ToPro: Token-Level Prompt Decomposition for Cross-Lingual Sequence Labeling Tasks

Bolei Ma[*♣,♡], **Ercong Nie**[*♣,♡], **Shuzhou Yuan**[♠], **Helmut Schmid**[♣], **Michael Färber**[♠], **Frauke Kreuter**[♣,♡,♢] **and Hinrich Schütze**[♣,♡]

[♣]LMU Munich, [♡]Munich Center for Machine Learning,
[♠]Karlsruhe Institute of Technology, [♢]University of Maryland, College Park

bolei.ma@lmu.de, nie@cis.lmu.de, shuzhou.yuan@kit.edu

Abstract

Prompt-based methods have been successfully applied to multilingual pretrained language models for zero-shot cross-lingual understanding. However, most previous studies primarily focused on sentence-level classification tasks, and only a few considered token-level labeling tasks such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging. In this paper, we propose **To**ken-Level **Pro**mpt Decomposition (**ToPro**), which facilitates the prompt-based method for token-level sequence labeling tasks. The ToPro method decomposes an input sentence into single tokens and applies one prompt template to each token. Our experiments on multilingual NER and POS tagging datasets demonstrate that ToPro-based fine-tuning outperforms Vanilla fine-tuning and Prompt-Tuning in zero-shot cross-lingual transfer, especially for languages that are typologically different from the source language English. Our method also attains state-of-the-art performance when employed with the mT5 model. Besides, our exploratory study in multilingual large language models shows that ToPro performs much better than the current in-context learning method. Overall, the performance improvements show that ToPro could potentially serve as a novel and simple benchmarking method for sequence labeling tasks.[1]

1 Introduction

As multilingual pretrained language models (MPLMs) continue to evolve (Devlin et al., 2019; Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021; Shliazhko et al., 2022), zero-shot cross-lingual transfer methods are gaining increasing popularity within the multilingual NLP domain (Lauscher et al., 2020; Nie et al., 2023a).

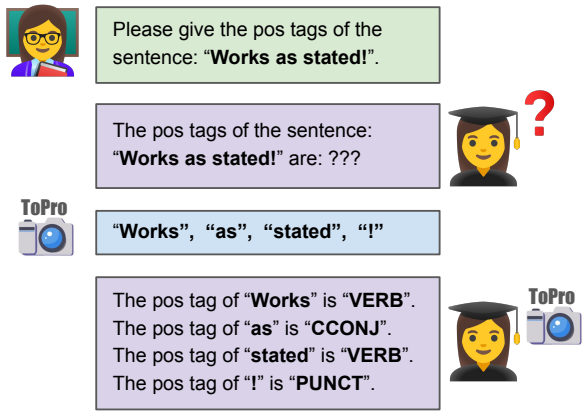

Figure 1: ToPro as a token-level prompting method for sequence labeling tasks. It decomposes the input sentence into single tokens and applies the prompt template to each token, inspired by human step-by-step logical thinking when solving this kind of task.

In light of the limited availability of training data in many low-resource languages, prior research (Artetxe et al., 2020; Hu et al., 2020) employed zero-shot cross-lingual transfer learning by fine-tuning an MPLM on a high-resource language such as English, and then directly applying the fine-tuned system to low-resource languages.

Prompt-based learning (Schick and Schütze, 2021a,b,c) is steadily garnering traction in recent NLP research. Prompt-based methods reformulate downstream tasks as language modeling tasks by using prompts comprising a template and a set of label words. The prompt can be either discrete in a textual format or continuous, performing prompting directly in the embedding space of the model (Liu et al., 2023). Much recent work highlights that applying prompt-based fine-tuning to MPLMs enables better zero-shot cross-lingual transfer performance (Zhao and Schütze, 2021; Huang et al., 2022; Nie et al., 2023b; Zhou et al., 2023). However, they focus on sentence-level clas-

[*] Equal contribution.

[1]Our source code is available in this GitHub repository: https://github.com/boleima/ToPro.

sification tasks such as sentiment analysis (Keung et al., 2020), XNLI (Conneau et al., 2018), and paraphrase detection (Zhang et al., 2019). Token-level sequence labeling tasks like Named Entity Recognition (NER) and Part-of-Speech (POS) Tagging rarely benefit from the advantages of prompt-based fine-tuning, primarily due to the intricate challenge of devising an appropriate prompt template.

To enhance the applicability of prompt-based learning to token-level sequence labeling tasks, we introduce the **To**ken-Level **Pro**mpt Decomposition (**ToPro**) method. ToPro splits the input sentence into tokens and creates a separate prompt for each token which asks for its label, following the human step-by-step logical thinking when solving these tasks, as shown in Figure 1. The evaluation on NER and POS tagging tasks shows that the ToPro-based fine-tuning achieves stronger zero-shot cross-lingual transfer performance than Vanilla fine-tuning and Prompt-Tuning, especially for languages that are typologically different from the source language (English).

Besides fine-tuning, in-context learning (ICL) is another common paradigm for applying large language models (LLMs) to downstream tasks. ICL prompts the LLM with few-shot demonstrations and/or natural language instructions to tackle a new task without updating any parameters (Brown et al., 2020; Min et al., 2022; Wei et al., 2023). ICL using instructions without demonstrations has also been applied to zero-shot cross-lingual transfer with multilingual large language models (MLLMs) (Muennighoff et al., 2023; Shi et al., 2023). Nevertheless, akin to cross-lingual fine-tuning, existing methodologies encounter limitations when applied to sequence labeling tasks (Ahuja et al., 2023; Asai et al., 2023). To this end, we delve deeper into the integration of ToPro with MLLMs and ascertain its effectiveness in an ICL paradigm. Empirical findings substantiate the potential of ToPro as a benchmarking method for evaluating the performance of MLLMs in sequence labeling tasks.

To sum up, our contributions are as follows:
(1) We propose a novel and simple method, called ToPro, which improves zero-shot cross-lingual transfer in token-level sequence labeling tasks by taking advantage of prompt-based learning.
(2) We substantiate the strength of ToPro in zero-shot cross-lingual fine-tuning through evaluations on NER and POS tagging tasks. Our method not only outperforms baselines for over 40 languages but also demonstrates efficacy in zero-shot English ICL, making it a promising benchmarking method for MLLMs in token-level tasks.
(3) We conduct a thorough cross-lingual analysis, revealing that ToPro exhibits particularly strong performance for languages that are typologically different from the source language English.

2 Related Work

MPLMs and Cross-Lingual Transfer The progress in MPLMs has established them as the basis for cross-lingual transfer. MPLMs typically adopt the architecture of monolingual Transformer-based language models and are pretrained on extensive unlabeled multilingual corpora. Examples of MPLMs are mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), Glot500 (ImaniGooghari et al., 2023), etc. Empirical studies (Karthikeyan et al., 2020; Turc et al., 2021) have showcased the remarkable cross-lingual prowess of MPLMs which are fine-tuned on English training datasets, and then used to predict on test datasets in other languages. Several benchmark datasets such as XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021), and Taxi1500 (Ma et al., 2023b) have been created to assess the capabilities of multilingual models. The increasing popularity of prompt learning has drawn the attention of researchers towards prompt-based methods for cross-lingual transfer (Tu et al., 2022; Ma et al., 2023a). Diverging from previous studies centered on sentence-level classification tasks, our work applies prompt-based fine-tuning to token-level sequence labeling tasks.

