Polyglot Prompting: Multilingual Multitask Prompt Training

Jinlan Fu
NUS
jinlan@nus.edu.sg

See-Kiong Ng
NUS
seekiong@nus.edu.sg

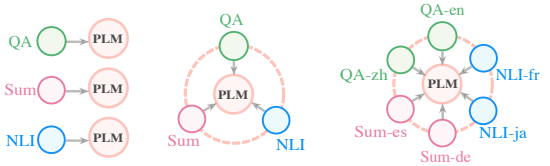
Pengfei Liu
CMU & Inspired Cognition
stefanpengfei@gmail.com

Abstract

This paper aims for a potential architectural improvement for multilingual learning and asks: *Can different tasks from different languages be modeled in a monolithic framework, i.e. without any task/language-specific module*? The benefit of achieving this could open new doors for future multilingual research, including allowing systems trained on low resources to be further assisted by other languages as well as other tasks. We approach this goal by developing a learning framework named *Polyglot Prompting* to exploit prompting methods for learning a unified semantic space for different languages and tasks with multilingual prompt engineering. We performed a comprehensive evaluation of 6 tasks, namely topic classification, sentiment classification, named entity recognition, question answering, natural language inference, and summarization, covering 24 datasets and 49 languages. The experimental results demonstrated the efficacy of multilingual multitask prompt-based learning and led to inspiring observations. We also present an interpretable multilingual evaluation methodology and show how the proposed framework, multilingual multitask prompt training, works. We release all datasets prompted in the best setting and code. [1]

1 Introduction

The emergence of multilingual pre-trained language models (Xue et al., 2021; Liu et al., 2020; Conneau et al., 2020; Conneau and Lample, 2019) enables different languages to be represented in a unified semantic space. As a result, a fine-tuned model of a data-rich language such as English can achieve decent transfer (e.g., zero-shot) performance in geographically, syntactically, or phonetically similar languages (Malaviya et al., 2017). The insufficient features learned by languages under lower-resource settings can thus be compensated through the higher-resource languages shared with them.



(a) Generic (b) MT (c) MLMT

Figure 1: Application of prompt technology in three different scenarios: Generic, Multitask (MT), Multilingual Multitask (MLMT). *QA*, *Sum*, and *NLI* represent different tasks, namely question answering, summarization, and natural language inference here. PLM represents pre-trained language model, and "zh", "en", "fr", "ja", "de","es" denote different languages.

Despite the preliminary success in the low-resource scenarios using shared knowledge across languages in multilingual language models (Wang et al., 2019b; Liu et al., 2019; K et al., 2020), the cross-lingual transfers have mostly occurred only within the boundary of the same task or similar tasks. *Can more conceptually-diverse tasks from different languages communicate together?* While researchers have made some preliminary progress towards this direction (Yang et al., 2016; Lin et al., 2018; Hasan et al., 2021; Dong et al., 2015; Mahabadi et al., 2021), the scope of the "different" tasks had remained relatively narrow, such as limiting to tasks with the same sequence labeling form (Yang et al., 2016) (e.g., named entity recognition, chunking, and part-of-speech), or different domains for the same task (Wang et al., 2020).

Unifying different tasks into one framework can be challenging if we are to avoid introducing additional task-specific parameterized modules. Recently, the success of the prompting methods (Liu et al., 2021b; Sanh et al., 2021) has provided us with new clues on unifying different tasks in the same framework without task-specific parameters by formulating all tasks as a pre-training problem

[1] https://github.com/jinlanfu/Polyglot_Prompt

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9919–9935
December 7-11, 2022

with various frameworks such as the mask language model (Devlin et al., 2019; Conneau et al., 2020) or the encoder-decoder model (Xue et al., 2021; Raffel et al., 2020; Liu et al., 2020; Chi et al., 2021).

In this paper, we leverage prompt techniques to cross the boundaries of different tasks and languages so that multiple tasks in different languages can be placed in a *monolithic* framework (as shown in Fig. 1-(c) as opposed to single task single language (Fig. 1-(a)) as well as multiple task single language learning (Fig. 1-(b))) to benefit from one another without requiring any task/language-specific modules.

We name this multilingual multitask training model as *Polyglot Prompting (PolyPrompt)*. Different tasks from different languages can then be seamlessly connected together by being reformulated as pre-training tasks. Architecturally, we choose the encoder-decoder pre-training framework so that more NLP tasks could be unified, as compared to other architectures such as the mask language model that favors classification-based tasks. Our explorations in this paper are driven by following research questions:

Q1: *Can different tasks from different languages benefit from each other by a monolithic framework?* If the answer is "yes", can the performance be further improved by introducing more high-resource datasets that are more readily available?[2] We develop *PolyPrompt*, a new multitask multilingual learning framework, and study the performance influenced by the introduction of 17 high-resource datasets. (Sec. 5.1)

Q2: *Can PolyPrompt benefit all languages in different datasets? If not, how do different characteristics of datasets and languages affect the performance of PolyPrompt?* We try to give answers by designing a multilingual interpretable evaluation methodology (Fu et al., 2020; Liu et al., 2021a) to analyze the strengths and weaknesses of the unified framework for different tasks, datasets, and languages. (Sec. 5.2)

Q3: *What makes a good prompt for multilingual multitask prompt training?* Applying the prompting method to a multilingual multitasking setting requires considering various difficulties of prompt engineering in the linguistic dimension. We study two aspects of the prompt designs for*PolyPrompt*: the language choice of prompt templates and the uniformity of prompt templates across tasks. (Sec. 5.3)

The main observations are listed in Sec. 6. Below, we summarize the main contributions. (1) To the best of our knowledge, this is the first architectural exploration for the learning of multiple conceptually-different tasks (e.g., classification, question answering and text generation) and multiple diverse languages, which relies solely on a monolithic model. (2) We introduce the concept of *multilingual prompt engineering* and provide empirical insights on what makes a good multilingual prompt. (3) We have conducted extensive experiments for *in-language training*, *cross-lingual zero-shot transfer*, and *cross-task & cross-lingual zero-shot transfer* scenarios, and designed an interpretable multilingual evaluation methodology to understand how multitask multilingual prompting works, which leads to interesting observations (Sec. 5.2).

2 Related Work

Multitask & Multilingual Learning The developments of neural networks have made it easier to share information across tasks or languages. As such, in the past few years, there has been much work on multitask learning within the same language (Liu et al., 2015, 2016; Søgaard and Goldberg, 2016; Kumar et al., 2016), or multilingual learning in the same or similar types of tasks such as sequence labeling (Yang et al., 2016; Lin et al., 2018) and machine translation (Dong et al., 2015; Mahabadi et al., 2021; Hu et al., 2021). However, the task of learning different languages and tasks simultaneously in a unified learning framework without task- or language-dependent parameters has remained unexplored.

Prompting Methods Prompting is a technique that aims to make better use of pre-trained knowledge by reformulating tasks at hand accordingly (Liu et al., 2021b) and thus allowing us to do more with one model by unifying signals cross tasks (Sanh et al., 2021), languages (Zhao and Schütze, 2021), even modality (Zhao et al., 2021). In this paper, we expand what a system can do by proposing multilingual multitask learning with prompting methods for connecting geographically diverse languages and linguistically different tasks, thereby allowing them to leverage one another effectively.

[2] It is relatively easy for us to obtain the training set of relevant tasks in real scenarios. Therefore, the purpose of this research question is to explore whether the data that is relatively easy to obtain from relevant tasks can bring benefits after being introduced into the multitask learning framework.

