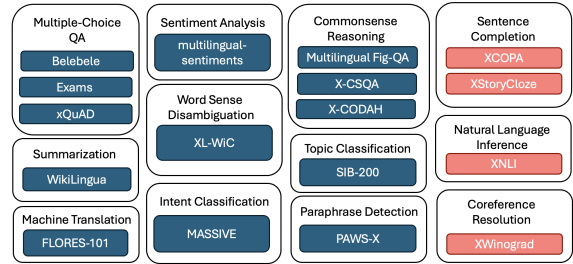


Figure 3: NLU datasets and their respective tasks used for multilingual statement fine-tuning.

the number of parameters and pre-training corpora in Table 2.

| Model | # Params | Pretraining Corpus |
|---|---|---|
| mBERT base | 110M | Wikipedia |
| mDeBERTa-v3 | 276M | CC-100 (Wenzek et al., 2020) |
| XLM-R large | 560M | |

Table 2: Multilingual encoder models with their parameter sizes and pretraining corpora.

### 4.2 Baselines

To assess the capabilities of encoder models with Multilingual Statement-Tuning, we compare them with several state-of-the-art instruction-tuned multilingual generative models (except Gemma 2 which is not specifically pretrained for multilingually but has some multilingual support) ranging from 500 million to 72 billion parameters in size. We

use models from the following language families: Qwen2 (Yang et al., 2024), Llama3.1 (Dubey et al., 2024), Gemma 2 (Rivière et al., 2024), and Aya 23 (Aryabumi et al., 2024).

We make our comparison through two forms. First, we finetune the base models on an instruction dataset we created using the same datasets and languages used for Multilingual Statement-Tuning (for details see Appendix D). The second setup involves using the instruction-tuned varieties released by the original team. We include the first setup as a fair comparison controlling for the same fine-tuning data and the second to gauge the performance against the publicly released instruction-tuned versions which employ more data and techniques as another strong baseline for performance.

We make use of 4 unseen multilingual NLU benchmarks in our analysis, XCOPA (commonsense reasoning) (Ponti et al., 2020), XNLI (sentence understanding) (Conneau et al., 2018), XStoryCloze (Lin et al., 2022), and XWinograd (Muennighoff et al., 2023; Tikhonov and Ryabinin, 2021). Some of the languages in these benchmarks are unseen by our Multilingual Statement-Tuning models, demonstrating cross-lingual generalization. For most of the analysis, we report averages across all the languages included in the evaluation datasets, more detailed figures with individual languages are included in Appendix E.

Although we have tried to select models and evaluation data to minimize the chance of leakage of the evaluation and (pre)training data, some (generative) models' pretraining is not completely open, and hence, this remains a limitation of our analysis that is difficult to control. As shown in Table 2, the pretraining of all encoder models is open and, to our knowledge, does not include the evaluation data. For all generative models, we employ the prompting templates provided by the Language Model Evaluation Harness (Gao et al., 2024).

### 4.3 Ablations

As part of our analysis, we ablate several design choices of an encoder-only cross-lingual generalization system. We experiment with encoder models of sizes ranging from 110 million parameters to 560 million (models are outlined in Section 3.2).

We explore several design choices for statement generation. First, we use a multilingual prompt template as opposed to just using an English template. To achieve this we machine translate the

English template to the language of the example using ChatGPT, specifically the GPT-3.5 version (OpenAI, 2023).

Furthermore, we are interested in the effect on cross-lingual generalization when more languages are used during the Statement-tuning step so we explore 3 linguistic setups: English-only (with and without machine translation in the task mixture), 11 languages, and 25 languages to be used during Statement Tuning (languages used for each setup are outlined in Appendix F).

Finally, we directly explore the effect of machine translation data in the Statement-Tuning training data mixture. For the rest of the design choices, such as the number of statements to use per dataset and number of templates to use, we follow the general guidelines recommended by Elshabrawy et al. (2025). Furthermore, we explore the advantages of using encoder models over generative models from an efficiency perspective by exploring the inference time of encoder models against generative models in Section 5.7.

## 5 Results and Analysis

In this section, we derive insights from our experimental results about the cross-lingual zero-shot generalization capabilities of encoder models.

### 5.1 Encoder Models are Cross-Task Generalizers

In Table 1, we report the average (over languages) unseen task performance of models trained with Multilingual Statement-Tuning in 11 languages. We contrast this with several instruction-tuned multilingual decoder models, ranging from 500 million to 72 billion parameters, which were instruction-tuned with the same data as our Multilingual Statement-Tuning models. The individual language performance is shown in Appendix E.

The results show that multilingual encoder models are capable of zero-shot cross-task task generalization over a variety of unseen commonsense reasoning and natural language understanding tasks. For XNLI and XStoryCloze, the best-performing encoder models (mDeBERTa and XLM-R Large) outperform most of the generative models examined. More impressively, XLM-R large has an average accuracy of 78.8 on XStoryCloze outperforming the best-performing LLM, Llama3.1 70B, by **10.5** points despite having ∼**130** times fewer parameters.

On XNLI the gap is not as large but quite impressively mDeBERTa is the best-performing model at only 276 million parameters outperforming both Qwen2 72B and Llama3.1 70B by around **5.7** and **6.1** points on average. For XCOPA, the same best-performing encoder models still maintain impressive results outperforming all the generative models of under 9 billion parameters, and outperforming one of the 70B+ parameter models (Llama3.1 70B).

In Appendix G, we perform the same analysis but on the instruction-tuned varieties of the models released by the teams who trained them. We largely draw the same conclusions with slight variations where certain models on certain tasks perform slightly better/worse.

## 5.2 Encoder Models are Cross-Lingual Generalizers

When examining individual language performance (see Appendix E for more details) we note that mDeBERTa had less variation **across** languages in the same task (i.e. there is less disparity between higher and lower resource languages) when compared with generative models. Moreover, mDeBERTa was able to generalize on **unseen** tasks on languages **completely unseen** during Multilingual Statement-Tuning if they were seen during pretraining. This further supports the use of state-of-the-art encoder-only models as alternatives to generative models for low-resource languages and cross-lingual generalization on NLU tasks.

Interestingly, mBERT and XLM-R base do not exhibit such performance; at first, it may seem to be an issue of size; however, mDeBERTa has a parameter size similar to XLM-R base, but significantly outperforms it. Hence, we believe that such generalization capabilities require effective pretraining.

Nevertheless, all encoder models fail to generalize on the XWinograd benchmark, a coreference resolution dataset, achieving mostly random baseline performance. We attribute this to task selection during the Multilingual Statement-Tuning stage, as most of the datasets used may not have sufficient relevance with coreference resolution tasks. The exception might be XL-WiC, which involves word sense disambiguation.

This aligns with the findings of Elshabrawy et al. (2025), who noted that dataset relevance significantly impact a model's ability to generalize effectively.
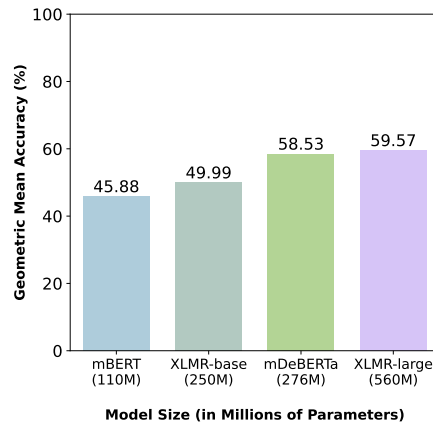


Figure 4: Geometric mean accuracy of multilingual encoder models (mBERT, XLMR-base, mDeBERTa, XLMR-large) across tasks.

## 5.3 Pretraining and Size Enable Cross-lingual Generalization Abilities

Interested in generalizing our finding across multiple encoder-only models, we train 4 different models (mBERT, XLMR-base, mDeBERTa, XLMR-large) using Multilingual Statement-Tuning using the exact same data setup and varying certain hyperparamters depending on model (see Appendix D) until convergence. We evaluate them on the same four unseen benchmark datasets outlined in Section 4.2.

In Figure 4, we compare the geometric mean of the task performance of the four multilingual encoder-only models we examine with Statement-Tuning, as discussed earlier we note that the two models mDeBERTa and XLM-R Large exhibit much higher task performance than the other two models mBERT and XLM-R base. Despite finetuning all the models until convergence and performing hyperparameter optimization with all models, this remains the case. Previous work has shown the relatively limited capabilities of mBERT in comparison to other models (Conneau et al., 2020) which has different pretraining data and regimes. However, the difference in abilities between XLM-R base and XLM-R large cannot be attributed to just pretraining, as XLM-R base fails to achieve zero-shot cross-lingual generalization with the same pretraining choices as XLM-R large at a different scale.