MLLMs and In-Context Learning BLOOMZ and mT0 (Muennighoff et al., 2023) stand out as two representative multilingual models in the era of LLMs. Both are fine-tuned on the xP3 dataset which contains multi-lingual multi-task prompts. BLOOMZ is built upon BLOOM (Scao et al., 2022) while mT0 is built upon mT5 (Xue et al., 2021). Brown et al. (2020) demonstrated that LLMs like GPT-3 can acquire task-solving ICL abilities. The emergence of MLLMs opens up the possibility for conducting zero-shot cross-lingual ICL, as demonstrated by recent benchmarking efforts, for example MEGA (Ahuja et al., 2023) and BUFFET (Asai et al., 2023). However, current ICL methods using text-to-text prompting with a fixed output template for sequence labeling tasks "*consistently exhibit*

extremely poor performance" (Asai et al., 2023) when applied to MLLMs, failing to exploit their real cross-lingual transfer abilities. Contrary to that, our proposed ToPro method better reflects the potential of MLLMs on token-level tasks.

Prompt Methods for Sequence Labeling Tasks Although prompt-based methods proved useful in sentence-level classification tasks, they were seldom employed for token-level labeling tasks. Cui et al. (2021) applied template-based prompting methods to the BART model (Lewis et al., 2020) for NER tasks. Their method is rank-based. They generate a sentence for each possible label and compute the probabilities of all generated sentences for the prediction, which can be expensive to decode. Ma et al. (2022) proposed a template-free prompting method for few-shot NER, called entity-oriented LM fine-tuning. However, they adopt the span-based task formulation of NER, resulting in more complexity, while our proposed method applies to NER tasks in the IOB (Inside-Outside-Beginning) tagging format. Blevins et al. (2023) proposed a structural prompting method for sequence labeling tasks, which was adapted for multilingual benchmarking large language models in recent work (Ahuja et al., 2023).

3 ToPro for Fine-Tuning

Problem Formulation In prompt-based learning, there is a pattern-verbalizer pair (PVP) (Schick and Schütze, 2021a) consisting of (i) a *prompt pattern* which converts the input text into a cloze-style question with a mask token, and (ii) a *verbalizer* which maps the labels onto representative words from the LM's vocabulary. This aligns well with the nature of text classification tasks where one label is predicted based on the input text. As Figure 2 shows, the input text X of a sentiment analysis task can be reformulated with a prompt pattern $P(\cdot)$ into a prompted input representation $P(X) =$ " Works as stated! In summary, the product was [MASK]. " The prompt $P(X)$ is processed by the LM to determine the most likely verbalizer word in the masked position. The label corresponding to this verbalizer is the prediction which is evaluated against the gold standard.

However, in sequence labeling tasks, each token of the input should receive a label. Thus, it is not possible to apply this type of prompt pattern with one mask token directly for token classification.

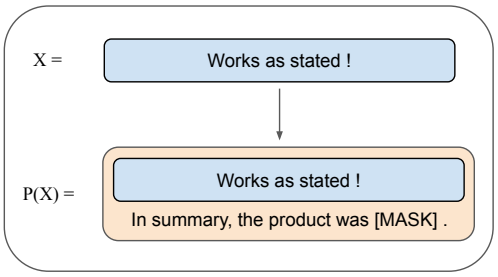

Figure 2: A prompt example for text classification.

Token-Level Prompt Decomposition (ToPro) When given such a token-level sequence labeling task, a human usually solves the task token by token. Inspired by this human process as well as the prompt design for sentence classification tasks, we propose a new prompting method ToPro for token classification which decomposes an input sentence into tokens and generates a series of prompts – one prompt for each token. Let $X = x_1, x_2, ..., x_m$ denote an input sentence consisting of m tokens. Our prompt generator function $P(T, X)$ generates m prompts by filling the template $T(\cdot, \cdot)$ with the sentence X and each of the tokens $x_1, x_2, ..., x_m$, respectively.

$$P(T, X) = \{T(X, x_1), ..., T(X, x_m)\} \quad (1)$$

Figure 3 shows the prompts generated by $P(T, X)$ for the input $X =$ "Works as stated !" and the template $T(X, x_i) =$ " X The POS tag of x_i is a kind of [MASK] ."

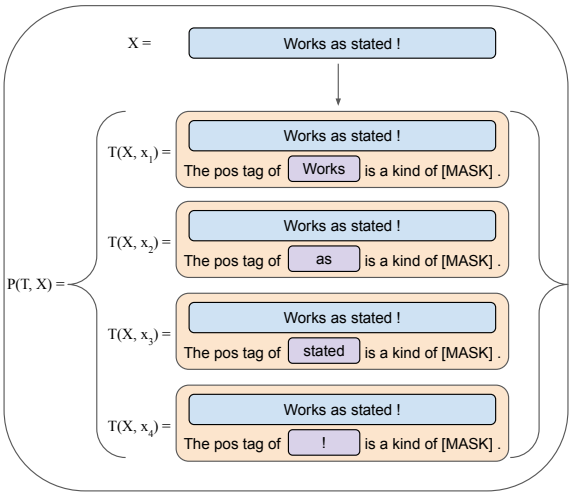

Figure 3: An example of ToPro framework for sequence labeling.

Prompt-Based Fine-Tuning and Cross-Lingual Transfer Following Ma et al. (2023a), we conduct prompt-based fine-tuning to evaluate our

ToPRO approach in a zero-shot cross-lingual context. Let $D = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ denote the set of training examples in the source language, where $X_1, ..., X_n$ are token sequences and $Y_1, ..., Y_n$ are tag sequences. Given $(X, Y) \in D$, the ToPRO function $P(T, X)$ reformulates the input sentence X into a set of cloze-style questions $\{T(X, x_1), ..., T(X, x_m)\}$ with masked tokens. The pretrained language model M with trainable parameters θ performs masked token prediction and returns the probabilities $p(\cdot) = M(T(X, x_i), \theta)$ of all candidate words for the masked token in the prompt $T(X, x_i)$. The verbalizer function $V(\cdot)$ is a bijective mapping from the set of class labels L to a set of verbalizers from the source language vocabulary. For each token, we predict the tag $\hat{y}$ whose verbalizer $V(\hat{y})$ receives the highest probability from model M:

$$\hat{y} = \arg\max_{y \in L} p(V(y)) \tag{2}$$

We fine-tune the parameters θ of model M by minimizing the cross-entropy loss function $\ell(D, \theta)$:

$$\ell(D, \theta) = -\sum_{(X,Y) \in D} \sum_{i=1}^{|Y|} \log M(T(X, x_i), \theta)(V(y_i)) \tag{3}$$

The fine-tuned model is used to predict the labels of the target language examples $\{X'_1, ..., X'_n\}$ using the same prompt pattern $T(\cdot, \cdot)$ and verbalizer $V(\cdot)$ as during fine-tuning . The best tags Y'_j for each example X'_j are predicted according to Eq. 2.