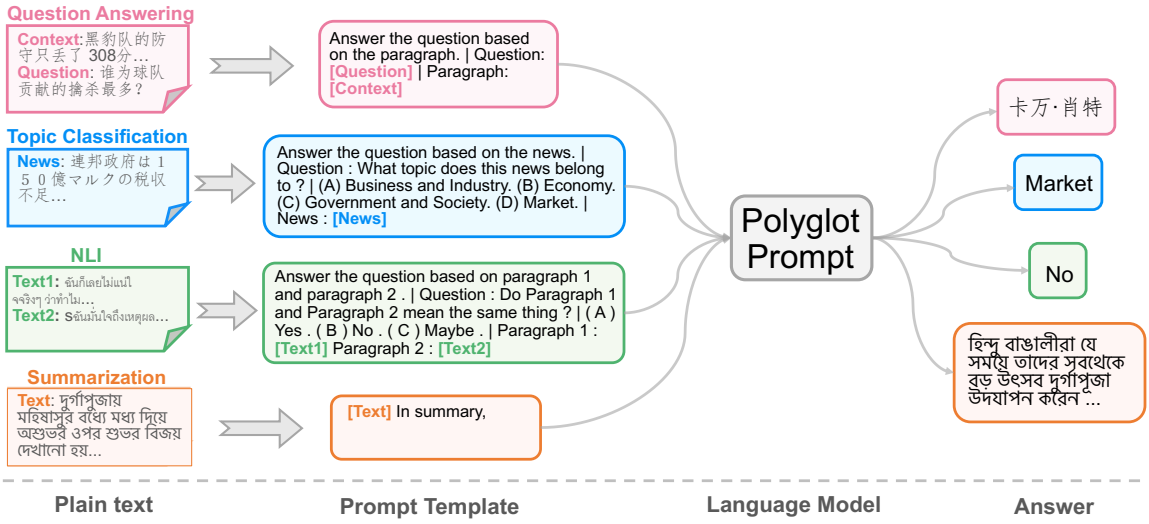


Figure 2: The proposed *PolyPrompt* framework for multilingual multitask prompt training.

3 Multitask Multilingual Prompt Training

We unify different tasks from different languages by reformulating each NLP task as a sequence-to-sequence problem (Sutskever et al., 2014; Bahdanau et al., 2015) so that they can be connected by a multilingual pre-trained language model (e.g., mT5 (Xue et al., 2021)) that also adopts a sequence-to-sequence training objective. Fig. 2 shows the overview of our proposed framework. Each sample from different tasks will be re-formatted as a (*prompt*, *answer*) pair using pre-defined *templates* and then be fed into a multilingual pre-trained language model.

Formally, given a task set $\mathcal{T} = \{T_1, T_2, \cdots, T_n\}$ with n tasks and corresponding prompt templates $\mathcal{K} = \{K_1, K_2, \cdots, K_n\}$ (without loss of generality, we shall assume that each task has one prompt for easy understanding). First, we transform the samples in each task into a form that is understandable by the language model based on the predefined templates $\mathcal{K}$. Assume that $(\mathrm{x}_{i,j}, \mathrm{y}_{i,j}) \in \mathcal{Z}$ is the input and output pair for the j-th sample of the i-th task, where $\mathcal{Z}$ contains all input and output sample pairs. The input-output pair $(\mathrm{x}_{i,j}, \mathrm{y}_{i,j})$ for the j-th sample of the i-th task can be converted to $(\hat{\mathrm{x}}_{i,j}, \hat{\mathrm{y}}_{i,j})$ through the predefined template K_i, which can be formulated as:

$$(\hat{\mathrm{x}}_{i,j}, \hat{\mathrm{y}}_{i,j}) = K_i(\mathrm{x}_{i,j}, \mathrm{y}_{i,j}) \quad (1)$$

We choose a sequence-to-sequence language model to achieve multilingual multitask prompt training, where samples from n tasks will be the input of the chosen language model. The loss function is to maximize the log-likelihood of the output text and can be defined as:

$$\mathrm{L} = \sum_{(\hat{\mathrm{x}},\hat{\mathrm{y}})\in(\mathrm{X},\mathrm{Y})} log(\prod_{m=1}^{|\hat{\mathrm{y}}|} P(\hat{\mathrm{y}}_m|\hat{\mathrm{y}}_{<m}, \hat{\mathrm{x}}; \theta)), \quad (2)$$

where $(\hat{\mathrm{x}}, \hat{\mathrm{y}}) \in \mathcal{Z}$ represents a sequence-to-sequence text pair for any task. $|\hat{\mathrm{y}}|$ is the number of tokens in the decoded text, and $\hat{y}_{<m}$ is the target tokens before the time step m.

4 Experiment Setup

4.1 Tasks & Datasets

The datasets, tasks, and evaluation metrics studied in this work are shown in Tab. 1. We call those datasets that provide training and test sets for multilingual multitask prompt training as **target datasets**. To explore the influence of introducing more high-resource English and multilingual datasets to *PolyPrompt*, we present the **expanding datasets**, which only provides training datasets for multilingual multitask prompt training (we do not evaluate *PolyPrompt* and its variants on expanding datasets). Overall, we study 7 multilingual target datasets covering 4 NLP tasks (*question answering, sentiment classification, topic classification, and sentence pair classification*), and 15 monolingual (English) and 2 multilingual expanding datasets covering 6 NLP tasks (*text summarization, named entity recognition, and 4 tasks covered by target datasets*). Further details of the target and expanding datasets can be found in App. B. The languages considered in this work can be seen in App. A

Task	Dataset	Domain	Lang.	Metric
Target Datasets				
QA	XQuAD	Wikipedia	11	F1 / EM
	TyDiQA	Wikipedia	9	F1 / EM
	MLQA	Wikipedia	7	F1 / EM
SC	MARC	Amazon	6	Acc.
TC	MLDOC	Reuters	8	Acc.
SPC	PAWS-X	Wikipedia/Quora	7	Acc.
	XNLI	Misc.	15	Acc.
Expanding Datasets				
Summ.	XL-Sum	BBC	45	-
NER	Wikiann	Wikipedia	40	-
QA	SQuAD 2.0	Wikipedia	1	-
	Quoref	Wikipedia	1	-
	NewsQA	CNN	1	-
	ROPES	Textbooks/Wikipedia	1	-
	MCTest	Misc.	1	-
	Social IQa	Misc.	1	-
TC	DBpedia-2014	DBpedia	1	-
	AG_News	Reuters	1	-
	YATC	Yahoo!	1	-
SC	IMDB	IMDb	1	-
	SST2	Rotten Tomatoes	1	-
	ARP	Amazon	1	-
SPC	Quora	Quora	1	-
	RTE	News/Wikipedia	1	-
	SNLI	Misc.	1	-

Table 1: The tasks and datasets studied in this work. *Lang.* and *Acc.* denote "Language" and "Accuracy". `Summ.`, `NER`, `QA`, `TC`, `SC`, and `SPC` are abbreviations for summarization, named entity recognition, question answering, topic classification, sentiment classification, and sentence pair classification. "-" indicates that the task is not used for evaluation. "Misc." indicates that the dataset was artificially constructed or of unclear origin.

4.2 Experimental Settings

Model We list 5 models explored in this work.
(1) **Vanilla mT5**: In the *cross-lingual zero-shot transfer* setting, mT5 is trained on the training set in English of the specific task (e.g. XNLI), while in the *in-lingual training* setting, mT5 is trained on the training samples in all languages for the particular task (e.g. XNLI).
(2) **Polyglot Prompt (PolyPrompt)** is a standard multilingual multitask prompt training model, which is trained on 7 target datasets covering 4 NLP tasks (e.g., QA).
(3) **PolyPrompt+Expand** is the *PolyPrompt* model trained on the 7 target datasets and 15 high-resource (English) expanding datasets.
(4) **PolyPrompt+Expand+PANX** is the *PolyPrompt* trained on the 7 target datasets, 15 high-resource datasets, and a multilingual NER dataset (PANX).
(5) **PolyPrompt+Expand+XLSum** is the *PolyPrompt* trained on the 7 target datasets, 15 high-resource datasets, and a multilingual summarization dataset (XL-Sum).