It is also difficult to attribute cross-lingual generalization capabilities purely to model size as mDeBERTa achieves similar performance and cross-lingual generalization capabilities as XLM-R large
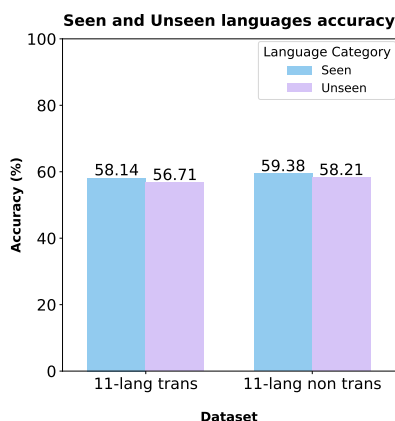
Figure 5: Mean accuracy of mDeBERTa on seen vs. unseen languages during Statement-Tuning across tasks and languages. 11-lang trans uses machine-translated prompt templates while 11-lang non trans shows performance using English-only prompt templates.
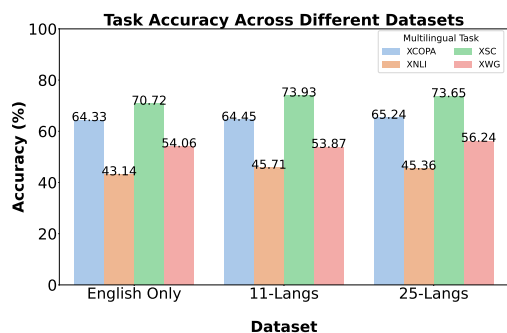


Figure 6: Task accuracy across different training datasets (English-only, 11-langs, and 25-langs) using mDeBERTa.

but at a size comparable to XLM-R base. Hence, we hypothesize that cross-lingual zero-shot generalization (being an inherently difficult task) emerges in encoder-only as a function of both size and pretraining. In general, encoder models that have shown state-of-the-art performance on general tasks are more likely to exhibit cross-lingual generalization capabilities but it is not strictly a matter of model "capacity" as would be implied by model size, or pretraining data.

### 5.4 English-only Prompting Templates are Sufficient to Enable Effective Statement-Tuning

As part of our investigation on assessing the cross-lingual zero-shot generalization capabilities of encoder-only models, we experiment to see if the use of machine-translated prompt templates
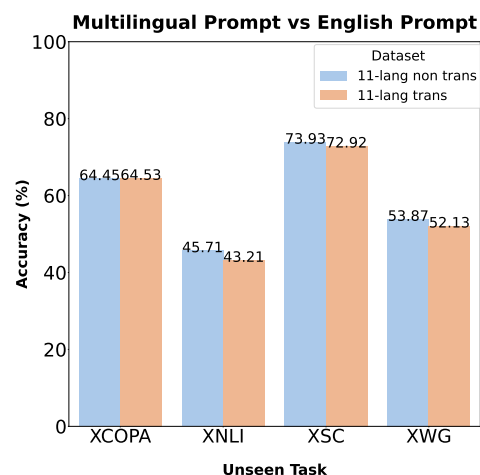


Figure 7: Mean task accuracy of mDeBERTa (finetuned on 11 languages during Statement-Tuning) over languages on the 4 evaluation tasks using machine translated prompt templates (11-lang trans) vs. English-only prompt templates (11-lang non trans).

would offer any improved performance over the use of English-only prompt templates for the various tasks. To achieve this, we utilized ChatGPT, specifically the GPT-3.5 version (OpenAI, 2023), to machine-translate each prompt to match the language of the example being turned into a statement. We then train and evaluate a model using these machine-translated examples to assess the impact on performance across different languages. An obvious limitation of this analysis could be the quality of the Machine Translation, however, we decided to use a commonly accessible model.

As seen in Figure 7, we do not observe any apparent benefit to translating prompts. This is consistent with observations from multilingual prompting of LLMs (Zhang et al., 2023a).

Curious to see if perhaps machine translating prompt templates helps benefit generalization capabilities we compare seen versus unseen (during Statement-Tuning) average language performance in Figure 5. Again, we observe no apparent benefit of unseen languages. This observation has the added benefit of simplifying prompt design and reducing the computational/time cost of having to machine translate prompts.

### 5.5 Multilingual Pretraining Sufficiently Enables Cross-lingual Generalization

To understand the effect of Multilinguality during Statement-Tuning on cross-lingual zero-shot task generalization we experiment by changing the number of languages included in the Statement-

Tuning data by using the same training tasks with a differing number of languages being included in the training set.

In Figure 6, we examine the effect of including more languages in the training set on cross-lingual capabilities. We examine 3 setups, English-only where the mDeBERTa model is only trained on the English subsets/equivalents of the training datasets (except for the machine translation task which **is** included), 11-langs where the model is trained on subsets of the data that belong to only 11 of the possible 25 languages and 25-langs which includes all possible languages in the training datasets. By sampling, we fix for training set sizes to be similar in size with the English-only and 11-langs to include 123k training examples and the 25-langs dataset including ∼185k examples (it needed to be slightly larger to representatively sample the languages).

Overall, we observe that most of the cross-lingual task performance can actually be obtained by finetuning a multilingual encoder model on a single language (English) multi-task statement dataset, with the English-only setup achieving **98.6%**, **95.1%**, and **96.0%** of the performance of 25-langs on XCOPA, XNLI, and XStoryCloze respectively. Increasing the number of represented languages to 11 during Statement-Tuning yields gains over English-only and manages to very slightly outperform using all languages on XNLI and XStoryCloze.

Furthermore, in Figure 5, we compare the average performance of the 11-lang model on seen versus unseen languages during Statement-Tuning. We only observe a slight performance gain on average (59.4 versus 58.2) when languages are seen during Statement-Tuning. This leads us to believe that most of the cross-lingual generalization capabilities are impressively due to the multilingual pre-training, rather than requiring cross-lingual exposure during Statement-Tuning. This opens up many potential use cases of encoder-only models for zero-shot task generalization for use with languages without necessitating any supervised Statement-Tuning in these languages which proves very useful given the abundance of task data in high-resource languages such as English.

Nevertheless, we report the performance of 11-langs on other experimental setups as an intermediate between both extremes while also allowing us to observe differences in performance between seen/unseen languages during Statement-Tuning.
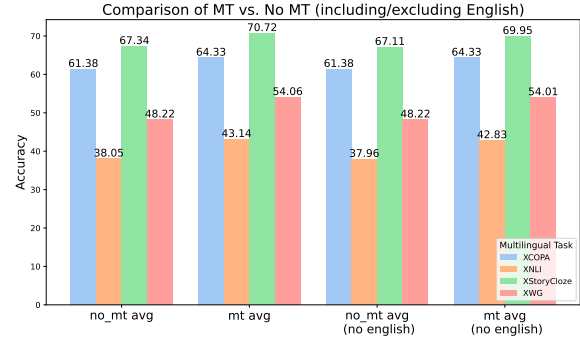


Figure 8: Mean task accuracy of mDeBERTa on an English-only task dataset (including and excluding MT statement data). On the left are the averages including English evaluation sets and on the right excluding them.

## 5.6 Including Machine Translation in Statement-Tuning Training Data improves Cross-lingual Transfer

In section 5.5, we observed that an English-only training setup that includes machine translation (MT) data can achieve most of the performance benefits of using multiple languages in the training task mixture. However, it remains unclear whether the inclusion of MT data itself is a key contributing factor. To investigate this, Figure 8 directly compares the effect of including versus excluding MT data in the English-only setup.

Interestingly, incorporating MT data leads to a notable performance increase across all tasks, both in the average performance across all languages and in the average performance excluding English (the "seen" language) from the evaluation. This suggests that MT data enhances cross-lingual transfer and is particularly beneficial when language-specific NLU task data is unavailable. In such cases, using English-only task data with MT achieves a significant portion of the multilingual performance. Additionally, since MT data is generally more accessible for lower-resource languages than NLU task data, it is relatively easy to incorporate into the Statement-Tuning training mix.

## 5.7 Multilingual Statement-Tuning Enables Efficient Inference for Zero-shot Cross-lingual Generalization

Though Statement-Tuning enables generalization into zero-shot settings with comparable performance against the zero-shot LLMs, increasing the number of candidate labels would also increase the model's computational overhead. In the case of a statement-tuned model, for a downstream task
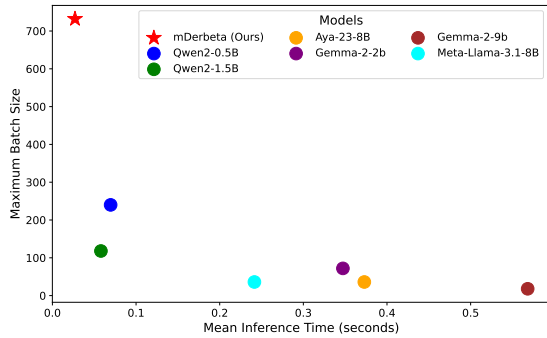
Figure 9: Mean inference time and maximum batch size of various models during a simulated text classification task on a single A100 GPU.

with $n$-possible labels, a naive way to perform a prediction is to iterate the model over $n$-times for each label. But in practice, it'd be more efficient to perform a batched prediction. Here, we compare the inference time and maximum batch sizes of our best statement-tuned model (mDeBERTa-v3-base; 0.276B) against zero-shot LLMs with varying parameter sizes, ranging from 0.5B to 9B.

We simulate a text classification task on each model, measuring the mean inference time per batch where we gradually increase the batch size. We perform this experiment on a single A100 GPU and show the result in Figure 9. As expected, due to our model size and non-autoregressive nature, it achieves the fastest mean inference time with the largest batch size capacity. Having a statement-tuned model that could handle $m$-batch size means that it could handle $m/n$ instances at once.