4 Experimental Setups

4.1 Datasets and Prompt Designs

We choose the following two representative datasets for sequence labeling tasks:

PAN-X (Pan et al., 2017), also called WikiANN, is a multilingual NER dataset based on Wikipedia articles including 282 languages. In our work, we use the subset of 48 languages which is part of the XTREME benchmark (Hu et al., 2020) to facilitate comparisons with related work.

For each token x_i of an input sequence X, we use the following prompt template $T(X, x_i)$:

> $T(X, x_i) = X \circ$ " The named entity of " $\circ\ x_i \circ$ " is a kind of: [MASK]."

The PAN-X dataset is annotated with location (LOC), person (PER), and organization (ORG) in IOB2 format. These labels are difficult to understand for the language model. Therefore, we replace them with real words and train the model to predict those instead. However, the model can only predict single words from its vocabulary as fillers for the [MASK] position. So, we choose the replacement words from the model's vocabulary.

As IOB2 annotates the beginning of a name and its remaining tokens with different tags, we use a word and its hyponym to represent the beginning of a name and its remaining tokens, respectively. For instance, we use the hypernym "location" for the beginning of the LOC and the hyponym "place" for the other words which should be semantically inside of the term "location". The verbalizer function $V(\cdot)$ for tag set Y is defined as follows:

> V(B-LOC) = location V(I-LOC) = place
> V(B-ORG) = organization V(I-ORG) = body
> V(B-PER) = person V(I-PER) = name
> V(O) = other

UDPOS This dataset was extracted from the Universal Dependency treebanks (Zeman et al., 2019). It contains 38 languages and is part of the XTREME benchmark (Hu et al., 2020).

Similarly as for the PAN-X dataset, we use the following prompt template $T(X, x_i)$ for token x_i of an input sequence X by paraphrasing the tags with semantically related words:

> $T(X, x_i) = X \circ$ " The pos tag of " $\circ\ x_i \circ$ " is a kind of: [MASK]."

The dataset has 14 labels. A detailed description of the tags can be found in Table 9 of the Appendix.

We define the verbalizer $V(\cdot)$ for the 14 tags as follows:

> V(ADJ) = modification V(ADP) = position
> V(ADV) = verbal V(AUX) = auxiliar
> V(CCONJ) = link V(DET) = determine
> V(INTJ) = mode V(NOUN) = thing
> V(NUM) = number V(PART) = functional
> V(PRON) = reference V(PROPN) = name
> V(PUNCT) = punct V(SCONJ) = condition
> V(SYM) = symbol V(VERB) = verb
> V(X) = other

We cannot select words like "adjective" and "adverb" which would better represent the meanings of the tags because the verbalizers have to come from the vocabulary of the PLM so that the masked language model is able to predict them as a single unit. Instead, we use semantically related words from the vocabulary as verbalizers.

4.2 Baselines

We compare our approach with the following baselines:

Vanilla Fine-Tuning (Vanilla) The vanilla fine-tuning method predicts the token labels through the hidden embeddings of each token in the output layer without using a prompt pattern. We use the cross-entropy loss as the objective function for fine-tuning and `AdamW` for optimization with a learning rate of 1e-5. The fine-tuned models are used to predict the test data.

Prompt-Tuning (PT) Prompt-Tuning only trains a small number of parameters, e.g., a continuous prompt or a task classifier (Lester et al., 2021; Liu et al., 2022). We implement the prompt-tuning method of Tu et al. (2022) for zero-shot cross-lingual transfer by tuning the prefix prompts and layer prompts for the two sequence labeling tasks.

4.3 Multilingual Models

The following MPLMs from the HuggingFace Transformers library (Wolf et al., 2020) are applied in our main experiments:

- **Encoder-Only Models:** including the multilingual BERT model (Devlin et al., 2019) `bert-base-multilingual-cased` (B) and the XLM-R model (Conneau et al., 2020) `xlm-roberta-base` (X), and
- **Encoder-Decoder Model:** the multilingual T5 model (Xue et al., 2021) `mt5-base` (T). We also include the encoder-decoder model mT5 in our experiments, as we wish to explore the potential of ToPro with different types of models. Details on how we trained mT5 for sequence labeling tasks are shown in §A.2.

5 Results and Analysis

5.1 Main Results

Table 1 gives an overview of the average results[2] on PAN-X and UDPOS. We find that ToPro Fine-Tuning outperforms Vanilla Fine-Tuning and Prompt-Tuning obviously on both tasks in mBERT and XLM-R settings: On PAN-X, the performance is improved by **19.18%** and **25.16%** compared to Vanilla and Prompt-Tuning respectively when trained with mBERT, and by **18.73%** and **26.98%** with XLM-R. On PAN-X, the performance is improved by **5.27%** and **6.24%** compared to Vanilla and Prompt-Tuning respectively when trained with mBERT, and by **3.74%** and **4.3%** with XLM-R.

In the mT5 setting, the ToPro Fine-Tuning outperforms Vanilla Fine-Tuning on both tasks as well, namely by **28.63%** on PAN-X, and by **14.72%** on UDPOS. We notice that the mT5 model performs even better than the two encoder-only models and achieves SOTA performance[3], showing the potential of ToPro with different model types. We find that Prompt-Tuning does not work well with mT5, as it requires more training epochs for the model to achieve subtle performance improvements, necessitating even longer training time compared to the Vanilla baselines. One possible reason for this could be the limited number of trainable parameters in mT5 with Prompt-Tuning, as only 0.002% of the parameters are updated with our current prompt settings (see §A.2). We exclude the results of Prompt-Tuning for mT5 because the increased training resources do not align with the efficiency-focused goals of Prompt-Tuning as a training methodology.

When comparing performances on the two tasks generally, we notice that the performance shows greater improvement on PAN-X with all three models, indicating that the PAN-X for NER task has a greater improvement potential.

Model	Method	PAN-X	UDPOS
mBERT	Vanilla Fine-Tuning	62.73	70.89
	Prompt-Tuning	56.76	69.91
	ToPro Fine-Tuning	**81.91**	**76.16**
XLM-R	Vanilla Fine-Tuning	61.30	72.42
	Prompt-Tuning	53.05	71.86
	ToPro Fine-Tuning	**80.03**	**76.16**
mT5	Vanilla Fine-Tuning	64.19	71.38
	Prompt-Tuning	-*	-*
	ToPro Fine-Tuning	**92.82**	**86.11**

Table 1: Overview of average results on PAN-X and UDPOS. ∗: The results of PT with mT5 are excluded from the comparison as the F1 scores are 0 for the current parameter settings.