Parameters The *PolyPrompt* model is built based on the mT5 (Xue et al., 2021), and our experiments are designed based on the mT5-base (Wolf et al., 2020) with 580 million parameters. We used token limits of size 512 and 64 for input and output sequences, respectively. All models have a learning rate of $1e-4$, with the batch size set to 18, and were trained for 20 epochs. During training, checkpoints were saved every $1,000$ steps. The model with the best performance on the validation set was selected.

Training Data Construction Some datasets have a large number of training samples, for example, XNLI has 4.5 million training samples. To reduce the expensive computational cost of our experiments, we randomly sampled $3,000$ samples from the training set for each language of the target datasets, and $5,000$ samples from each expanding dataset. These selected samples will serve as the training set for multilingual multitask prompt training with different experiment scenarios.

Experimental Scenario We consider three experimental scenarios: (1) **In-language training**, fine-tuned on golden data in all target languages. Like Hu et al. (2020), we use the translations from English released by Hu et al. (2020) as the golden training samples for the target language for the XQuAD, MLQA, XNLI, and PAWS-X datasets, which have only English training sets. (2) **Cross-lingual zero-shot transfer** (Hu et al., 2020), where the model is fine-tuned only on the training set in English. (3) **Cross-task & cross-lingual zero-shot transfer**, where a model is evaluated on tasks and languages that did not appear in its training dataset.

5 Results & Analysis

5.1 Exp-I: Effect of Multitask Prompt Training

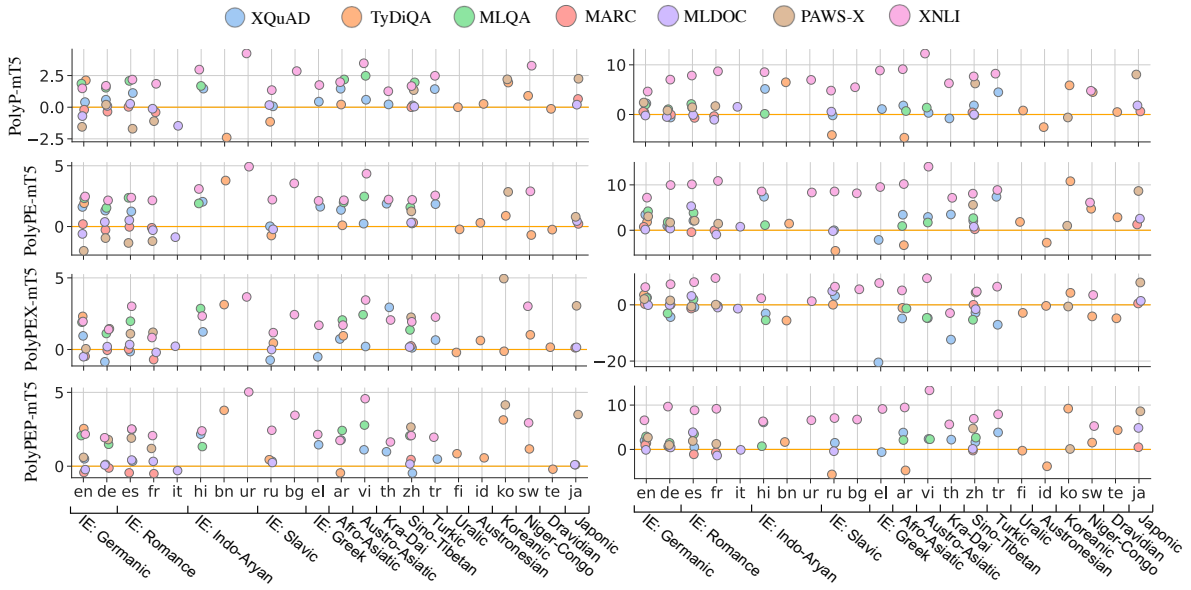
The experiment in this section is designed to answer the research question **Q1** in Sec.1, namely to investigate whether multilingual multitask prompt training (*PolyPrompt*) can achieve improvement, and whether the performance can be further improved by introducing more high-resource datasets.

5.1.1 Approach

Significance tests: To examine whether the *PolyPrompt* and its variants are significantly better

	Question Answering						Sentiment	Topic	Sentence Pair			
	XQuAD		TyDiQA		MLQA		MARC	MLDOC	PAWS-X	XNLI	Avg.	Sig.
Metrics	F1	EM	F1	EM	F1	EM	Acc.	Acc.	Acc.	Acc.		
In-language training												
Vanilla mT5	72.93	57.22	81.44	70.78	62.93	44.61	91.71	93.99	84.85	69.52	73.00	-
PolyPrompt	73.65	58.17	81.63	70.32	64.90	46.44	91.66	93.80	85.09	71.82	73.75	1.91E-03
+Expand	**74.15**	**58.93**	82.00	70.69	64.95	46.57	**91.77**	93.95	84.76	**72.28**	74.00	1.54E-03
+Expand+XLSum	73.35	58.01	82.37	**71.47**	64.88	46.36	91.57	94.04	86.88	71.71	74.06	1.03E-04
+Expand+PANX	73.73	58.43	**82.75**	71.70	**65.02**	**46.60**	91.55	**94.09**	**87.10**	72.12	**74.31**	1.03E-04
Cross-lingual zero-shot transfer												
Vanilla mT5	62.49	44.51	64.67	47.46	57.16	38.92	89.75	85.74	78.24	55.54	62.45	-
PolyPrompt	64.01	46.33	65.47	49.57	58.19	39.92	89.85	86.01	81.10	62.95	64.34	3.96E-05
+Expand	**65.31**	**48.07**	**66.11**	**50.39**	**59.48**	**41.71**	**90.14**	**86.84**	**81.60**	**64.57**	**65.42**	3.96E-05
+Expand+XLSum	57.50	40.56	63.45	46.94	54.97	37.51	89.60	86.43	80.35	60.93	61.82	0.18
+Expand+PANX	64.67	47.51	65.08	48.38	59.40	41.41	89.75	86.66	81.13	63.44	64.74	8.63E-05

Table 2: Overall results of the models explored in this work on 7 multilingual datasets from 4 NLP tasks. Values in **bold** represent the best performance in a particular setting (e.g. in-language training). *"Avg."* denotes the average performance of the 7 datasets, and "-" means not applicable. *"Sig."* is the *"significance test"*, where gray values indicate that the evaluated model failed the significance test ($p > 0.05$).



(a) In-language training

(b) Cross-lingual zero-shot transfer

Figure 3: The relative performance improvement of *PolyPrompt* and its variants over the *vanilla mT5 (mT5)* at the language-level. "IE" denotes the "Indo-European". *PolyP*, *PolyPE*, *PolyPEX*, and *PolyPEP* are abbreviations for *PolyPrompt*, *PolyPrompt+Expand*, *PolyPrompt+Expand+XLSum*, and *PolyPrompt+Expand+PANX*.

than the *vanilla mT5*, we perform the significance test with Wilcoxon's Signed-rank Test (Wilcoxon et al., 1970) at $p = 0.05$. The null hypothesis is that the performance of *PolyPrompt* and its variants is indistinguishable from that of *vanilla mT5*.