## 6 Conclusion

While large generative models dominate multilingual NLP, the potential of encoder-only models for cross-lingual generalization remains underexplored. We show that a well-designed finetuning setup enables state-of-the-art pretrained encoder-only models to match, or even surpass, generative models in three of four unseen cross-lingual NLU tasks, despite using far fewer parameters. Additionally, these models generalize across languages even when finetuned only on a monolingual multitask dataset, leveraging their multilingual pretraining. Our findings position encoder-only models as a memory-efficient alternative for multilingual multitask NLU. Future work can further optimize finetuning, extend cross-lingual generalization, and refine encoder architectures for large-scale multilingual learning.

## Limitations

Statement-Tuning highly relies on the training task selection and the proximity of the chosen tasks to the target task, hence the utility of the approach may still be limited if the training task selection fails to include similar enough examples to the target task. Due to this, some tasks like XWinograd may not be sufficiently addressed.

Statement-Tuning requires the use of verbalization which requires extra effort and careful prompt design. Furthermore, requiring a statement for each potential target makes this method infeasible for tasks with an extremely large hypothesis class.

Not all the encoder models we studied were capable of cross-lingual generalization and we were not able to pinpoint the exact mechanism during pretraining which enables such capabilities. We leave this for future work.

We were not able to control for the pretraining/instruction-tuning of all the models explored due to the lack of transparency regarding exact training data for some models. Hence, our analysis may include models which are not completely blind to the task data.

## References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giridharan Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeffrey Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Veselin Stoyanov. 2022. Efficient large scale language modeling with mixtures of experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11699–11732, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Viraat Aryabumi, John Dang, Dwarak Taluparu, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnsamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

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

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information*

*Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Ahmed Elshabrawy, Yongxin Huang, Iryna Gurevych, and Alham Fikri Aji. 2025. Enabling natural zero-shot prompting on encoder models via statement-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8302–8321, Albuquerque, New Mexico. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh,

Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. 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 1274–1287, Online. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Bolei Ma, Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023. Is prompt-based finetuning always better than vanilla finetuning? insights from cross-lingual language understanding. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 1–16, Ingolstadt, Germany. Association for Computational Lingustics.

Bolei Ma, Ercong Nie, Shuzhou Yuan, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schuetze. 2024. ToPro: Token-level prompt decomposition for cross-lingual sequence labeling tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2685–2702, St. Julian's, Malta. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 165–172, New York, NY, USA. Association for Computing Machinery.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt (gpt-3.5). https://openai.com/chatgpt. Accessed: August 2024.

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.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16339–16347, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Sebastian Ruder. 2022. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Saleh Soltan, Andy Rosenbaum, Tobias Falke, Qin Lu, Anna Rumshisky, and Wael Hamza. 2023. Recipes for sequential pre-training of multilingual encoder and Seq2Seq models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9380–9394, Toronto, Canada. Association for Computational Linguistics.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva

Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, X. Li, Zhi Yuan Lim, S. Soleman, R. Mahendra, Pascale Fung, Syafri Bahar, and A. Purwarianti. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Haike Xu, Zongyu Lin, Jing Zhou, Yanan Zheng, and Zhilin Yang. 2023. A universal discriminator for zero-shot generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10559–10575, Toronto, Canada. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing

Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023b. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

# A Statement Templates

## A.1 Multiple-Choice QA Templates

### A.1.1 Belebele Templates

| Task | Statement Template |
|---|---|
| Belebele | {{context}} {{question}} {{correct_answer/other_answer}}<br>{{context}} According to the passage above, the answer of {{question}} is {{correct_answer/other_answer}}<br>Passage: {{context}} Question: {{question}} Answer: {{correct_answer/other_answer}}<br>{{context}} Q: {{question}} A: {{correct_answer/other_answer}}<br>Content: "{{context}}"\n Inquiry: "{{question}}"\n Response: "{{correct_answer/other_answer}}"<br>Text: "{{context}}"\n Query: "{{question}}"\n Solution: "{{correct_answer/other_answer}}"<br>Passage content: "{{context}}"\n What is asked: "{{question}}"\n The answer is: "{{correct_answer/other_answer}}"<br>Here is the passage: "{{context}}"\n The question is: "{{question}}"\n The provided answer is: "{{correct_answer/other_answer}}"<br>The passage reads: "{{context}}"\n Asked: "{{question}}"\n The correct answer is: "{{correct_answer/other_answer}}"<br>From the text: {{context}}\n As stated above, the response to "{{question}}" is "{{correct_answer/other_answer}}"<br>Based on the passage: {{context}}\n The answer to "{{question}}" according to the text is "{{correct_answer/other_answer}}"<br>In reference to the passage: {{context}}\n Thus, the answer to "{{question}}" is "{{correct_answer/other_answer}}"<br>In reference to the passage: {{context}}\n According to the text, the answer for "{{question}}" is "{{correct_answer/other_answer}}"<br>Given the text: {{context}}\n Therefore, the answer to "{{question}}" is "{{correct_answer/other_answer}}"<br>Text: {{context}}\n Inquiry: {{question}}\n Response: {{correct_answer/other_answer}}<br>Content: {{context}}\n Question asked: {{question}}\n Given Answer: {{correct_answer/other_answer}}<br>Passage: {{context}}\n Question: {{question}}\n Answer Provided: {{correct_answer/other_answer}}<br>The text reads: {{context}}\n Query: {{question}}\n Solution: {{correct_answer/other_answer}}<br>Content: {{context}}\n What is the question: {{question}}\n Answer: {{correct_answer/other_answer}}<br>{{context}}\n Question: {{question}}\n Answer: {{correct_answer/other_answer}}<br>{{context}}\n Inquiry: {{question}}\n Response: {{correct_answer/other_answer}}<br>{{context}}\n What is being asked: {{question}}\n The answer is: {{correct_answer/other_answer}}<br>{{context}}\n The question posed is: {{question}}\n The correct answer is: {{correct_answer/other_answer}}<br>{{context}}\n The text asks: {{question}}\n The response provided is: {{correct_answer/other_answer}} |

### A.1.2 Exams Templates

| Task | Statement Template |
|---|---|
| Exams | Q: {{question}}. A: {{correct_answer/other_answer}}<br>{{question}}. Answer: {{correct_answer/other_answer}}<br>Question: {{question}} Answer: {{correct_answer/other_answer}} |

### A.1.3 xQuAD Templates

| Task | Statement Template |
|---|---|
| xQuAD | {{context}} Question: {{question}} Answer: {{correct_answer/other_answer}}<br>Passage: {{context}} Question: {{question}} Answer: {{correct_answer/other_answer}}<br>{{context}} Q: {{question}} A: {{correct_answer/other_answer}}<br>{{context}} According to the passage above, the answer of {{question}} is {{correct_answer/other_answer}}<br>Text: {{context}}\n Question: {{question}}\n Reply: {{correct_answer/other_answer}}<br>Passage text: {{context}}\n What is the solution: {{question}}\n Answer: {{correct_answer/other_answer}}<br>{{context}}\n In reference to the text, what is the answer: {{question}}\n Answer: {{correct_answer/other_answer}}<br>From the given passage: {{context}}\n Query: {{question}}\n Solution: {{correct_answer/other_answer}}<br>Passage: {{context}}\n Q: {{question}}\n A: {{correct_answer/other_answer}}<br>Text: {{context}}\n What is the response: {{question}}\n Answer: {{correct_answer/other_answer}}<br>Context: {{context}}\n Answer for {{question}} is: {{correct_answer/other_answer}}<br>According to the context: {{context}}\n Solution to {{question}} is: {{correct_answer/other_answer}}<br>Text: {{context}}\n Answer to {{question}} is: {{correct_answer/other_answer}}<br>{{context}}\n In reference to the passage, {{question}} has the answer: {{correct_answer/other_answer}}<br>{{context}}\n Answer to the question {{question}} based on the passage is: {{correct_answer/other_answer}}<br>{{context}}\n From the passage, the response to {{question}} is: {{correct_answer/other_answer}}<br>Text: {{context}}\n Question: {{question}}\n Response: {{correct_answer/other_answer}}<br>Passage: {{context}}\n Query: {{question}}\n Answer: {{correct_answer/other_answer}}<br>Context: {{context}}\n What is the answer to {{question}}?\n Solution: {{correct_answer/other_answer}}<br>From the text: {{context}}\n What is the solution to {{question}}?\n Answer: {{correct_answer/other_answer}} |

## A.2 Summarization Templates

### A.2.1 WikiLingua Templates

| Task | Statement Template |
|---|---|
| WikiLingua | Passage: {{source}}, Summary: {{correct_target/random_target}}<br>The summary of "{{source}}" is {{correct_target/random_target}}<br>Context: {{source}}, Summary: {{correct_target/random_target}}<br>Q: Summarize the following: {{source}}, A: {{correct_target/random_target}}<br>The answer of "Summarize the following {{source}}" is {{correct_target/random_target}} |

## A.3 Machine Translation Templates

### A.3.1 FLORES-101 Templates

| Task | Statement Template |
|---|---|
| FLORES-101 | The {{target_lang}} translation of {{lang}} sentence {{sentence}} is {{target_sentence}}<br>The {{target_lang}} translation of {{lang}} sentence {{sentence}} is not {{target_sentence}} |