[2]Since we are interested in the zero-shot cross-lingual transfer performance, we do not include the English results in the average performance. Our evaluation metric is the weighted average F1-score.

[3]Based on the evaluation results available at https://sites.research.google/xtreme/dataset, as of Jan. 23, 2024, the SOTA performance in structured prediction, calculated as the mean value of PANX-X and UDPOS, is 84.6. Our mT5 model, when used with ToPro, achieves an impressive score of 89.47.

langs	B (Vanilla)	B (PT)	X (Vanilla)	X (PT)	T (Vanilla)
en	8.96	13.71	10.90	16.27	19.38
af	12.81	19.50	15.00	20.10	19.82
ar	18.52	23.10	20.57	24.09	39.14
az	17.73	21.83	22.65	25.46	32.78
bg	11.29	16.33	11.17	16.05	23.13
bn	8.24	19.52	3.10	18.65	30.42
de	13.30	18.26	17.15	23.13	21.01
el	18.03	26.54	16.28	27.09	19.34
es	10.99	16.88	13.13	18.42	26.08
et	12.11	16.23	17.53	22.83	21.55
eu	19.91	24.36	26.52	36.62	27.50
fa	27.10	34.67	14.09	24.17	47.48
fi	12.51	17.24	15.29	20.42	20.78
fr	6.75	12.13	10.39	17.06	20.87
gu	33.33	55.16	30.99	40.57	31.99
he	27.47	31.26	30.94	38.85	24.09
hi	12.70	18.50	11.18	18.71	30.79
hu	14.76	20.04	14.96	21.21	22.97
id	16.79	19.60	21.31	24.03	26.95
it	10.15	13.13	11.78	17.81	19.03
ja	41.04	45.53	47.61	49.89	43.52
jv	18.70	23.05	16.43	32.80	22.80
ka	19.31	25.80	20.48	30.28	25.85
kk	33.74	34.89	42.36	42.48	28.63
ko	22.33	25.43	31.42	37.05	33.16
lt	13.58	18.13	14.24	21.01	23.48
ml	26.57	32.21	25.70	34.47	32.55
mr	25.15	31.75	21.01	33.32	31.70
ms	14.50	18.38	8.25	28.53	17.64
my	29.29	39.47	31.69	40.16	48.96
nl	10.12	14.57	12.32	17.12	19.50
pa	25.38	28.31	19.42	35.89	32.47
pl	10.12	13.49	13.02	17.62	21.03
pt	7.48	13.25	9.15	15.87	23.98
qu	12.97	31.44	17.08	32.22	25.16
ro	7.91	22.31	13.15	24.11	25.06
ru	19.37	26.56	18.10	25.80	27.92
sw	9.25	18.76	7.81	19.75	25.30
ta	24.48	28.73	26.68	33.46	29.84
te	32.97	36.06	36.53	43.85	28.14
th	67.60	67.84	16.48	15.89	50.10
tl	11.40	11.00	8.52	16.21	27.06
tr	12.63	20.13	13.77	24.87	26.93
uk	14.63	20.74	12.30	24.53	23.51
ur	29.96	36.69	1.63	22.93	51.31
vi	16.35	18.87	14.26	20.49	31.65
yo	15.41	27.00	16.13	30.82	23.30
zh	24.88	27.66	40.81	41.58	39.50
avg.	19.18	25.16	18.73	26.98	28.63

Table 2: Performance difference (−) of ToPRO to Vanilla or Prompt Tuning (PT) with mBERT (B), XLM-R (X) and mT5 (T) on PAN-X.

langs	B (Vanilla)	B (PT)	X (Vanilla)	X (PT)	T (Vanilla)
en	0.54	0.87	0.41	0.87	7.90
af	3.27	3.31	2.00	1.96	7.16
ar	16.51	14.43	4.65	4.37	15.23
bg	2.80	2.63	0.56	0.69	14.34
de	3.11	3.44	1.58	1.85	12.48
el	3.80	4.86	-0.49	-0.64	13.06
es	-0.86	0.89	-1.23	-0.89	5.73
et	3.86	7.90	0.94	1.94	11.58
eu	10.11	8.87	1.88	5.24	14.14
fa	2.23	1.42	0.82	1.47	15.12
fi	2.63	5.21	0.48	1.21	11.68
fr	-0.16	4.76	-5.36	-5.00	8.41
he	24.40	24.55	14.12	14.38	23.68
hi	9.61	8.62	3.63	3.80	18.68
hu	0.56	1.08	-2.07	-1.82	13.90
id	4.63	4.83	4.10	4.43	13.81
it	-2.01	-0.33	-1.25	-2.35	8.28
ja	5.06	5.34	29.16	32.61	27.31
kk	4.45	4.79	0.46	1.67	15.61
ko	13.85	13.04	11.39	10.85	25.57
lt	3.75	6.61	2.49	4.05	12.81
mr	5.39	8.51	-2.52	-1.13	17.19
nl	0.50	1.01	0.29	0.60	9.15
pl	3.20	3.82	1.84	1.50	13.76
pt	-1.07	-0.66	-0.79	-0.75	7.96
ro	3.15	4.01	1.46	1.89	14.07
ru	4.20	3.44	1.54	2.19	11.36
ta	13.43	12.89	10.85	10.88	20.38
te	-0.93	1.45	-0.59	1.68	12.00
th	15.83	19.92	25.28	29.21	15.61
tl	-0.59	4.12	-4.68	-6.53	19.38
tr	1.82	5.11	-0.37	1.11	16.88
uk	5.91	5.62	1.95	2.32	13.51
ur	12.04	11.07	7.94	6.14	20.46
vi	2.79	4.08	1.30	2.44	20.61
wo	2.45	4.19	-10.70	-9.42	-0.88
yo	5.74	7.79	-6.59	-5.54	5.97
zh	9.56	8.29	44.30	42.58	38.53
avg.	5.27	6.24	3.74	4.30	14.72

Table 3: Performance difference (−) of ToPRO to Vanilla or PT with mBERT (B), XLM-R (X) and mT5 (T) on UDPOS.

5.2 Cross-Lingual Transfer Analysis

The detailed results of the cross-lingual transfer performance of ToPRO compared to the baselines for each target language are documented in Table 11 and Table 12 in §A.6 of the appendix. Table 2 and Table 3 above show the performance improvements of ToPRO compared to the baselines for each language. Overall, ToPRO-based Fine-Tuning outperforms Vanilla Fine-Tuning and Prompt-Tuning on average. However, we can notice individual performance differences between the languages.

On both tasks, we find that the performance gain of ToPRO for English (en) is among the lowest across all languages. Since English is the language on which the models have been fine-tuned, we conclude that ToPRO is particularly effective in cross-lingual zero-shot scenarios. The reason could be that the models are only fine-tuned on the English dataset. Therefore, ToPRO's potential performance improvement is smaller for English than for other languages.