5.1.2 Results

We detail main observations in Tab. 2 and Fig. 3:

(1) ***PolyPrompt* can achieve improvement, especially with the introduction of high-resource datasets.** Compared to *Vanilla mT5*, the average performance of *PolyPrompt* and its variants (e.g. *PolyPrompt+Expand*) was greatly improved on the 7 datasets of 4 tasks with both the *in-language training* and *cross-lingual zero-shot transfer* settings, other than the *PolyPrompt+Expand+XLSum* with the *cross-lingual zero-shot transfer* setting ($p = 0.18 > 0.05$). Furthermore, the best systems for the *in-language training* and *cross-lingual zero-shot transfer* scenarios are *PolyPrompt+Expand+PANX* (5 out of 7) and *PolyPrompt+Expand* (7 out of 7), respectively, illustrating the effectiveness of intro-

ducing high-resource expanding datasets.

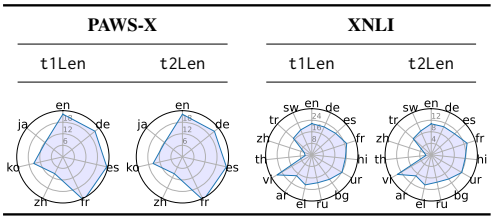


Figure 4: Dataset bias of PAWS-X and XNLI characterized by ϕ_p defined in Sec. 5.2.1.

(2) **It is more beneficial for *PolyPrompt* to introduce high-resource English datasets in the cross-language zero-shot transfer.** In Fig. 3-(b), we observe that *PolyPrompt* achieved performance gains in many languages belonging to different datasets in the *cross-language zero-shot transfer* scenarios. When the external English datasets (*PolyPrompt+Expand*) are introduced, more languages gained performance improvements. However, when multilingual datasets (e.g. PANX and XLSum) were introduced (*PolyPrompt+Expand+PANX* and *PolyPrompt+Expand+XLSum*), the overall performance dropped (observed from Tab. 2) and there were more languages with negative relative performance gains (compared to *PolyPrompt+Expand*).

5.2 Exp-II: Multilingual Interpretable Evaluation

This section aims at the research question **Q2** (*How do different characteristics of datasets and languages affect the performance of PolyPrompt?*) by introducing a multilingual interpretable evaluation.

5.2.1 Approach

Interpretable evaluation (Liu et al., 2021a; Fu et al., 2020; Ruder et al., 2021) aims to breakdown the holistic performance (e.g., F1) to a more fine-grained level based on predefined features (e.g., text length) to interpret the model's behavior better.

Below, we list some features of each task explored in this work. XQuAD, TyDiQA, MLQA: the length of context (cLen), question (qLen), and answer (aLen). The BLUE score of the answer and context (BLUE_AC). PAWS-X, XNLI: the length of sentence$_1$ (t1Len) and sentence$_2$ (t2Len). The BLUE score of the sentence pair (BLUE_t1t2). Further detailed interpretable evaluation definition can be found in App. C. We also measure the **dataset-level** feature ϕ_p, the average feature value over a dataset. For example, $\phi_{\text{aLen(MLQA)}}$ denotes the average answer length of MLQA dataset. Further details can be found in App. E

5.2.2 Results

Here are the main observations in Tab. 3 and Fig. 4:

(1) **Dataset Perspective:** the strengths of *PolyPrompt* in the co-occurring languages on different datasets are inconsistent due to dataset bias. For example, en, de, and fr co-occur on PAWS-X and XNLI. *PolyPrompt* was better in the short sentence$_2$ (t2Len:XS) in PAWS-X, while excelling in the long sentence$_2$ (t2Len:XL) of XNLI. This inconsistency results from the dataset bias shown in Fig. 4: the ϕ_{t2Len}(XNLI-[en,de,fr]) < 12 while ϕ_{t2Len}(PAWS-X-[en,de,fr]) > 20. Therefore, ϕ_{t2Len}(PAWS-X-[en,de,fr]) on bucket t2Len:XS was close to ϕ_{t2Len}(XNLI-[en,de,fr]) on bucket t2Len:XL.

(2) **Model Perspective:** *PolyPrompt* achieves overall performance improvements on the 7 target datasets, but it cannot perform well on all samples (e.g., worse performance on long sentences). *PolyPrompt* is better at short context samples for MLQA (cLen:XS/S), long context samples for XQuAD (cLen:XL/L), long sentence$_2$ for XNLI (t2Len:XL/L), which is valid for most languages. **Disadvantage analysis:** *PolyPrompt* is worse at handling long question samples (qLen:XL/L) for XQuAD, TyDiQA, and MLQA, long sentence$_1$ samples for XNLI-es, and long sentence$_1$ and sentence$_2$ samples in zh,ko,ja of PAWS-X.

(3) **Language Perspective:** it is difficult for *PolyPrompt* to bring gains for languages that appear only once in the 7 target datasets unless high-resource datasets are introduced. For example, *PolyPrompt* showed a slight performance improvement over *vanilla mT5* in languages bn, fi, id, and te that only appeared in the TyDiQA dataset. When introducing high-resource English datasets, the performance of bn is significantly improved especially for long context and short answers samples, while the performance improvement of fi, id, and te is still limited until a high-resource multilingual training dataset PANX is introduced. The reason may be that most of the languages in the 7 tasks belong to the IE language family (findings from Fig. 3), and so does the bn language. Therefore, compared to fi, id, and te, it is easier for bn to get knowledge from neighbor languages in multitask training.

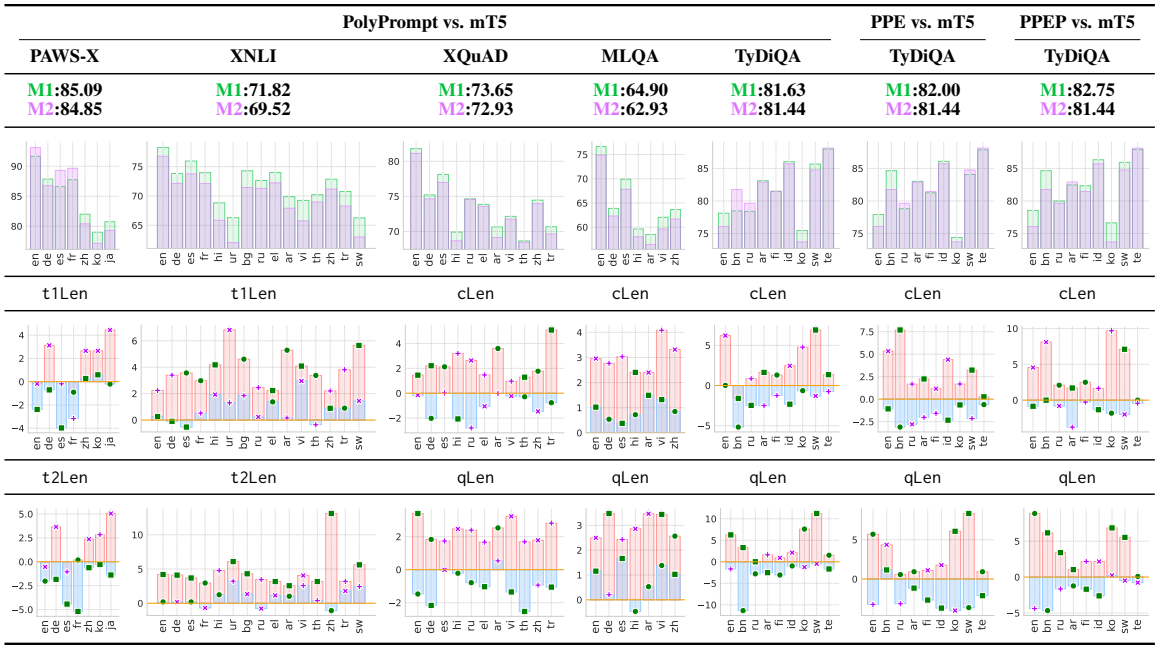
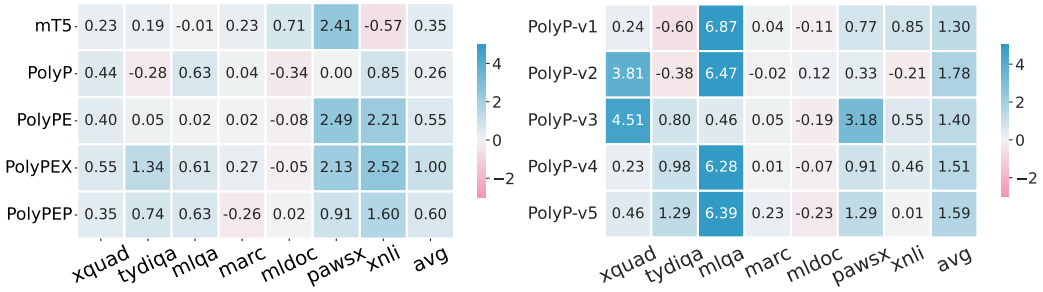