## A.4 Sentiment Analysis Templates

### A.4.1 multilingual-sentiments Templates

| Task | Statement Template |
|---|---|
| multilingual-sentiments | The text "{{text}}" is {{correct_label/other_label}}.<br>Sentence: '{{text}}'. Label: {{correct_label/other_label}}<br>Sentiment Analysis:\nText: {{text}}\nResult: {{correct_label/other_label}}<br>The sentiment of the text {{text}} is {{correct_label/other_label}}<br>Text: {{text}} has a sentiment labeled as {{correct_label/other_label}}<br>The text {{text}} conveys a sentiment of {{correct_label/other_label}}<br>The analysis reveals that {{text}} is characterized by a sentiment of {{correct_label/other_label}}<br>For the text {{text}}, the sentiment is identified as {{correct_label/other_label}}<br>The sentiment associated with {{text}} is {{correct_label/other_label}}<br>In terms of sentiment, {{text}} reflects {{correct_label/other_label}} |

## A.5 Word Sense Disambiguation Templates

### A.5.1 XL-WiC Templates

| Task | Statement Template |
|---|---|
| XL-WiC | "{{target_word}}" means the same in "{{context_1}}" and "{{context_2}}"<br>"{{target_word}}" does not mean the same in "{{context_1}}" and "{{context_2}}"<br>The meaning of "{{target_word}}" is consistent across "{{context_1}}" and "{{context_2}}"<br>The meaning of "{{target_word}}" is inconsistent across "{{context_1}}" and "{{context_2}}"<br>The interpretation of "{{target_word}}" remains unchanged in both "{{context_1}}" and "{{context_2}}"<br>The interpretation of "{{target_word}}" changes in "{{context_1}}" and "{{context_2}}"<br>The sense of {{target_word}} is identical between {{context_1}} and {{context_2}}<br>The sense of {{target_word}} differs between {{context_1}} and {{context_2}}<br>The interpretation of {{target_word}} is the same in both {{context_1}} and {{context_2}}<br>The sense of {{target_word}} varies between {{context_1}} and {{context_2}}<br>{{target_word}} has the same meaning in both {{context_1}} and {{context_2}}<br>The meaning of {{target_word}} is different in {{context_1}} and {{context_2}} |

## A.6 Intent Classification Templates

### A.6.1 MASSIVE Templates

| Task | Statement Template |
|---|---|
| MASSIVE | The utterance "{{utt}}" is under the "{{scenario}}" scenario.<br>Utterance: "{{utt}}" Scenario: "{{scenario}}"<br>User: "{{utt}}". The best scenario for the user query is "{{scenario}}".<br>The scenario of user's utterance "{{utt}}" is "{{scenario}}". |

## A.7 Commonsense Reasoning Templates

### A.7.1 Multilingual Fig-QA Templates

| Task | Statement Template |
|---|---|
| Multilingual Fig-QA | "{{startphrase}}" "{{ending1/ending2}}"<br>"{{startphrase}}" therefore "{{ending1/ending2}}"<br>Startphrase: "{{startphrase}}" ending: "{{ending1/ending2}}"<br>"{{startphrase}}" then "{{ending1/ending2}}"<br>if "{{startphrase}}" then "{{ending1/ending2}}"<br>"{{startphrase}}" means "{{ending1/ending2}}" |

### A.7.2 X-CSQA Templates

| Task | Statement Template |
|---|---|
| X-CSQA | Question: "{{question}}". Answer: "{{correct_answer/other_answer}}"<br>Q: "{{question}}". A: "{{correct_answer/other_answer}}"<br>"{{question}}". Ans: {{correct_answer/other_answer}}"<br>Inquiry: {{question}}\nResponse: "{{correct_answer/other_answer}}"<br>The question: {{question}} has the answer: "{{correct_answer/other_answer}}"<br>Question posed: "{{question}}". Possible response: "{{correct_answer/other_answer}}"<br>In response to {{question}}, the answer is "{{correct_answer/other_answer}}"<br>Query: {{question}}\nResponse: "{{correct_answer/other_answer}}"<br>The query: {{question}} yields the answer: "{{correct_answer/other_answer}}"<br>The answer to {{question}} could be: "{{correct_answer/other_answer}}"<br>For the question: {{question}}, the answer is "{{correct_answer/other_answer}}"<br>The inquiry {{question}} could receive the answer: "{{correct_answer/other_answer}}"<br>The query: {{question}} has the answer: "{{correct_answer/other_answer}}"<br>When posed with the question: {{question}}, the answer provided is "{{correct_answer/other_answer}}"<br>Upon inquiry: {{question}}, the answer provided is "{{correct_answer/other_answer}}" |

### A.7.3 X-CODAH Templates

| Task | Statement Template |
|---|---|
| X-CODAH | The statement {{correct_text}} makes more sense than the statement {{other_text}}.<br>Statement {{correct_text}} is more logical than {{other_text}}.<br>The statement {{correct_text}} makes sense.<br>The statement {{correct_text}} is clearer compared to {{other_text}}.<br>{{correct_text}} is more reasonable than {{other_text}}.<br>Between the two, {{correct_text}} is the more sensible statement over {{other_text}}.<br>{{correct_text}} presents a clearer rationale than {{other_text}}.<br>When comparing, {{correct_text}} is more coherent than {{other_text}}.<br>Statement {{correct_text}} exhibits greater logic than {{other_text}}.<br>In terms of logic, {{correct_text}} surpasses {{other_text}}.<br>{{correct_text}} shows a higher degree of logical reasoning than {{other_text}}.<br>Compared to {{other_text}}, statement {{correct_text}} is the more logical choice.<br>Between the two, {{correct_text}} is the more logical statement compared to {{other_text}}.<br>The statement {{correct_text}} is sensible and coherent.<br>Clearly, the statement {{correct_text}} is logical.<br>It is evident that the statement {{correct_text}} is reasonable.<br>Undoubtedly, the statement {{correct_text}} holds logic.<br>There is clarity in the statement {{correct_text}}. |

## A.8 Topic Classification Templates

### A.8.1 SIB-200 Templates

| Task | Statement Template |
|------|--------------------|
| SIB-200 | Sentence: {{text}}. Label: {{label}}. |
| | The sentence {{text}} is considered a {{label}} sentence. |
| | The sentence {{text}} is not considered a {{label}} sentence. |
| | The sentence {{text}} is about {{label}}. |
| | The sentence {{text}} is not about {{label}}. |
| | The sentence {{text}} is a {{label}} sentence. |
| | The sentence {{text}} is not a {{label}} sentence. |
| | Text: {{text}} \n Category: {{label}}. |
| | The text: "{{text}}" is labeled as {{label}}. |
| | Sentence: "{{text}}" \n Topic: {{label}}. |
| | The given sentence "{{text}}" belongs to the category: {{label}}. |
| | The sentence describes a {{label}} topic: {{text}}. |
| | The text "{{text}}" discusses {{label}}. |
| | "{{text}}" talks about the topic: {{label}}. |
| | This sentence, "{{text}}", revolves around {{label}}. |
| | "{{text}}" is centered on {{label}}. |
| | The topic of "{{text}}" is {{label}}. |
| | The text "{{text}}" does not discuss {{label}}. |
| | "{{text}}" does not talk about {{label}}. |
| | This sentence, "{{text}}", does not revolve around {{label}}. |
| | "{{text}}" is not related to {{label}}. |
| | The topic of "{{text}}" is not {{label}}. |
| | The text "{{text}}" is regarded as a {{label}} sentence. |
| | "{{text}}" is classified as a {{label}} sentence. |
| | This sentence, "{{text}}", is viewed as {{label}}. |
| | The text is recognized as {{label}}: "{{text}}". |
| | The classification of "{{text}}" is {{label}}. |
| | The text "{{text}}" is not regarded as a {{label}} sentence. |
| | "{{text}}" is not classified as a {{label}} sentence. |
| | This sentence, "{{text}}", is not viewed as {{label}}. |
| | The text is not recognized as {{label}}: "{{text}}". |
| | The classification of "{{text}}" is not {{label}}. |
| | The sentence "{{text}}" falls under the category of {{label}}. |
| | "{{text}}" is labeled as a {{label}} sentence. |
| | This sentence, "{{text}}", is classified as {{label}}. |
| | The text "{{text}}" belongs to the {{label}} category. |
| | "{{text}}" is a sentence of the {{label}} type. |
| | The sentence "{{text}}" does not fall under the category of {{label}}. |
| | "{{text}}" is not labeled as a {{label}} sentence. |
| | This sentence, "{{text}}", is not classified as {{label}}. |
| | The text "{{text}}" does not belong to the {{label}} category. |
| | "{{text}}" is not a sentence of the {{label}} type. |

## A.9 Paraphrase Detection Templates

### A.9.1 PAWS-X Templates

| Task | Statement Template |
|------|--------------------|
| PAWS-X | "{{text1}}" can be stated as "{{text2}}". |
| | "{{text1}}" can not be stated as "{{text2}}". |
| | "{{text1}}" can't be stated as "{{text2}}". |
| | "{{text1}}" duplicates "{{text2}}". |
| | "{{text1}}" does not duplicate "{{text2}}". |
| | "{{text1}}" doesn't duplicate "{{text2}}". |
| | "{{text1}}" is a duplicate of "{{text2}}". |
| | "{{text1}}" is not a duplicate of "{{text2}}". |
| | "{{text1}}" is the same as "{{text2}}". |
| | "{{text1}}" is not the same as "{{text2}}". |
| | "{{text1}}" is unrelated to "{{text2}}". |
| | "{{text1}}" is a paraphrase of "{{text2}}". |
| | "{{text1}}" is not a paraphrase of "{{text2}}". |
| | "{{text1}}" isn't a paraphrase of "{{text2}}". |