On PAN-X, ToPRO outperforms Vanilla and Prompt-Tuning across all target languages, with some language-independent variations. The improvements in languages such as Persian (fa), Gujarati (gu), Hebrew (he), Japanese (ja), Kazakh (kk), Burmese (my), Telugu (te), Thai (th), Urdu (ur), and Chinese (zh) are above the average. All these languages are from different language groups to English and have different writing systems. We

Input sequence & Its Gloss & Tags (True, Vanilla, ToPro) & Translation
Case 1 (zh)
Input: 温暖 海岸 是 指 西班牙 穆爾西亞 自治 區 長 約 250 公里 的 地中 海 海岸 線 。
Gloss: Warm coast be refer Spain Murcia autonomous region long approximately 250 km 's Mediterranean sea coast line .
True: propn noun verb verb propn propn verb part adj adv num noun part propn part noun part punct
Vanilla: propn punct punct punct punct propn punct punct punct punct num num punct noun noun punct punct punct (0.19 F1)
ToPro: propn propn aux verb propn propn propn propn adj adv num noun adp noun noun propn noun punct (0.47 F1)
Translation: The Costa Cálida is the 250-kilometer-long Mediterranean coastline of the Autonomous Region of Murcia, Spain.
Case 2 (ja)
Input: 政府 が もっと 宗教法人 の 定義 を 厳しく し , こういう 団体 は 排除 す べき 。
Gloss: government subject more religious organisation of definition object strictly that , such association topic exclude do must .
True: noun adp adv noun adp noun adp adj aux punct adj noun adp verb aux aux punct
Vanilla: det punct punct punct punct noun punct punct punct punct punct punct punct verb punct punct punct (0.29 F1)
ToPro: noun part adv noun part noun pron adj verb punct adj noun pron verb verb adj punct (0.64 F1)
Translation: "The government should tighten the definition of religious corporations and eliminate such organizations."
Case 3 (de)
Input: „ Mich interessiert etwas , wenn es mich zu Teilnahme zu erregen weiß ", sagte er in einem deutschen Fernsehen .
Gloss: " me interest something , if it me to attendance to irritate knows " , said he in a German television .
True: punct pron verb pron punct sconj pron pron adp noun part verb verb punct punct verb pron adp det adj noun punct
Vanilla: punct pron verb pron punct sconj pron pron adp noun part verb verb punct punct verb pron adp det adj noun punct (1.00 F1)
ToPro: punct pron verb pron punct sconj pron pron part noun part verb verb punct punct verb pron adp det adj noun punct (0.95 F1)
Translation: "I am interested in something if it knows how to excite me to participate", he said on German television."
Case 4 (nl)
Input: „ We hebben een concept nodig voor verandering ", zei Djindjic in een interview met de Duitse televisie .
Gloss: " we have a concept necessary for change " , said Djindjic in a interview with the German television .
True: punct pron verb det noun adj adp noun punct punct verb propn adp det noun adp det adj noun punct
Vanilla: punct pron verb det noun verb adp noun punct punct verb propn adp det noun adp det adj noun punct (0.95 F1)
ToPro: punct pron verb det noun verb adp noun punct punct verb propn adp det noun adp det adj noun punct (0.95 F1)
Translation: "We need a concept for change", Djindjic said in an interview with German television.

Table 4: Comparison of the output of ToPro and Vanilla for selected UDPOS examples with XLM-R. The interesting tokens and their tags are marked red. The sentences were translated into English using www.deepl.com.

can conclude that the performance improvement of ToPro is particularly high for languages that differ a lot from English, further indicating the cross-lingual ability of our prompt-based method.

On UDPOS, ToPro outperforms Vanilla and PT in most of the languages, although there are some languages for which ToPro performs slightly worse and the overall performance gain is not as high as on PAN-X. Typically, the improvements for languages such as Arabic (ar), Basque (eu), Hebrew (he), Korean (ko), Tamil (ta), Thai (th), Urdu (ur), and Chinese (zh) are above average. The improvements over Vanilla in Chinese reach 44.3% and 38.53% for XLM-R and mT5, respectively, and the improvement over PT in Chinese is 42.58%.

Overall, the results show that ToPro outperforms Vanilla and PT on both sequence labeling tasks, indicating that the ToPro method has a better ability to transfer knowledge cross-lingually. And the NER performance is even better than the performance for POS tagging. When analyzing the performances for individual languages, we find that ToPro has a strong performance for zero-shot cross-lingual transfer, particularly in languages with low similarity to English and different writing systems. The prompt-based approach seems to mitigate the language barriers and facilitate cross-lingual transfer. Additionally, the results vary across target languages, highlighting the importance of language typology and writing systems in determining the effectiveness of ToPro.

5.3 Error Analysis

In this section, we analyze selected instances from the UDPOS task with typical annotation errors by the models in Table 4.

The first example is a sentence in Chinese (zh), which is typologically and orthographically quite different from the training language English. The first two tokens marked red in this example 是 ("be") 指 ("refer to") are a pair of verbs, a so-called double-verb structure. They are both predicted by Vanilla as PUNCT (punctuation), but by ToPro

as AUX (auxiliary) and VERB which are quite close to the corrected tags, as the auxiliary itself is a special kind of verb. The tokens 長 ("long") 約 ("around") are predicted by Vanilla again as PUNCT, but correctly by ToPro as ADJ and ADV. Moreover, the tokens 海岸 ("coast") 線 ("line") are predicted by Vanilla still as PUNCT, and by ToPro as PROPN (proper noun) and NOUN, which, though not the same as the original tags NOUN and PART (particle), are already close to the original tags as they are all a kind of noun. In this case, we notice that the Vanilla model tends to predict PUNCT for the majority of the tokens, whereas the ToPro method often predicts the correct tags or at least semantically related tags.

The second example is in Japanese (ja), which is also typologically and orthographically quite different from English. The first token 政府 ("government") is predicted by Vanilla as DET (determiner) which is somehow close to its original tag, and correctly by ToPro as NOUN. The token もっと ("more") is predicted by Vanilla as PUNCT, but correctly by ToPro as ADV. The token し ("that") is originally AUX, and it is predicted by Vanilla again as PUNCT, but by ToPro as VERB, which is already close to the meaning of AUX. And the token pair 排除 ("exclude") す ("do") has original labels VERB AUX, and is predicted by ToPro as VERB VERB, which are still very close to their original labels. However, Vanilla predicts them again as PUNCT PUNCT. Similar to the Chinese example, the Vanilla method tends to predict PUNCT for unfamiliar tokens, whereas ToPro generates the correct tags or at least tags close to the correct tags.

The third example is an input sentence in German, which is very close to the English language. We reach therefore very high F1 scores both with Vanilla and ToPro approaches. Noticeably, in this example, there is one token zu ("to") with two different kinds of POS: ADP and PART. The Vanilla correctly detects the difference, while ToPro classifies both tokens as PART. This is a shortcoming of TOPRO's token-wise prompting strategy which generates identical prompts for both occurrences of "zu".