Table 3: The pairwise model diagnosis of *PolyPrompt* (and its variants) and *vanilla mT5 (mT5)*. The model name on the left side of "vs." denotes M1, and the right side represents M2. The bar charts in the first row are the overall performance of M1 (green bars) and M2 (purple bars) across different languages. The bar charts after the second row represent the relative performance improvement of *PolyPrompt* and its variants over *vanilla mT5*, where the heights of the red and blue bars represent the maximum positive and maximum negative gains. ×, +, ●, and ■ denotes the "extra-small (XS)","small (S)","large (L)", and "extra-large (XL)", respectively. *PPE* and *PPEP* denote *PolyPrompt+Expand* and *PolyPrompt+Expand+PANX*, respectively.



(a) Cross-lingual (CL) − In-lingual (IL)

(b) *PolyPrompt* − *PolyPrompt-v(x)*

Figure 5: The exploration of the language and uniformity of prompt design. (a) is the performance gap between cross-lingual (CL) and in-lingual (IL) prompt templates, where *PolyP*, *PolyPE*, *PolyPEX*, and *PolyPEP* are abbreviations for *PolyPrompt*, *PolyPrompt+Expand*, *PolyPrompt+Expand+XLSum*, and *PolyPrompt+Expand+PANX*. (b) is the relative performance improvement of *PolyPrompt* with unified prompt templates versus diversified prompt templates (e.g. *PolyP-v1*). *PolyP-v(x)* ($x \in [1, 5]$) represent x-th version of diversified prompt templates. The bluer color indicates that the model with the cross-lingual (unified) prompts outperforms the in-lingual (diversified) prompts more, while the redder color has the opposite meaning. The last column is the average relative improvement.

5.3 Exp-III: Effect of Prompt

In this section, we try to find out what prompts or prompt combinations are suitable for multilingual and multitask scenarios (**Q3**).

5.3.1 Prompt Design

Although prompting methods have proven effective in many NLP scenarios, its effectiveness comes at the cost of prompt engineering (Liu et al., 2021b), as there are usually a multitude of factors that influence the prompt design process, and the situation is clearly more complicated in the multilingual situation. Existing works have studied manual prompt (Schick and Schütze, 2021), soft (trainable) prompt (Lester et al., 2021), and mix prompt (mixing the manual and soft prompt) (Gu et al., 2021; Zhao and Schütze, 2021) design approaches. In this work, we take particular care of *language* and

Prompt Design		Dataset	Prompt Template
Language Choice	In-lingual (zh)	TyDiQA	根据段落的内容回答问题。| 问题：`[Q-zh]` | 段落：`[C-zh]`
	Cross-lingual (zh)	TyDiQA	Answer the question based on the content of the paragraph. | `[Q-zh]` | Paragraph: `[C-zh]`
Uniformity of Templates	Unified	XQuAD	Answer the question based on the paragraph. | Question: `[Q-xx]` | Paragraph: `[C-xx]`
		MLQA	Answer the question based on the paragraph. | Question: `[Q-xx]` | Paragraph: `[C-xx]`
		XNLI	Answer the question based on paragraph 1 and paragraph 2 . | Question : Do Paragraph 1 and Paragraph 2 mean the same thing ? | (A) Yes . (B) No . (C) Maybe . | Paragraph 1 : `[T1-xx]` Paragraph 2 : `[T2-xx]`
		MARC	Answer the question based on the review. | Question: Can we conclude that the buyer is satisfied with the product based on his review? | (A) Yes. (B) No. | Review: `[T1-xx]`
	Diversified	XQuAD	I have always wondered: `[Q-xx]` | I searched Wikipedia and this is what I found. What's the answer? | `[C-xx]`
		MLQA	Context: `[C-xx]` | I am trying to figure out the answer to the question from the above context. Can you tell me the answer? | Question: `[Q-xx]` Answer:
		XNLI	Given that `[T1-xx]` Therefore, it must be true that `[T2-xx]`? Yes, no, or maybe?
		MARC	I am reading a review that says `[T1-xx]`. Do you think the review is positive or negative?

Table 4: Examples of prompt design in this work. "Q", "C", "T1", and "T2" denotes the placeholders for question, context, sentence$_1$, and sentence$_2$ field, respectively. The format "[text field - language]" is used to represent text in a specific language, such as "[Q-zh]" denotes a question text in Chinese (zh). "xx" denotes any language, and "|" represents the separator of the input field (e.g., question and context). The tokens in pink are the co-occurring words in most templates.

uniformity of prompt templates designed for multilingual multitask setting. The examples of the considered prompt design can be seen in Tab. 4.

Language Choice: we consider both the *in-lingual* and *cross-lingual* prompts. *In-lingual prompts* are those in which the language of the prompt is the same as the target language (Zhao and Schütze, 2021). *Cross-lingual prompts* denote those in which the language of the prompt template is different from the target language. In this work, we keep the language of the prompt template in English (en) (Lin et al., 2021) regardless of the target language (e.g., zh).

Uniformity of Templates: Previous studies (Caruana, 1997; Evgeniou and Pontil, 2004; Argyriou et al., 2008) have shown that similar tasks would benefit from multitask training. In this work, we study *unified prompts* versus *diversified prompts*. *Unified prompts* indicates that prompt templates of different datasets have similar structures and cooccurrences. *Diversified prompts* means that the prompt templates for each task did not consider the same structure and multiple co-occurrence words. In practice, for each dataset, we designed 5 different prompts and then randomly selected one prompt for each task to build a set of diverse prompts for multitask prompt training. In total, we created 5 groups of diversified templates. The list of templates can be found in App. F.

5.3.2 Results

(1) **Cross-lingual prompt can help better retrieve knowledge encoded in language model.** We can observe from Fig. 5-(a) that the average overall performance of the 5 models equipped with CL prompts outperformed IL prompts, which holds for all the seven datasets. We think they might be because mT5 was pre-trained on a larger body of English corpus, it can understand the English template well. This makes it easier for downstream NLP tasks to retrieve knowledge from mT5.