## A.10 Sentence Completion Templates

### A.10.1 XCOPA Templates

| Task | Statement Template |
|------|--------------------|
| XCOPA | The cause of {{premise}} is that {{choice1/choice2}}. |
| | {{premise}} due to {{choice1/choice2}}. |
| | The effect of {{premise}} is that {{choice1/choice2}}. |
| | {{premise}} therefore {{choice1/choice2}}. |
| | {{premise}}, so {{choice1/choice2}}. |

### A.10.2 XStoryCloze Templates

| Task | Statement Template |
|------|--------------------|
| XStoryCloze | {{text1}} entails {{text2}}. |
| | {{text1}}? yes, {{text2}}. |
| | Premise: {{text1}}, Hypothesis: {{text2}}, label: Entailment. |
| | {{text1}} is neutral with regards to {{text2}}. |
| | {{text1}}? maybe, {{text2}}. |
| | Premise: {{text1}}, Hypothesis: {{text2}}, label: Neutral. |
| | {{text1}} contradicts {{text2}}. |
| | {{text1}}? no, {{text2}}. |
| | Premise: {{text1}}, Hypothesis: {{text2}}, label: Contradiction. |

## A.11 Natural Language Inference Templates

### A.11.1 XNLI Templates

| Task | Statement Template |
|------|--------------------|
| XNLI | In {{sentence}}, _ is: {{option1/option2}}. |
| | Q:{{sentence}}, A: {{option1/option2}}. |
| | The missing word in {{sentence}} is {{option1/option2}}. |
| | _ in: {{sentence}} is {{option1/option2}}. |
| | {{sentence}}, _ is: {{option1/option2}}. |

## A.12 Coreference Resolution Templates

### A.12.1 XWinograd Templates

| Task | Statement Template |
|------|--------------------|
| XWinograd | {{input_sentence_1}} {{input_sentence_2}} {{input_sentence_3}} {{input_sentence_4}} The right way to close this story is: {{sentence_quiz1/sentence_quiz2}}. |
| | {{input_sentence_1}} {{input_sentence_2}} {{input_sentence_3}} {{input_sentence_4}} The proper ending to this story is: {{sentence_quiz1/sentence_quiz2}}. |
| | {{input_sentence_1}} {{input_sentence_2}} {{input_sentence_3}} {{input_sentence_4}} The correct ending to this story is: {{sentence_quiz1/sentence_quiz2}}. |

## B Training Datasets

The following datasets where used to create the statement dataset of Multilingual Statement-Tuning and the instruction dataset for the decoder models: Belebele (reading comprehension) (Bandarkar et al., 2024), Exams (Question Answering) (Hardalov et al., 2020), xQuAD (Question Answering) (Artetxe et al., 2020) for multiple-choice question answering; WikiLingua (Ladhak et al., 2020) for summarization; FLORES-101 (Goyal et al., 2022) for machine translation; Multilingual Sentiments (IndoNLU, Multilingual Amazon Reviews, GoEmotions, Offenseval Dravidian, SemEval-2018 Task 1: Affect in Tweets, Emotion, IMDB, Amazon Polarity, Yelp Reviews, Yelp Polarity) (Wilie et al., 2020; Keung et al., 2020; Demszky et al., 2020; Chakravarthi et al., 2021; Hande et al., 2020; Chakravarthi et al., 2020b,a; Mohammad et al., 2018; Saravia et al., 2018; Maas et al., 2011; McAuley and Leskovec, 2013; Zhang et al., 2015a,b) for sentiment analysis; XL-WiC (Raganato et al., 2020) for word sense disambiguation; MASSIVE (FitzGerald et al., 2023) for intent classification; Multilingual Fig-QA (Kabra et al., 2023), X-CSQA and X-CODAH (Lin et al., 2021) for commonsense reasoning; SIB-200 (Adelani et al., 2024) for topic classification; and PAWS-X (Yang et al., 2019) for paraphrase detection.

# C   Languages

| ISO | Language | Family | Subgrouping | Script | Resource |
|-----|----------|--------|-------------|--------|----------|
| af | Afrikaans | Indo-European | Germanic | Latin | High |
| ar | Arabic | Afro-Asiatic | Semitic | Arabic | High |
| de | German | Indo-European | Germanic | Latin | High |
| en | English | Indo-European | Germanic | Latin | High |
| es | Spanish | Indo-European | Italic | Latin | High |
| fr | French | Indo-European | Italic | Latin | High |
| ga | Irish | Indo-European | Celtic | Latin | Low |
| gu | Gujarati | Indo-European | Indo-Aryan | Gujarati | Low |
| ha | Hausa | Afro-Asiatic | Chadic | Latin | Low |
| hi | Hindi | Indo-European | Indo-Aryan | Devanagari | High |
| id | Indonesian | Austronesian | Malayo-Polynesian | Latin | High |
| ig | Igbo | Atlantic-Congo | Benue-Congo | Latin | Low |
| is | Icelandic | Indo-European | Germanic | Latin | High |
| it | Italian | Indo-European | Italic | Latin | High |
| kk | Kazakh | Turkic | Common Turkic | Cyrillic | High |
| ky | Kyrgyz | Turkic | Common Turkic | Cyrillic | Low |
| lo | Lao | Tai-Kadai | Kam-Tai | Lao | Low |
| mt | Maltese | Afro-Asiatic | Semitic | Latin | High |
| ny | Nyanja | Atlantic-Congo | Benue-Congo | Latin | Low |
| pt | Portuguese | Indo-European | Italic | Latin | High |
| ru | Russian | Indo-European | Balto-Slavic | Cyrillic | High |
| si | Sinhala | Indo-European | Indo-Aryan | Sinhala | Low |
| tr | Turkish | Turkic | Common Turkic | Latin | High |
| vi | Vietnamese | Austroasiatic | Vietic | Latin | High |
| zh | Chinese | Sino-Tibetan | Sinitic | Han | High |

Table 3: Languages used in this study in alphabetical order of ISO 639-1 Code. Information on language family, subgrouping, script, and resource level is drawn from (Costa-jussà et al., 2022).

## D   Finetuning Setup

We include the finetuning setup of the Statement Tuned Encoder models in Table 4.

| Model | #Epochs | Batch Size | Learning Rate | Weight Decay | Warmup Ratio |
|---|---|---|---|---|---|
| google-bert/bert-base-multilingual-cased (mBERT) | 20 | | 1.00e-6 | | |
| microsoft/mdeberta-v3-base (mDeBERTa) | | 16 | 2.00e-6 | 0.1 | 0.1 |
| FacebookAI/xlm-roberta-base (XLM-R base) | 15 | | 1.00e-6 | | |
| FacebookAI/xlm-roberta-large (XLM-R base) | | | 2.00e-6 | | |

Table 4: Finetuning Setup and Hyperparameters for each encoder model.

Additionally, the decoder models were Instruction finetuned. All models above 2B parameters are finetuned using QLoRA (Dettmers et al., 2023), while all models under 2B parameters are finetuned using full finetuning. We include the specific hyperparameters in Table 5. We used a custom instruction dataset of 150K examples constructed from the same task mixture as Statement-Tuning. The instruction templates are outlined in Appendix H.

| Model | #Epochs | Mode | Batch Size | Learning Rate | Weight Decay | Warmup Ratio |
|---|---|---|---|---|---|---|
| Llama3.1 8B | | | | 0.00001 | | steps=10 |
| Qwen2 72B | | | | 0.0002 | | steps=10 |
| Llama3.1 70B | | | | 0.00001 | | steps=10 |
| Gemma 2 9B | | QLoRA | | 0.0002 | | 0.1 |
| Gemma 2 27B | | | | 0.0002 | | 0.1 |
| Aya 23 8b | 1 | | 4 | 0.00001 | 0 | steps=10 |
| Aya 23 35b | | | | 0.00001 | | steps=10 |
| Gemma 2 2b | | | | 0.0002 | | steps=10 |
| Qwen2 1.5B | | FFT | | 0.0002 | | steps=10 |
| Qwen2 0.5B | | | | 0.0002 | | steps=10 |

Table 5: Finetuning Setup and Hyperparameters for each decoder model.

## E  Language Level Performance

In Figure 10 (fine-tuned on the same data) and Figure 11 (fine-tuned on custom data mixtures by the teams who developed the models), we report the individual language performance on all 4 evaluation tasks of all generative models and mDeBERTa fine-tuned using Statement-Tuning. There are several interesting observations.

In both cases of instruction finetuning setup, we observe largely the same trends. First, on XCOPA, XNLI, and XStoryCloze we notice that mDeBERTa tends to perform more equitably than the LLMs, meaning that there is less variation **across** languages in a single task/dataset. For example, in XCOPA, Qwen2 72B and Llama3.1 70B perform strongly on Indonesian, Italian, Vietnamese, and Chinese but have lackluster performance on most of the other languages. While mDeBERTa seems to have less deviation between the best performing languages and the others. We see this in XNLI and XStoryCloze as well (for example Arabic, Swahili, and Urdu in XNLI, and Swahilli, Telugu, and Basque in StoryCloze). This adds more support for the use of our method with lower resource/tail-end languages.