The fourth example is a Dutch (nl) input sentence. Dutch is also closely related to English. We reach a high F1 score of 0.95 with both Vanilla and ToPro approaches. The two approaches make the same error by predicting the token nodig ("necessary") as VERB, which should be ADJ.

In conclusion, the first two examples show that ToPro works much better than Vanilla for languages that are typologically different from the source language of training. And even when predicting false POS tags, ToPro tends to predict tags semantically close to the correct tags. The last two examples show the slightly worse performance of ToPro for languages that are close to the source language of training. These findings support our claim in §5.2.

5.4 Exploratory Study in MLLMs

Previous benchmarking work on MLLMs has encountered difficulties in the sequence labeling tasks. The BUFFET work by Asai et al. (2023) evaluates the multilingual capability of large generative language models across a wide variety of tasks. The benchmark contains NER tasks using the PAN-X test dataset. They employ a text-to-text prompting method to define the prompts for sequence labeling tasks. In this method, a model extracts all named entities with named entity tags from the input sentences. The target output string has to follow a fixed template. If the actual output produced by the system during evaluation does not exactly follow this format, it cannot be evaluated against the gold standard.

As Table 5 shows, the text-to-text prompting approach works badly for MLLMs. The BUFFET work evaluated with the MLLMs BLOOMZ and mT0, but achieved very poor performance. The performance of the mT0 model was reported as 0.0, suggesting that this prompting approach fails to work for mT0. Therefore, such text-to-text prompting as a benchmarking method cannot properly reflect the cross-lingual capability of currently existing MLLMs for sequence labeling tasks.

	BUFFET	ToPro (Ours)
bloomz-7b1	7.6*	13.98
mt0-xxl	0.0*	18.09

Table 5: Results of bloomz-7b1 and mt0-xxl models on the PAN-X dataset. ∗: Values were directly extracted from the original paper.

In contrast, our proposed ToPro prompting method performs much better as Table 5 shows. We apply the ToPro prompt to MLLMs in a zero-shot ICL paradigm without parameter updating. The results of our exploratory study in MLLMs indicate that, compared to the current commonly

used prompting method in MLLM benchmarking work, ToPro is a more promising prompting one for MLLMs in sequence labeling tasks. Detailed information on the experimental settings and results regarding ToPro applied to MLLMs is provided in §A.5.

6 Conclusion

In our work, we introduce ToPro for token-level sequence labeling tasks, a novel and simple method that adopts the basic framework of prompting from sentence classification tasks and applies the prompt template to each token in a sentence. We evaluate the ToPro-based fine-tuning for zero-shot cross-lingual transfer and compare it to Vanilla fine-tuning and Prompt-Tuning baselines. We apply ToPro with three MPLMs on two representative sequence labeling tasks: NER and POS tagging. Our experiments show that ToPro outperforms the baselines with the MPLMs and achieves SOTA performance with mT5. We further discovered that the performance improvement of ToPro is generally more obvious in the cross-lingual context, especially for languages that are linguistically very different from the source language English, highlighting its cross-lingual ability. Additionally, we applied the ToPro method to MLLMs and noticed better performances of ToPro as well, compared to existing benchmarking work. Overall, ToPro shows a noticeable performance improvement and could serve as a potential benchmark for sequence labeling tasks for future studies in prompt-based learning.

Limitations

Our method has a shortcoming which we observed in the third error example in §5.3: ToPro generates identical prompts when a token occurs multiple times in a sentence. This results in errors unless all occurrences have the same goldstandard label. We plan to explore in future work whether marking the currently annotated token in the input sentence solves this problem.

While ToPro outperforms the baseline methods, its training takes much longer because the model is prompted with each token. For example, the Vanilla fine-tuning on PANX with mBERT takes around 30 minutes for 5 epochs, while the ToPro fine-tuning takes around 125 minutes, i.e. training is 4 times slower. Future work could pay more attention to improving the efficiency of the method.

Furthermore, the current study only considers prompt patterns and verbalizers that have been manually designed by the authors. Future work could focus on methods that automatically find a suitable prompt. Also, dynamic prompt applications could be taken into account, to look for a best-performing prompt.

Ethics Statement

This research was conducted in accordance with the ACM Code of Ethics. Both datasets (Pan et al., 2017; Zeman et al., 2019) that we use are publicly available. We report only aggregated results in the main paper. We have not intended or do not intend to share any Personally Identifiable Data with this paper.

Acknowledgements

We want to thank the anonymous reviewers for their invaluable contributions and constructive feedback. This work was partially supported by funding from BERD@NFDI, Munich Center for Machine Learning (MCML), and China Scholarship Council (CSC).

Besides, we express our heartfelt gratitude to Mengmeng Xu for her efforts in designing the ToPro logo: 

.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

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.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 6649–6663, Toronto, Canada. 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

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

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 (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

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

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

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

Bolei Ma, Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023a. 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.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Ehsaneddin Asgari, and Hinrich Schütze. 2023b. Taxi1500: A multilingual dataset for text classification in 1500 languages. *arXiv preprint arXiv:2305.08487*.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang. 2022. Template-free prompt tuning for few-shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5721–5732, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023a. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2023b. Unleashing the multilingual encoder potential: Boosting zero-shot performance via probability calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15774–15782, Singapore. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla A. Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Sai-

ful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Francois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, A. Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Timo Schick and Hinrich Schütze. 2021a. 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.

Timo Schick and Hinrich Schütze. 2021b. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021c. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

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.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

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

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

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin

Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2023. Enhancing cross-lingual prompting with dual prompt augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11008–11020, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Hyperparameter Settings

To avoid random effects on training, we trained each experiment with 5 different random seeds $\{10, 42, 421, 510, 1218\}$ and took the average results. The applied hyperparameters for the three models are documented in Table 6. As Prompt-Tuning only tunes a few parameters, we increase the training epochs by 10 times more than Vanilla and ToPro Fine-Tuning.

Hyperparameter	B & X	T
epochs	5 (PT: 50)	10 (PT: 100)
learning_rate	1e-5	3e-5
batch_size	8	24
grad_acc_steps	4	4
max_seq_length	128	128
max_target_length	-	150
num_beam_search	-	3

Table 6: Hyperparameters

A.2 mT5 Training

To align with the text-to-text transformer format, we have redefined the output structure for both NER and POS tasks, drawing inspiration from the prior work of mT5 (Xue et al., 2021). For the input text, we introduce task descriptions as prompts, specifically "NER tagging:" for the PAN-X dataset and "POS tagging:" for the UDPOS dataset. Regarding the target text, we append tags to each token and insert delimiters between tokens to create a coherent sequence of text. The following example illustrates our preprocessing procedure using a sample from the UDPOS dataset for Vanilla fine-tuning:

> - **Input text:** POS tagging: On the other hand, it looks pretty cool .
> - **Target text:** ADP: On $$ DET: the $$ ADJ: other $$ NOUN: hand $$ PUNT: , $$ PRON: it $$ VERB: looks $$ ADV: pretty $$ ADJ: cool $$ PUNT: .