(2) **The unified template outperforms the diversified template** In Fig. 5-(b), we observed that the *PolyPrompt* with uniform templates outperforms any diverse templates (e.g. *PolyP-v1*), especially on the QA task. The reason may be that unified prompts helped eliminate the boundaries between tasks, thereby reducing the distance between tasks and making the interaction between tasks easier.

5.4 Exp-IV: Cross-task Cross-lingual zero-shot transfer

To investigate whether *PolyPrompt* is better at retrieving relevant knowledge from pre-trained language models for tasks and languages unseen in training stage, we investigate *vanilla mT5*, *PolyPrompt*, and *PolyPrompt+Expand* fine-tuned on the English datasets and evaluate these three models on the PANX dataset, a named entity recognition task with 40 languages. We then subtract the performance of *vanilla mT5* from *PolyPrompt* and *PolyPrompt+Expand* in the same language, and the results are shown in Fig. 6.

Results: (1) **Almost all languages benefit from both *PolyPrompt* and its variants.** *PolyPrompt* brings gains for 34 of the 40 languages, and more languages will benefit when *PolyPrompt* is enhanced with high-resource English training datasets. Interestingly, *PolyPrompt+Expand* per-

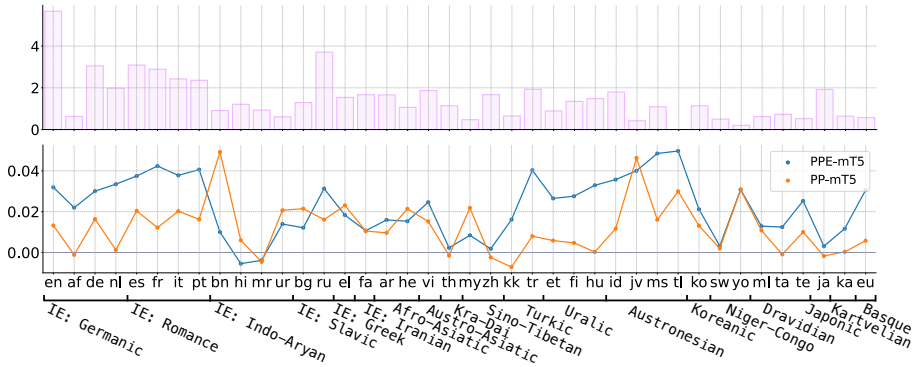


Figure 6: The performance improvement of *PolyPrompt* (*PP*, orange line) and *PolyPrompt+Expand* (*PPE*, blue line in line chart) relative to *vanilla mT5* in the cross-task & cross-lingual zero-shot setting. The bar chart is the proportion (%) of different languages in the training corpus participating in mT5 pre-training, where `tl` is a language that never participated in mT5 pre-training.

formed much better than *PolyPrompt* in languages belonging to *IE: Germanic* and *IE: Romance* language families, which made up a large proportion of samples in the pre-training corpus of mT5.

(2) ***PolyPrompt* significantly improves performance on languages that have never appeared in the pre-training corpus of *mT5*.** Both *PolyPrompt* and *PolyPrompt+Expand* improve a lot over mT5 on `tl`, a language that never appeared in mT5's pre-training corpus. Furthermore, *PolyPrompt+Expand* achieves the best performance gain on `tl`. The reasons can be attributed to (1) we unify different tasks into a monolithic framework (including NER), which effectively shortens the distance between different tasks; (2) `English (en)` and `tl` share the same semantic space, NER knowledge in `English (en)` can be effectively transferred to `tl`.

6 Conclusions

We can provide the following preliminary empirical answers to our research questions.

(1) ***Can different tasks from different languages benefit from each other by a monolithic framework?*** Yes. What's more, introducing more high-resource datasets can further improve the tasks' performance involved in multitask prompt training.

(2) ***How do different characteristics of datasets and languages affect the performance of PolyPrompt?*** *PolyPrompt* cannot benefit all languages in all datasets. For example, (a) languages that appear only once in target datasets have benefits when *PolyPrompt* is enhanced by high-resource datasets; (b) *PolyPrompt* is better in short context samples for MLQA, long context samples for XQuAD, while poor in long question samples for XQuAD, TyDiQA, and MLQA.

(3) ***What makes a good prompt for multilingual multitask prompt training?*** The best performance is achieved when the model is equipped with cross-lingual prompts (i.e., using English as prompt templates regardless of what the language of training samples is) and prompts with unified templates across tasks.

7 Limitations

Although in this paper, we try to cover as many languages and tasks as possible, some tasks (e.g., semantic parsing, machine translation) and languages are still not considered. In addition, due to limited computational resources, we adopt a relatively small pre-trained language model, and the results on the larger pre-trained language models are also worth expecting. In addition, there are a variety of factors affecting the design of prompts in a multilingual setting. This paper only considers two (language choice and uniformity of prompt templates), so more comprehensive studies in this direction could be conducted.

Acknowledgements

We thank Graham Neubig and Junjie Hu for their useful discussion and suggestions on this work. This work was supported by the National Research Foundation of Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the National Research Foundation of Singapore.

References

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2008. Convex multi-task feature learning. *Machine learning*, 73(3):243–272.

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, ACL 2020, Online, July 5-10, 2020*, pages 4623–4637. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021. mt6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1671–1683. Association for Computational Linguistics.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5924–5931. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 109–117. ACM.

Jinlan Fu, Pengfei Liu, and Graham Neubig. 2020. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6058–6069. Association for Computational Linguistics.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. PPT: pre-trained prompt tuning for few-shot learning. *CoRR*, abs/2109.04332.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *CoRR*, abs/2106.13822.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2021. Deep: Denoising entity pre-training for neural machine translation. *arXiv preprint arXiv:2111.07393*.

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 *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A Smith. 2020. The multilingual amazon reviews corpus. *arXiv preprint arXiv:2010.02573*.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387. PMLR.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7315–7330. Association for Computational Linguistics.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 58–62. 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 T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809, Melbourne, Australia. Association for Computational Linguistics.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021a. EXPLAINABOARD: an explainable leaderboard for NLP. *CoRR*, abs/2104.06387.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Qianchu Liu, Diana McCarthy, Ivan Vulic, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 33–43. Association for Computational Linguistics.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, Denver, Colorado. 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. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

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 *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150. The Association for Computer Linguistics.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2529–2535. 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1946–1958. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 151–164. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 193–203. ACL.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya 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, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

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, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019b. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5720–5726. Association for Computational Linguistics.

Frank Wilcoxon, SK Katti, and Roberta A Wilcox. 1970. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3685–3690. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*.

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.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. 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.

Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. 2021. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. *CoRR*, abs/2111.00865.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. *CoRR*, abs/2109.03630.

A Languages

In this work, we studied 49 languages that appear in 24 datasets covering 6 NLP tasks. For brevity, the languages are shown in ISO 639-1 codes[3] as follows: af, am, ar, az, bg, bn, cy, de, el, en, es, et, eu, fa, fi, fr, gu, ha, he, hi, hu, id, ig, it, ja, jv, ka, kk, ko, ml, mr, ms, my, nl, np, pa, pt, ru, si, sw, ta, te, th, tl, tr, ur, vi, yo, zh. Among them, **zh, ja, th, te, km** are languages that do not use space separation for words.

B Datasets

B.1 Target Datasets

XQuAD (Artetxe et al., 2020) is a cross-lingual question answering dataset, including 11 languages. Its English dataset is a subset of the development set of SQuAD v1.1. The other 10 languages are professional translations of the English dataset. Therefore, the dataset in 11 languages is completely parallel.

MLQA (Lewis et al., 2020) is another multi-language extractive QA dataset, including 7 languages. Each QA instance is paralleled between 4 languages on average. Since MLQA and XQuAD lack training sets, following (Hu et al., 2020), we use the training data of SQuAD v1.1 as their training set.