Second, we notice that our method can generalize to languages/language families that are unseen during Statement-Tuning if they are seen during pretraining. For example, Turkish (tr) and all Turkic languages for that matter are completely unseen during Statement-Tuning, but are seen during pretraining, the model was still able to generalize on Turkish on XNLI performing on par with Aya 23 35b. Moreover, Burmese (my) and all its closely related languages are completely unseen during Statement-Tuning while being seen during pretraining, but the performance on XStoryCloze was exceptionally strong far outperforming even the strongest generative model (**72.93** of mDeBERTa vs. **52.81** of Llama3.1 70B fine-tuned on the same data mixture). On the other hand, a language where our model fails to generalize Quechua (qu) on XCOPA was completely unseen during Statement-Tuning and pretraining. This is encouraging as it further supports our hypothesis that multilingual pretraining is what powers Multilingual Statement-Tuning. This should encourage the development of more powerful encoder-only models with support for more languages.

## F  Languages Used during Statement Tuning

As a subset of the potential 25 languages from the training set we choose the following 11 languages as an intermediate subset:

- Chinese
- English
- French
- Vietnamese
- Swahili
- Russian
- Arabic
- Hindi
- German
- Indonesian
- Italian

We make the choice of these specific languages as they span a variety of language families, scripts, and resource availability and hence could potentially help with cross-lingual generalization.
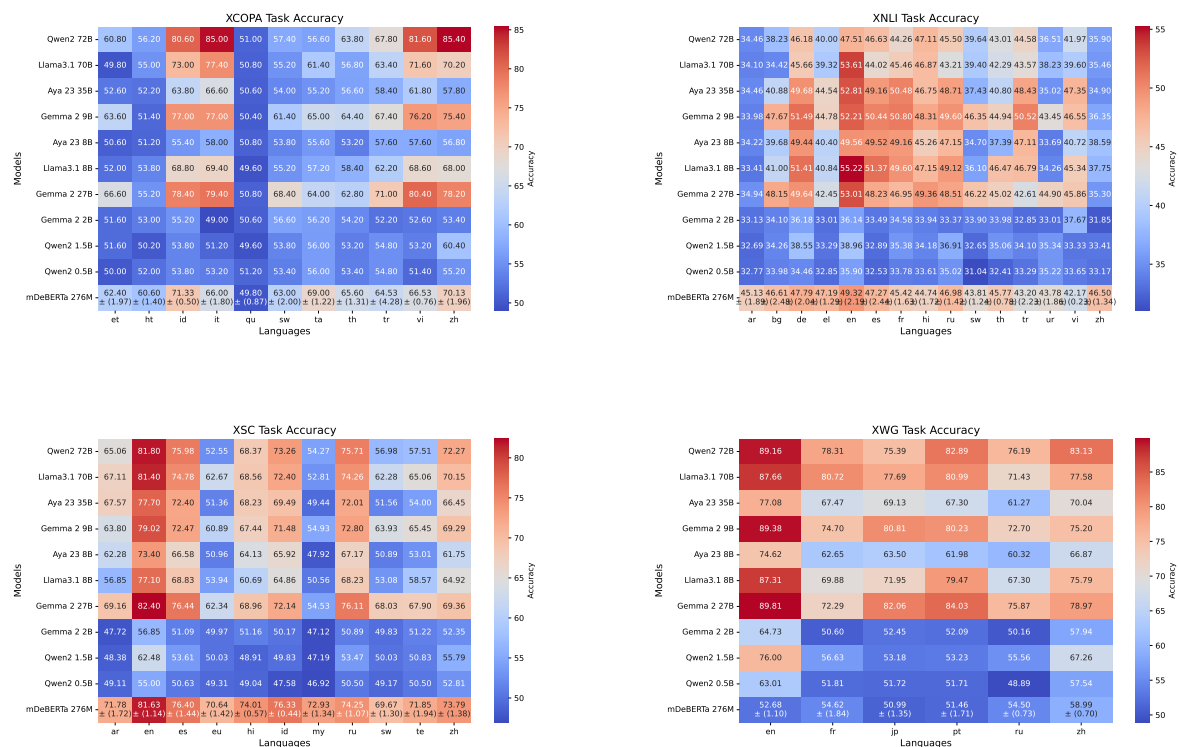
Figure 10: Individual language subset performance on all 4 evaluation tasks and decoder models finetuned on the same data mixture as Statement-Tuningand mDeBERTa trained on 11-langs. We include the standard deviation in performance over 3 training runs for mDeBERTa.
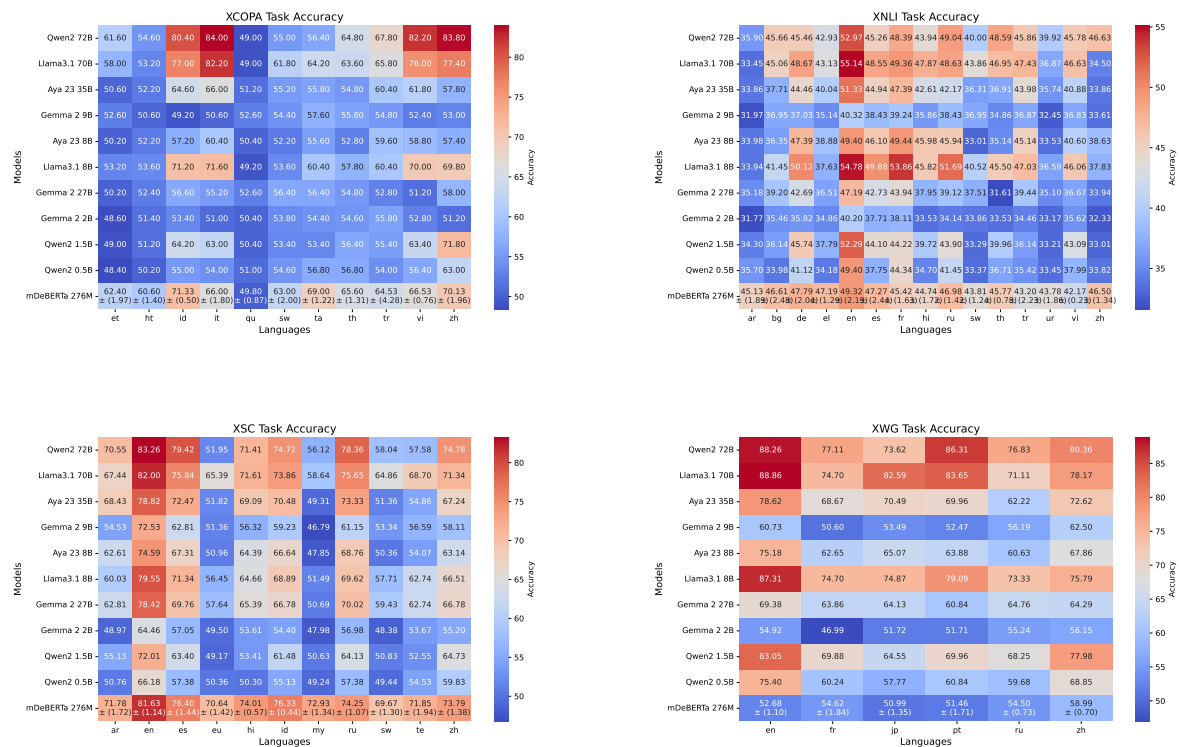


Figure 11: Individual language subset performance on all 4 evaluation tasks and decoder models (instruction-tuned on custom data by the teams who released the models) and mDeBERTa trained on 11-langs. We include the standard deviation in performance over 3 training runs for mDeBERTa.

16245

# G  Performance of Instruction-Tuned Model Variants

| Model | Parameters | XCOPA | XNLI | XStoryCloze | XWinoGrad |
|---|---|---|---|---|---|
| Qwen2 | 72B | 67.24 | 45.09 | 68.74 | 80.42 |
| Llama3.1 | 70B | 66.20 | 45.07 | 70.48 | 79.85 |
| Gemma 2 | 9B | 53.00 | 36.33 | 57.52 | 56.00 |
| Llama3.1 | 8B | 60.98 | 44.84 | 64.45 | 77.52 |
| Aya 23 | 8B | 55.16 | 41.30 | 60.97 | 65.88 |
| Aya 23 | 35B | 57.31 | 40.81 | 64.29 | 70.43 |
| Gemma 2 | 27B | 54.24 | 38.59 | 64.59 | 64.54 |
| Gemma 2 | 2B | 52.49 | 34.97 | 53.65 | 52.79 |
| Qwen2 | 1.5B | 57.42 | 39.79 | 57.95 | 72.28 |
| Qwen2 | 500M | 54.56 | 37.56 | 54.59 | 63.80 |
| mBERT(base) | 110M | 52.47 | 34.51 | 48.30 | 50.68 |
| XLMR-base | 250M | 56.69 | 35.33 | 60.71 | 51.34 |
| XLMR-large | 560M | 64.36 | 45.76 | 78.78 | 54.26 |
| **mDeBERTa (Best)** | 276M | $65.52_{(1.64)}$ | $47.84_{(1.65)}$ | $73.53_{(1.25)}$ | $54.75_{(1.24)}$ |

Table 6: **Accuracy of the existing instruction-tuned varieities of multilingual decoder and Statement-Tuned encoder models** on XCOPA, XNLI, XStoryCloze, and XWinoGrad tasks. Results in grey highlight performances that are below the best-performing encoder model, mDeBERTa (276M). Additionally, we report the average standard deviation across languages over 3 training runs only for mDeBERTa to quantify the random deviation due to Statement-Tuning training.