As for the ToPro method, we use the same prompt pattern as for encoder-only models:

> - **Input text:** On the other hand, it looks pretty cool . The pos tag of On is:
> - **Target text:** ADP

A.3 Dataset Statistics

In Table 7 we show a basic statistic view of the PAN-X (Pan et al., 2017) and UDPOS (Zeman et al., 2019) datasets from the XTREME benchmark (Hu et al., 2020). We use the original train-dev-test split from the datasets based one the XTREME benchmark. For training and validation,

we use the English train and dev dataset, and for test we use the test sets of all languages.

Task	Size			#Labels
	| Train |	| Dev |	| Test |	
PAN-X	20 000	10 000	10 000	7
UDPOS	21 253	3 974	3 973	17

Table 7: Overview of the three datasets. Train and dev data size refers to the number of samples (sentences) for English. Test data size refers to the average number of samples for each target language.

A.4 Tags and Meanings

The tags of the two datasets and their detailed meanings are documented in Table 8 and Table 9.

Tags	Meaning
B-LOC	location (beginning)
B-ORG	organization (beginning)
B-PER	person (beginning)
I-LOC	location (inside)
I-ORG	organization (inside)
I-PER	person (inside)
O	other

Table 8: IOB2 tags

Tags	Meaning
ADJ	adjective
ADP	adposition
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
INTJ	interjection
NOUN	noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

Table 9: Universal POS tags

A.5 Details of Experiments on MLLMs

We apply our ToPro method to two MLLMs including:

- BLOOMZ (Muennighoff et al., 2023), a multi-task fine-tuned version of the BLOOM (Scao et al., 2022) model created by the BigScience community. We use `bloomz-7b1`, the version with 7.1 billion parameters, in our experiment.
- mT0 (Muennighoff et al., 2023), a multi-task fine-tuned version of the mT5 (Xue et al., 2021) model. We use parameters `mt0-xxl`, the version with 13 billion, in our experiment.

In our exploratory study with MLLMs, we employ zero-shot English ICL without demonstrations (Shi et al., 2023). The prompt template used for MLLMs is as follows:

> $T(X, x_i) =$
> Named entity type: location organisation person place body name other
> Sentence: X
> Named entity type of x_i in the sentence is

The full experimental results of the two MLLMs on the PAN-X task are shown in Table 10.

A.6 Detailed Results

We present the detailed results of the cross-lingual evaluation performance of Vanilla, Prompt Tuning, and ToPro in Table 11 (PAN-X) and Table 12 (UDPOS).

Lang.	bloomz-7b1	mt0-13b
en	14.81	17.48
af	9.71	15.97
ar	20.63	19.37
az	10.00	17.25
bg	11.66	20.23
bn	23.27	26.68
de	10.96	15.32
el	8.70	14.11
es	17.20	20.70
et	12.16	18.12
eu	11.26	17.86
fa	19.51	20.08
fi	11.77	19.19
fr	20.55	20.11
gu	6.09	13.33
he	10.01	16.85
hi	17.98	23.10
hu	11.57	17.57
id	16.85	21.13
it	16.50	17.74
ja	7.58	4.79
jv	13.66	18.11
ka	8.11	18.23
kk	9.95	18.90
ko	11.32	19.10
lt	13.48	17.81
ml	13.37	22.53
mr	14.53	19.92
ms	22.08	19.10
my	3.37	18.59
nl	14.80	17.70
pa	12.74	17.14
pl	14.04	17.89
pt	19.35	20.46
qu	16.50	17.49
ro	21.19	20.18
ru	12.48	18.36
sw	17.02	23.72
ta	13.06	19.52
te	11.82	17.15
th	7.65	0.57
tl	25.10	22.20
tr	13.62	19.53
uk	11.74	19.69
ur	18.63	22.24
vi	16.86	17.05
yo	19.73	22.21
zh	6.86	5.25
avg.	**13.98**	**18.09**

Table 10: Full results of MLLMs on the PAN-X task.

lang.	en	af	ar	az	bg	bn	de	el	es	et	eu	fa	fi
B (Vanilla)	83.83	78.07	44.09	67.58	78.31	70.09	79.10	71.85	73.95	77.96	65.44	42.43	78.74
B (PT)	79.09	71.37	39.52	63.47	73.28	58.81	74.14	63.35	68.05	73.84	61.00	34.86	74.01
B (ToPro)	**92.80**	**90.87**	**62.62**	**85.30**	**89.61**	**78.33**	**92.40**	**89.88**	**84.94**	**90.07**	**85.35**	**69.52**	**91.25**
X (Vanilla)	81.31	75.03	47.26	61.37	77.02	68.97	74.07	74.93	70.51	70.73	58.07	48.73	75.44
X (PT)	75.94	69.92	43.75	58.57	72.15	53.42	68.09	64.12	65.21	65.43	47.97	38.65	70.31
X (ToPro)	**92.21**	**90.02**	**67.84**	**84.02**	**88.20**	**72.06**	**91.22**	**91.22**	**83.63**	**88.26**	**84.59**	**62.82**	**90.72**
T (Vanilla)	77.14	76.94	49.99	62.00	72.98	60.32	76.19	76.88	67.81	74.25	67.12	40.46	75.93
T (ToPro)	**96.52**	**96.76**	**89.13**	**94.78**	**96.11**	**90.74**	**97.21**	**96.22**	**93.90**	**95.80**	**94.62**	**87.93**	**96.71**

lang.	fr	gu	he	hi	hu	id	it	ja	jv	ka	kk	ko	lt
B (Vanilla)	80.40	53.89	55.80	68.17	76.16	61.21	81.10	28.25	61.58	67.94	47.21	61.60	74.41
B (PT)	75.02	32.07	52.00	62.38	70.88	58.39	78.11	23.76	57.23	61.45	46.06	58.51	69.86
B (ToPro)	**87.15**	**87.22**	**83.27**	**80.88**	**90.91**	**77.99**	**91.24**	**69.29**	**80.28**	**87.25**	**80.95**	**83.94**	**87.99**
X (Vanilla)	75.81	57.12	51.54	68.11	76.42	48.04	77.58	19.26	57.86	67.02	40.79	50.36	73.85
X (PT)	69.14	47.54	43.64	60.58	70.17	45.33	71.55	16.98	41.49	57.22	40.66	44.73	67.08
X (ToPro)	**86.20**	**88.11**	**82.49**	**79.28**	**91.38**	**69.35**	**89.36**	**66.87**	**74.29**	**87.50**	**83.14**	**81.78**	**88.09**
T (Vanilla)	73.68	64.18	68.83	61.90	74.01	64.28	77.33	46.19	67.79	70.17	65.10	60.24	72.09
T (ToPro)	**94.55**	**96.17**	**92.93**	**92.69**	**96.98**	**91.22**	**96.35**	**89.71**	**90.59**	**96.02**	**93.73**	**93.40**	**95.57**