TyDiQA-GoldP (TyDiQA) (Clark et al., 2020) is the gold passage version of the TyDiQA benchmark, including 9 languages for training, development, and testing. TyDiQA-GoldP is a simplified

[3] https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

version of the primary task, discarding Thai and Japanese languages and samples without answers. Like XQuAD and MLQA, TyDiQA is evaluated with SQuAD 1.1 (Rajpurkar et al., 2016) metrics.

XNLI (Conneau et al., 2018) is a cross-lingual natural language inference dataset containing annotated development and test sets in 15 languages, and an annotated English training set. The English training set has been translated into the other 14 languages through a machine translator.[4]

PAWS-X (Yang et al., 2019) is a cross-lingual paraphrase adversary from a word scrambling dataset with 7 languages. The goal of this task is to predict whether two sentences are paraphrases. The training set of PAWS-X is the PAWS's training data, and the subset of PAWS's development and test sets are translated into 6 other non-English datasets for evaluation.

MARC (multilingual Amazon Reviews Corpus) (Keung et al., 2020) is a multilingual text classification dataset with 6 different languages. Here, we use the binarized classification task that is defined by Keung et al. (2020).

MLDOC (Schwenk and Li, 2018) is a multilingual document classification dataset with six topics.

B.2 Expanding Datasets

Expanding tasks simply provide training sets for the multitask prompt training. In summary, we studied 15 English and 2 multilingual datasets.

Extractive Question Answering is the task of finding an answer to a given question from the context. We adopt SQuAD 2.0 (Rajpurkar et al., 2016), Quoref (Dasigi et al., 2019), NewsQA (Trischler et al., 2017), and ROPES (Lin et al., 2019).

Multiple-choice Question Answering aims to select an answer from candidate options based on the context and question. In this work, we study MCTest (Richardson et al., 2013) and Social IQa (Sap et al., 2019).

Natural Language Inference aims to determine the inference relation (e.g. entailment) between two texts. The datasets used in this work are Quora,[5] RTE (Wang et al., 2019a), and SNLI (Bowman et al., 2015).

[4] https://console.cloud.google.com/storage/browser/xtreme_translations

[5] https://huggingface.co/datasets/quora

Topic Classification is a task to predict a suitable topic (e.g., health) for a given text. We use the following topic classification datasets: DBpedia-2014 (Zhang et al., 2015), AG_News (Zhang et al., 2015), and YATC (Yahoo! Answers Topic Classification Dataset) (Zhang et al., 2015).

Sentiment Classification aims to identify the sentiment polarity of a given text. We studied datasets IMDB (Maas et al., 2011), Amazon Review Polarity (ARP) (Zhang et al., 2015), and SST2 (Socher et al., 2013).

XL-Sum (Hasan et al., 2021) is a multilingual summarization dataset covering 45 low- to high-resource languages. We randomly select 32 out of 45 languages for multitask prompt training. The ISO-639-1 codes for the chosen languages are *en, ar, vi, ko, es, zh, ru, fr, tr, hi, id, fa, pt, mr, th, az, bn, np, sr, sw, ta, te, ur, cy, am, my, gu, ha, ig, pa, si, yo*.

PANX (Pan et al., 2017) is a multilingual named entity recognition dataset in 40 languages constructed based on Wikipedia corpus. Following Hu et al. (2020), we use the version with balanced train, development, and test splits from Rahimi et al. (2019).

C Interpretable Multilingual Evaluation

For interpretable evaluation, the first step is attribute definition, and the second is sample breakdown. Assume that $\phi_{\text{Len}}(x)$ is a function to calculate the number of tokens in the given text x, and $\phi_{\text{BLUE}}(x_1, x_2)$ is to compute the BLUE score of two given texts x_1 and x_2. The following are the features tailored for the 7 multilingual datasets in this paper:

- **XQuAD, TyDiQA, MLQA**: `cLen`=$\phi_{\text{Len}}(X_c)$, `qLen`=$\phi_{\text{Len}}(X_q)$, `aLen`=$\phi_{\text{Len}}(X_a)$, and `BLUE_AC`= $\phi_{\text{BLUE}}(X_a, X_c)$, where X_c, X_q, and X_a denote the context, question, and answer sequence, respectively.
- **PAWS-X, XNLI**: `t1Len` = $\phi_{\text{Len}}(X_{t1})$, `t2Len` = $\phi_{\text{Len}}(X_{t2})$, `t1Len/t2Len` = $\phi_{\text{Len}}(X_{t1})/\phi_{\text{Len}}(X_{t2})$, and `BLUE_t1t2` = $\phi_{\text{BLUE}}(X_{t1}, X_{t2})$, where X_{t1} and X_{t2} denote the premise and hypothesis (sentence-1 and sentence-2 for PAWS-X) sequence.
- **MARC, MLDOC**: `t1Len`=$\phi_{\text{Len}}(X_{t1})$, `t1basic` = $\phi_{\text{basic}}(X_{t1})$, and `t1eNum` = $\phi_{\text{eNum}}(X_{t1})$, where

X_{t1} denotes a sequence of review (news for MLDOC). $\phi_{\text{basic}}(x)$ and $\phi_{\text{eNum}}(x)$ are functions to calculate the proportion of words belonging to the 1000 essential English words [6] and entities, respectively.

We then follow Fu et al. (2020) and breakdown the samples into four buckets, XS (extra-small), S (small), L (large), and XL (extra-large), according to their feature values, and calculate the performance for each bucket.

D Main Observations

Due to the space limitation, we summarize some main observations here.

(1) Whether a language that appears in only one task could gain improvement depends on the difficulty of the task. In Fig. 3-(a), we can observe that some languages in XNLI-[ur,bg], TyDiQA-[bn,fi,id,te] and MLDOC-[it] were not present in other tasks (e.g. it is only present in MLDOC). These languages that appeared only once in multitask training have significant performance gains on the XNLI dataset, while performance dropsped significantly on the TyDiQA and MLDOC datasets. The reason could be that XNLI is a task that relies more on fundamental knowledge (Yin et al., 2019), which is relatively easier to acquire from other tasks. On the contrary, tasks such as TyDiQA need to understand more, for example, the semantics of sentences and the position of the answer.

(2) ***PolyPrompt* improves the performance of non-Indo-European languages a lot in the *in-language training*.** From Fig. 3-(a), we can observe that languages belonging to *non-Indo-European* language families (e.g. *Sino-Tibet* and Niger-Congo) always have performance gains no matter which datasets were employed. However, in languages belonging to the *Indo-European*-related language families, the relative performance gains varied widely across datasets. For example, languages belonging to the XNLI and XQuAD datasets consistently achieved positive relative performance, while languages belonging to the PAWS-X and MLDOC datasets mainly achieved negative relative performance. However, this problem was found to be alleviated after introducing additional high-resource datasets (e.g. *PolyPrompt*+Expand).

(3) For low-resource languages, *PolyPrompt* with in-language prompts will bring more gains, while cross-lingual prompts bring more gains when introducing high-resource training datasets. From Fig. 7, we can observe that *PolyPrompt* with in-lingual prompts outperform with cross-language prompts in low-resource languages. However, when the external English dataset was introduced (*PolyPrompt+Expand*), cross-language prompts have more gains in both low and high resource languages. With the introduction of multilingual datasets (*PolyPrompt+Expand+XLSum*), the relative advantages of cross-lingual prompts increased.