# H Instruction Templates

## H.1 Multiple-Choice QA Templates

### H.1.1 Belelebe Templates

| Task | Instruction Template |
|---|---|
| Belebele | {{flores_passage}}\n\nQuestion: {{question}}\n{{options_}}<br>{{flores_passage}}\n\n{{question}}\n{{options_}}<br>{{flores_passage}}\n\nAnswer the following question: {{question}}\n{{options_}}<br>{{flores_passage}}\n\nBased on the preceding passage, answer the following question {{question}}\n{{options_}}<br>{{flores_passage}}\n\nGive an answer to the following question using evidence from the above passage: {{question}}\n{{options_}}<br>Context: {{flores_passage}}\n\nQuestion {{question}}\n\nAnswer:\n{{options_}}<br>Read the following passage and answer the question.\n\n{{flores_passage}}\n\n{{question}}\n{{options_}}<br>Answer the question about the text:\n\n{{flores_passage}}\n\n{{question}}\n{{options_}}<br>{{flores_passage}}\n\nWhat is the correct answer to the following question based on the text?\n{{question}}\n{{options_}}<br>Refer to the passage below:\n\n{{flores_passage}}\n\nWhat is the answer to this question?\n{{question}}\n{{options_}} |

### H.1.2 Exams Templates

| Task | Instruction Template |
|---|---|
| Exams | Question: {{stem}}\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}<br>{{stem}}\n\nOptions:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}<br>Read the question:\n\n{{stem}}\n\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}<br>What is the correct answer to this question?\n\n{{stem}}\n\nOptions:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}<br>Based on the question below, choose the most accurate answer:\n\n{{stem}}\n\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}<br>Select the right option for the following question:\n\n{{stem}}\n\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}<br>Question:\n{{stem}}\n\nOptions:\n1. {{choice1}}\n2. {{choice2}}\n3. {{choice3}}\n4. {{choice4}}<br>Answer the following question by selecting the best choice:\n\n{{stem}}\n\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}<br>{{stem}}\n\nAnswer options:\n1. {{choice1}}\n2. {{choice2}}\n3. {{choice3}}\n4. {{choice4}}<br>Below is a question and four possible answers. Choose one.\n\nQuestion: {{stem}}\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}} |

### H.1.3 xQuAD Templates

| Task | Instruction Template |
|---|---|
| xQuAD | Please answer a question about the following article:\n\n{{context}}\n\n{{question}}<br>Read this and answer the question:\n\n{{context}}\n\n{{question}}<br>{{context}}\n{{question}}<br>Answer a question about this article:\n{{context}}\n{{question}}<br>Here is an article: {{context}}\n\nWhat is the answer to this question: {{question}}<br>Article: {{context}}\n\nQuestion: {{question}}<br>Article: {{context}}\n\nNow answer this question: {{question}}<br>{{context}}\n\nQ: {{question}}<br>Read the following article and answer the question:\n\n{{context}}\n\nQuestion: {{question}}<br>Please read this passage and provide an answer to the following question:\n\n{{context}}\n\nQuestion: {{question}} |

## H.2 Summarization Templates

### H.2.1 WikiLingua Templates

| Task | Instruction Template |
|---|---|
| WikiLingua | Summarize:\n\n{{source}}<br>Summarize the following:\n\n{{source}}<br>Summarize this passage:\n\n{{source}}<br>Provide a concise summary of this passage:\n{{source}}<br>What is a summary of the following passage?\n{{source}}<br>Describe the key points of the following passage:\n{{source}}<br>Passage: {{source}}\n\nWhat is a summary?<br>Passage: {{source}}\n\nWhat is a summary of what this passage is about?<br>Provide a brief overview of the following passage:\n\n{{source}}<br>Condense the key information from the following passage:\n\n{{source}} |

## H.3 Machine Translation Templates

### H.3.1 FLORES-101 Templates

| Task | Instruction Template |
|---|---|
| FLORES-101 | Translate the following sentence from {{ori_lang}} to {{target_lang}}:\n\n{{ori_sen}}<br>Please translate this sentence from {{ori_lang}} into {{target_lang}}:\n\n{{ori_sen}}<br>Translate the {{ori_lang}} sentence to {{target_lang}}:\n\n{{ori_sen}}<br>Convert the following {{ori_lang}} sentence into {{target_lang}}:\n\n{{ori_sen}}<br>Translate this sentence from {{ori_lang}} into {{target_lang}}:\n\n{{ori_sen}}<br>Please provide the translation of the following {{ori_lang}} sentence into {{target_lang}}:\n\n{{ori_sen}}<br>Translate this {{ori_lang}} sentence to {{target_lang}}:\n\n{{ori_sen}}<br>Convert this {{ori_lang}} sentence to {{target_lang}}:\n\n{{ori_sen}}<br>Given the following {{ori_lang}} sentence, translate it into {{target_lang}}:\n\n{{ori_sen}}<br>How do you say the following sentence in {{target_lang}}?\n{{ori_sen}} |

## H.4 Sentiment Analysis Templates

### H.4.1 multilingual-sentiments Templates

| Task | Instruction Template |
|---|---|
| multilingual-sentiments | {{text}}\nWhat is the sentiment of this text? (positive, neutral, negative)<br>{{text}}\nIs the sentiment of this text positive, neutral, or negative?<br>What sentiment does this text express (positive, neutral, negative)?\n\n{{text}}<br>Describe the sentiment of the following text (positive, neutral, negative):\n\n{{text}}<br>How would you classify the sentiment of this text?\n\n{{text}} (positive, neutral, negative)<br>What sentiment best describes this text? (positive, neutral, negative)\n\n{{text}}<br>Analyze the sentiment of the following text:\n\n{{text}} (positive, neutral, negative)<br>{{text}}\nLabel the sentiment as positive, neutral, or negative:<br>Classify the sentiment expressed in this text as positive, neutral, or negative:\n\n{{text}}<br>Determine whether the sentiment of this text is positive, neutral, or negative:\n\n{{text}} |

## H.5 Word Sense Disambiguation Templates

### H.5.1 XL-WiC Templates

| Task | Instruction Template |
|---|---|
| XL-WiC | Context 1: {{context_1}}\nContext 2: {{context_2}}\nDoes the word '{{target_word}}' have the same meaning in both contexts? (Yes or No)<br>Given the word '{{target_word}}', compare the following two contexts\nContext 1: {{context_1}}\nContext 2: {{context_2}}\nIs the word used in the same sense in both contexts? (Yes or No)<br>Context 1: {{context_1}}\nContext 2: {{context_2}}\nDoes the target word '{{target_word}}' have the same meaning in both contexts? (Yes or No)<br>Context 1: {{context_1}}\nContext 2: {{context_2}}\nIs the word '{{target_word}}' used in the same sense in both contexts? (Yes or No)<br>Check if the word '{{target_word}}' has the same meaning in the following contexts\nContext 1: {{context_1}}\nContext 2: {{context_2}}\nAnswer with 'Yes' or 'No'<br>Does the word '{{target_word}}' have the same sense in both contexts?\nContext 1: {{context_1}}\nContext 2: {{context_2}}\nAnswer 'Yes' or 'No'<br>Context 1: {{context_1}}\nContext 2: {{context_2}}\nDoes the word '{{target_word}}' share the same meaning in both? (Yes/No)<br>Given the following contexts, is the meaning of '{{target_word}}' the same in both?\nContext 1: {{context_1}}\nContext 2: {{context_2}}\nAnswer 'Yes' or 'No'<br>Is the word '{{target_word}}' used with the same meaning in both contexts?\nContext 1: {{context_1}}\nContext 2: {{context_2}}\nAnswer 'Yes' or 'No'<br>Context 1: {{context_1}}\nContext 2: {{context_2}}\nDo the two contexts convey the same meaning for the word '{{target_word}}'? (Yes/No) |

## H.6 Intent Classification Templates

### H.6.1 MASSIVE Templates

| Task | Instruction Template |
|---|---|
| MASSIVE | {{utt}}\nWhat is the scenario of this utterance? (Choose from {{options_}})<br>Given the following utterance: {{utt}}\nWhat scenario does it belong to? (Select from {{options_}})<br>{{utt}}\nWhich scenario best describes this utterance? (Choose one from {{options_}})<br>What is the appropriate scenario for this utterance?\n{{utt}}\nSelect from the following scenarios: {{options_}}<br>Context: {{utt}}\nWhat scenario is being described? (Options: {{options_}})<br>Utterance: {{utt}}\nDetermine the scenario of this utterance from the following options: {{options_}}<br>{{utt}}\nIn which scenario does this utterance fit? (Choose from {{options_}})<br>Based on the following utterance, classify the scenario: {{utt}}\nPossible options: {{options_}}<br>{{utt}}\nIdentify the scenario that best fits this utterance (options: {{options_}})<br>What scenario does the following utterance belong to?\n{{utt}}\nAvailable options: {{options_}} |