lang.	ml	mr	ms	my	nl	pa	pl	pt	qu	ro	ru	sw	ta
B (Vanilla)	56.00	57.77	67.05	53.36	82.23	34.29	80.74	79.77	64.53	73.97	65.33	70.08	53.33
B (PT)	50.35	51.17	63.17	43.18	77.78	31.36	77.38	74.00	46.06	59.57	58.14	60.57	49.08
B (ToPro)	**82.57**	**82.93**	**81.55**	**82.65**	**92.35**	**59.67**	**90.87**	**87.25**	**77.50**	**81.88**	**84.71**	**79.33**	**77.81**
X (Vanilla)	59.85	60.74	66.13	53.41	79.67	50.31	77.64	76.83	60.49	70.45	62.54	69.51	54.62
X (PT)	51.08	48.43	45.86	44.94	74.88	33.83	73.04	70.12	45.36	59.48	54.84	57.57	47.83
X (ToPro)	**85.55**	**81.75**	**74.39**	**85.10**	**92.00**	**69.72**	**90.66**	**85.99**	**77.57**	**83.60**	**80.65**	**77.32**	**81.30**
T (Vanilla)	62.21	61.71	68.06	44.70	77.43	53.71	75.31	70.83	62.18	69.10	66.16	66.60	62.69
T (ToPro)	**94.77**	**93.42**	**85.70**	**93.66**	**96.93**	**86.18**	**96.34**	**94.81**	**87.35**	**94.16**	**94.07**	**91.90**	**92.52**

lang.	te	th	tl	tr	uk	ur	vi	yo	zh	avg.
B (Vanilla)	50.86	0.77	71.14	74.66	71.30	33.22	69.69	49.29	43.51	62.73
B (PT)	47.77	0.54	71.54	67.16	65.20	26.49	67.17	37.71	40.73	56.76
B (ToPro)	**83.83**	**68.37**	**82.54**	**87.29**	**85.94**	**63.18**	**86.04**	**64.70**	**68.39**	**81.91**
X (Vanilla)	48.20	3.09	69.84	75.58	73.43	59.48	67.92	50.25	25.28	61.30
X (PT)	40.89	3.67	62.14	64.48	61.21	38.17	61.68	35.57	24.51	53.05
X (ToPro)	**84.73**	**19.56**	**78.35**	**89.35**	**85.74**	**61.11**	**82.18**	**66.38**	**66.09**	**80.03**
T (Vanilla)	66.67	29.23	63.28	69.28	69.94	37.75	61.28	61.24	50.87	64.19
T (ToPro)	**94.82**	**79.33**	**90.34**	**96.21**	**93.45**	**89.06**	**92.94**	**84.54**	**90.37**	**92.82**

Table 11: Detailed results of the cross-lingual evaluation on PAN-X.

lang.	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he
B (Vanilla)	95.28	86.10	53.51	85.65	86.36	81.92	**86.79**	80.78	58.75	66.10	80.33	**84.59**	56.27
B (PT)	94.96	86.06	55.59	85.81	86.03	80.87	85.04	76.74	59.99	66.91	77.75	79.67	56.12
B (ToPro)	**95.82**	**89.37**	**70.02**	**88.45**	**89.46**	**85.72**	85.93	**84.64**	**68.86**	**68.33**	**82.96**	84.43	**80.68**
X (Vanilla)	95.64	87.88	65.41	88.48	88.03	86.63	**88.31**	85.96	70.07	69.22	85.32	**86.57**	66.38
X (PT)	95.18	87.92	65.69	88.35	87.76	**86.78**	87.98	84.96	66.71	68.57	84.60	86.21	66.12
X (ToPro)	**96.05**	**89.88**	**70.06**	**89.04**	**89.61**	86.14	87.08	**86.90**	**71.95**	**70.04**	**85.80**	81.21	**80.50**
T (Vanilla)	89.67	85.02	63.56	78.38	79.86	75.44	83.99	78.35	68.49	66.47	77.50	82.10	64.19
T (ToPro)	**97.57**	**92.18**	**78.79**	**92.72**	**92.35**	**88.50**	**89.72**	**89.93**	**82.63**	**81.59**	**89.18**	**90.51**	**87.87**

lang.	hi	hu	id	it	ja	kk	ko	lt	mr	nl	pl	pt	ro
B (Vanilla)	63.54	78.92	71.67	**88.46**	47.05	70.53	50.82	79.79	70.35	89.00	81.77	**86.44**	78.00
B (PT)	64.54	78.40	71.47	86.78	46.77	70.19	51.63	76.93	67.24	88.49	81.15	86.02	77.14
B (ToPro)	**73.16**	**79.48**	**76.30**	86.45	**52.10**	**74.98**	**64.68**	**83.54**	**75.75**	**89.50**	**84.97**	85.36	**81.15**
X (Vanilla)	69.35	**82.97**	72.83	87.79	25.62	76.14	52.75	84.67	82.61	89.26	83.91	**87.16**	84.23
X (PT)	69.19	82.72	72.50	**88.88**	22.17	74.93	53.29	83.11	81.22	88.95	84.24	87.11	83.80
X (ToPro)	**72.98**	80.90	**76.93**	86.53	**54.78**	**76.61**	**64.14**	**87.16**	**80.09**	**89.54**	**85.74**	86.37	**85.69**
T (Vanilla)	69.21	76.85	72.11	82.71	50.81	71.57	51.22	76.92	72.58	83.85	77.39	82.78	74.51
T (ToPro)	**87.89**	**90.75**	**85.92**	**90.99**	**78.12**	**87.18**	**76.79**	**89.73**	**89.76**	**93.01**	**91.16**	**90.74**	**88.58**

lang.	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	avg.
B (Vanilla)	85.82	58.78	**76.63**	41.08	**82.30**	69.46	81.04	55.04	55.98	30.93	59.56	62.62	70.89
B (PT)	86.58	59.33	74.25	37.00	77.59	66.17	81.32	56.01	54.69	29.19	57.50	63.88	69.91
B (ToPro)	**90.02**	**72.21**	75.70	**56.92**	81.71	**71.29**	**86.95**	**67.08**	**58.77**	**33.38**	**65.29**	**72.17**	**76.16**
X (Vanilla)	89.16	61.94	**84.38**	44.73	86.80	**74.22**	85.22	58.88	58.48	**30.07**	**26.12**	32.08	72.42
X (PT)	88.50	61.91	82.11	40.80	**88.64**	72.74	84.85	60.68	57.34	28.79	25.07	33.81	71.86
X (ToPro)	**90.70**	**72.78**	83.79	**70.01**	82.11	73.85	**87.17**	**66.82**	**59.79**	19.38	19.53	**76.38**	**76.16**
T (Vanilla)	82.12	62.85	78.65	64.06	73.73	68.58	77.17	64.63	58.43	**54.89**	66.74	43.90	71.39
T (ToPro)	**93.48**	**83.23**	**90.65**	**79.67**	**93.11**	**85.46**	**90.67**	**85.10**	**79.03**	54.01	**72.71**	**82.43**	**86.11**

Table 12: Detailed results of the cross-lingual evaluation on UDPOS.