E Dataset Bias

Dataset-level Features We also obtain the dataset-level features. Given a dataset D and a feature p (e.g. qLen), the dataset-level feature can be defined as:

$$\phi_p(D) = \frac{1}{|D^{te}|} \sum_{d \in D^{te}} \phi_p(d), \tag{3}$$

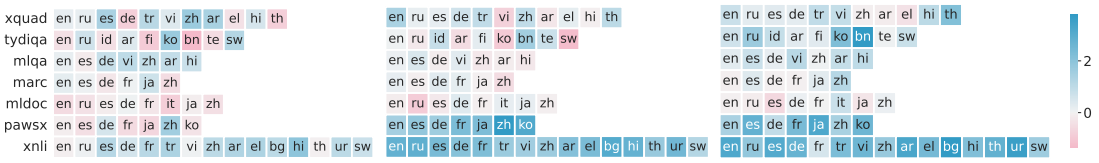
where d is a sample of the test set $D^{te} \in D$, and $\phi_p(\cdot)$ is a function that computes the feature value for a given sample. For example, $\phi_{\text{qLen}}(\text{MLQA})$ denotes the average question length of the MLQA.

Dataset bias is measured by ϕ_p, the dataset-level feature defined in Eq. 3. Tab. 8 shows five target datasets explored in Sec. 5.2.

F Prompt Template

Tab. 5 presents the cross-lingual (English) prompt templates explored in this work. We designed 5 templates for each of the 7 tasks.

[6] https://simple.wikipedia.org/wiki/Wikipedia:List_of_1000_basic_words



(a) CL-IL *PolyPrompt* (b) CL-IL *PolyPrompt+Expand* (c) CL-IL *PolyPrompt+Expand+XLSum*

Figure 7: The relative performance improvement at the language level for *PolyPrompt* with cross-lingual prompts (CL) versus in-lingual prompts (IL). Languages are sorted in descending order according to the sample size of the languages in the mT5 training set (high-resource to low-resource from left to right). The bluer (redder) the color, the greater the improvement (decrease) of CL over IL.

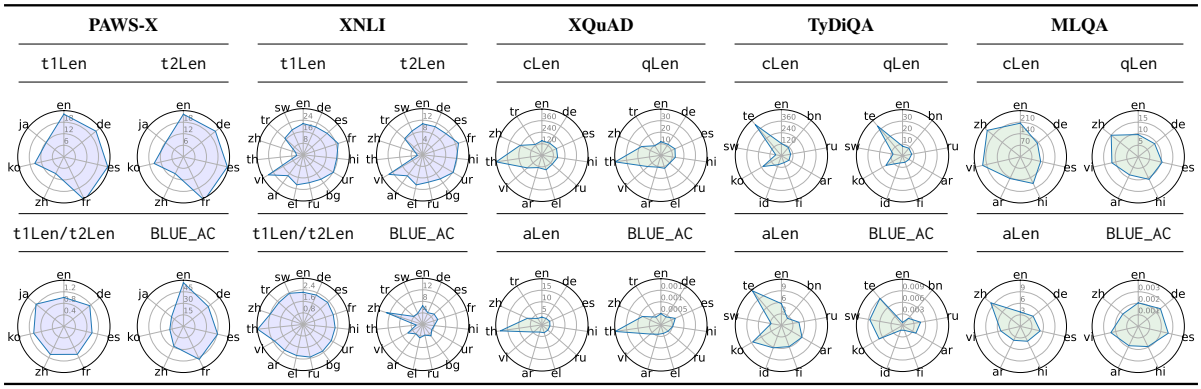


Figure 8: Dataset bias characterized by ϕ_p defined in Eq. 3 (the average of feature values over a specific language of the dataset).

Task	Prompt Template
XQuAD	(1) Answer the question based on the paragraph. | Question: [Q] | Paragraph: [C] (2) Answer the question based on the information contained in the paragraph. | Question: [Q] | Paragraph: [C] (3) [C] | With reference to the above context, [Q] (4) I have always wondered: [Q] | I searched Wikipedia and this is what I found. What's the answer? | [C] (5) Context: [C] | I am trying to figure out the answer to the question from the above context. Can you tell me the answer? | Question: [Q] Answer:
TyDiQA	(1) Answer the question based on the paragraph. | Question: [Q] | Paragraph: [C] (2) Answer the question based on the information contained in the paragraph. | Question: [Q] | Paragraph: [C] (3) I have always wondered: [Q] | I searched Wikipedia and this is what I found. What's the answer? | [C] (4) [C] | With reference to the above context, [Q] (5) Context: [C] | I am trying to figure out the answer to the question from the above context. Can you tell me the answer? | Question: [Q] Answer:
MLQA	(1) Answer the question based on the paragraph. | Question: [Q] | Paragraph: [C] (2) Answer the question based on the information contained in the paragraph. | Question: [Q] | Paragraph: [C] (3) Context: [C] | I am trying to figure out the answer to the question from the above context. Can you tell me the answer? | Question: [Q] Answer: (4) [C] | With reference to the above context, [Q] (5) I have always wondered: [Q] | I searched Wikipedia and this is what I found. What's the answer? | [C]
XNLI	(1) Answer the question based on paragraph 1 and paragraph 2. | Question: Do Paragraph 1 and Paragraph 2 mean the same thing? | (A) Yes. (B) No. (C) Maybe. | Paragraph 1: [T1] | Paragraph 2: [T2] (2) [T1] | Based on the previous passage, is it true that [T2]? Yes, no, or maybe? (3) Suppose [T1] Can we infer that [T2]? | Option: (A) Yes (B) No (C) Maybe? (4) Paragraph: [T1] | Question: Does this imply that "[T2]"? | Yes, no, or maybe? (5) Given that [T1] Therefore, it must be true that [T2]? Yes, no, or maybe?
PAWS-X	(1) Answer the question based on paragraph 1 and paragraph 2. | Question: Do Paragraph 1 and Paragraph 2 express the same meaning? | (A) Yes. (B) No. | Paragraph 1: [T1] | Paragraph 2: [T2] (2) Paragraph 1: [T1] | Paragraph 2: [T2] | Question: Do Paragraph 1 and Paragraph 2 express the same meaning? | Yes or No? (3) Suppose [T1] Can we infer that [T2]? | Option: (A) Yes (B) No? (4) [T1] | Based on the previous passage, is it true that [T2]? Yes or no? (5) Given that [T1] Therefore, it must be true that [T2]? Yes or no?
MARC	(1) Answer the question based on the review. | Question: Can we conclude that the buyer is satisfied with the product based on his review? | (A) Yes. (B) No. | Review: [T1] (2) I am reading a review that says [T1]. Do you think the review is positive or negative? (3) Review: [T1] | Did the reviewer find this product good or bad? (4) Review: [T1] | Is this review positive or negative? (5) Review: [T1] | Is the buyer satisfied with the product purchased?
MLDOC	(1) Answer the question based on the news. | Question: What topic does this news belong to? | (A) Business and Industry. (B) Economy. (C) Government and Society. (D) Market. | News: [T1] (2) Is this a piece of news regarding (A) Business and Industry (B) Economy (C) Government and Society (D) Market? | [T1] (3) Article: [T1] | Which of the following sections of a newspaper would this article likely appear in? | Options: (A) Business and Industry (B) Economy (C) Government and Society (D) Market (4) What topic does this news belong to? | (A) Business and Industry. (B) Economy. (C) Government and Society. (D) Market. | News: [T1] (5) Would you recommend the following article to a (A) Business and Industry (B) Economy (C) Government and Society (D) Market? | [T1]

Table 5: The cross-lingual (English) prompt templates studied in this work. "Q", "C", "T1", and "T2" denotes the placeholders for question, context, sentence$_1$, and sentence$_2$ field, respectively.