## H.7 Commonsense Reasoning Templates

### H.7.1 Multilingual Fig-QA Templates

| Task | Instruction Template |
|---|---|
| Multilingual Fig-QA | Given the phrase: '{{startphrase}}', which of the following is correct?\n1. {{ending1}}\n2. {{ending2}}\nAnswer (1 or 2):<br>Based on the phrase: '{{startphrase}}', which answer is more accurate?\n1. {{ending1}}\n2. {{ending2}}\nSelect 1 for the first option, or 2 for the second:<br>Phrase: '{{startphrase}}'\nWhich of the following best matches the phrase?\nOption 1: {{ending1}}\nOption 2: {{ending2}}\nAnswer (1 or 2):<br>Here is a phrase: '{{startphrase}}'\nWhich statement is correct?\nA. {{ending1}}\nB. {{ending2}}\nChoose 1 for first, 2 for the second:<br>Context: '{{startphrase}}'\nWhich option is true?\n1. {{ending1}}\n2. {{ending2}}\nChoose the correct number (1 or 2):<br>The phrase is: '{{startphrase}}'\nWhich of the following is correct?\n1. {{ending1}}\n2. {{ending2}}\nRespond with 1 or 2:<br>For the phrase '{{startphrase}}', which answer is accurate?\n1. {{ending1}}\n2. {{ending2}}\nAnswer 1 if correct, or 2 if correct:<br>'{{startphrase}}'\nWhich one of the following is correct?\nOption 1: {{ending1}}\nOption 2: {{ending2}}\nChoose 1 or 2:<br>Given the statement: '{{startphrase}}', which of the following is the correct conclusion?\n1. {{ending1}}\n2. {{ending2}}\nAnswer (1 or 2):<br>Consider the phrase: '{{startphrase}}'\nWhich option is the right one?\n1. {{ending1}}\n2. {{ending2}}\nSelect your answer (1 or 2): |

### H.7.2 X-CSQA Templates

| Task | Instruction Template |
|---|---|
| X-CSQA | Question: {{stem}}\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nAnswer:<br>{{stem}}\n\nOptions:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nCorrect choice:<br>Read the question and select the correct answer:\n\n{{stem}}\n\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nAnswer:<br>What is the correct answer to this question?\n\n{{stem}}\n\nOptions:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nAnswer:<br>Based on the question below, choose the most accurate answer:\n\n{{stem}}\n\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nCorrect option:<br>Select the right option for the following question:\n\n{{stem}}\n\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nCorrect option:<br>Question:\n{{stem}}\n\nOptions:\n1. {{choice1}}\n2. {{choice2}}\n3. {{choice3}}\n4. {{choice4}}\n5. {{choice5}}\nCorrect option:<br>Answer the following question by selecting the best choice:\n\n{{stem}}\n\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nYour answer:<br>{{stem}}\n\nAnswer options:\n1. {{choice1}}\n2. {{choice2}}\n3. {{choice3}}\n4. {{choice4}}\n5. {{choice5}}\nChoose the correct one:<br>Below is a question and five possible answers. Choose the correct one.\n\nQuestion: {{stem}}\nChoices:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nE. {{choice5}}\nAnswer: |

### H.7.3 X-CODAH Templates

| Task | Instruction Template |
|---|---|
| X-CODAH | Which option makes sense?\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Select the most reasonable option:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Choose the most logical statement:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Which of the following is the most plausible?\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Pick the statement that makes the most sense:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Which statement is most consistent with common sense?\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>What makes the most sense in this context?\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Choose the statement that fits best:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Which option is the most reasonable?\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer:<br>Select the option that best aligns with common sense:\nA. {{choice1}}\nB. {{choice2}}\nC. {{choice3}}\nD. {{choice4}}\nAnswer: |

## H.8 Topic Classification Templates

### H.8.1 SIB-200 Templates

| Task | Instruction Template |
|---|---|
| SIB-200 | Given the following text, choose the correct category:\n\n{{text}}\n\nCategories:\n{{options_}}\nAnswer:<br>What is the topic of the following text?\n\n{{text}}\n\nPossible categories:\n{{options_}}\nAnswer:<br>Classify the following text into one of the categories:\n\n{{text}}\n\nCategories:\n{{options_}}\nAnswer:<br>Identify the correct category for the text below:\n\n{{text}}\n\nAvailable categories:\n{{options_}}\nAnswer:<br>Which category does the following text belong to?\n\n{{text}}\n\nCategories:\n{{options_}}\nAnswer:<br>Read the text and select its category:\n\n{{text}}\n\nCategories to choose from:\n{{options_}}\nAnswer:<br>Determine the most suitable category for the following text:\n\n{{text}}\n\nCategories:\n{{options_}}\nAnswer:<br>What is the correct classification of the following text?\n\n{{text}}\n\nPossible options:\n{{options_}}\nAnswer:<br>Choose the category that best describes the following text:\n\n{{text}}\n\nCategories:\n{{options_}}\nAnswer:<br>Given the text below, what is its main topic?\n\n{{text}}\n\nPossible categories:\n{{options_}}\nAnswer: |

## H.9 Paraphrase Detection Templates

### H.9.1 PAWS-X Templates

| Task | Instruction Template |
|---|---|
| PAWS-X | Are the following two sentences paraphrases of each other?\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAnswer (Yes or No):<br>Do these two sentences mean the same thing?\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAnswer (Yes or No):<br>Determine if the following sentences are paraphrases:\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAnswer (Yes or No):<br>Paraphrase detection task:\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAre they paraphrases? Answer: (Yes or No)<br>Do the sentences below convey the same meaning?\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nYour answer: (Yes or No)<br>Decide if the two sentences express the same idea:\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAnswer: (Yes or No)<br>Check if the two sentences are semantically equivalent:\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAnswer: (Yes or No)<br>Sentence similarity task:\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAre they the same in meaning? Answer: (Yes or No)<br>Based on the given sentences, are they paraphrases?\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nAnswer: (Yes or No)<br>Compare the following sentences:\n\nSentence 1: {{sentence1}}\nSentence 2: {{sentence2}}\nDo they convey the same meaning? Answer: (Yes or No) |

# I   Inference Time Comparison

We report the average examples/sec processed for each of the datasets in Table 7. It is important to note that all models are run on a single GPU, except for Meta-Llama-3-70B-Instruct and Llama-2-13B-chat which were run on 4 and 2 GPUs, respectively.

We also present the maximum number of samples each model can handle during inference, alongside the average time taken to process a single batch, all while fully utilizing a single GPU. These results, detailed in Table 8, provide a clear understanding of each model's efficiency in handling larger batch sizes under optimal GPU utilization.

| Model | BCOPA | MRPC | FigQA | Amazon Polarity | StoryCloze | YA Topic | Emotion | Avg |
|---|---|---|---|---|---|---|---|---|
| Qwen1.5-0.5B-Chat | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| phi-2 | 1.4 | 1.4 | 1.4 | 1.4 | 1.5 | 1.4 | 1.4 | 1.4 |
| Meta-Llama-3-70B-Instruct* | 2.9 | 1.3 | 3.2 | 0.9 | 4.9 | 1.1 | 2.1 | 2.3 |
| flan-t5-large | 8.2 | 13.2 | 13.2 | 13.2 | 13.2 | 13.2 | 13.2 | 12.5 |
| Llama-2-13b-chat-hf* | 8.7 | 5.7 | 12.8 | 4.3 | 15.7 | 4.4 | 6.9 | 8.3 |
| Our Approach (roberta-large) | 9.3 | 14.5 | 15.0 | 15.0 | 14.7 | 3.1 | 5.1 | 11.0 |
| bart-large-mnli | 9.7 | 14.1 | 14.0 | 14.2 | 14.1 | 13.7 | 13.8 | 13.4 |
| pythia-6.9b | 12.0 | 0.6 | 4.6 | 0.4 | 0.6 | 2.2 | 0.4 | 3.0 |
| Llama-2-7b-chat-hf | 12.5 | 0.6 | 4.6 | 0.4 | 0.6 | 2.3 | 0.5 | 3.1 |
| Mistral-7B-Instruct-v0.2 | 12.8 | 0.5 | 2.7 | 0.3 | 0.5 | 1.7 | 0.4 | 2.7 |
| pythia-2.8b | 13.6 | 16.7 | 24.9 | 15.2 | 27.2 | 15.1 | 20.9 | 19.1 |
| flan-t5-small | 13.9 | 39.2 | 39.1 | 39.3 | 39.4 | 39.3 | 39.3 | 35.6 |
| Our Approach (roberta-base) | 17.9 | 49.8 | 50.0 | 49.8 | 49.9 | 10.3 | 17.0 | 34.9 |

Table 7: The average examples per second processed by each model on each task. * indicates that the model required the use of more than one GPU.

| Model | Maximum Batch Size | Mean Inference Time Per Batch (s) |
|---|---|---|
| Qwen2-0.5B-Chat | 240 | 0.0696 |
| Qwen2-1.5B-Chat | 118 | 0.0580 |
| aya-23-8B | 36 | 0.3729 |
| gemma-2-2B | 72 | 0.3473 |
| gemma-2-9B | 18 | 0.5682 |
| Meta-Llama-3.1-8B | 36 | 0.2415 |
| Our Approach (mdeberta-base) | 732 | 0.0270 |

Table 8: The maximum number of samples each model can handle during inference while fully utilizing GPU memory (Nvidia A100 80GB).