Language Models are Few-shot Multilingual Learners

Genta Indra Winata[1*], Andrea Madotto[1,3*], Zhaojiang Lin[1],
Rosanne Liu[2,3], Jason Yosinski[3], Pascale Fung[1]
[1]The Hong Kong University of Science and Technology
[2]Google Brain [3]ML Collective
{giwinata, amadotto, zlinao}@connect.ust.hk

Abstract

General-purpose language models have demonstrated impressive capabilities, performing on par with state-of-the-art approaches on a range of downstream natural language processing (NLP) tasks and benchmarks when inferring instructions from very few examples. Here, we evaluate the multilingual skills of the GPT and T5 models in conducting multi-class classification on non-English languages without any parameter updates. We show that, given a few English examples as context, pre-trained language models can predict not only English test samples but also non-English ones. Finally, we find the in-context few-shot cross-lingual prediction results of language models are significantly better than random prediction, and they are competitive compared to the existing state-of-the-art cross-lingual models and translation models.

1 Introduction

The progress in language model (LM) pre-training (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019; Liu et al., 2019a; Brown et al., 2020; Liu et al., 2020a; Lewis et al., 2020; Raffel et al., 2020; Gao et al., 2020a) has led to the possibility of conducting few-shot learning, that is, learning a new task using a small number of examples without any further training or gradient computation. Few-shot learning alleviates the cost for extensive labeled data, which is beneficial since collecting high-quality labeled data is resource-intensive and expensive. It also reduces the cost for model fine-tuning, which requires tremendous GPU or TPU resources. Few-shot learning can be seen as a *one-for-all plug-and-play* computational model that can be applied to various natural language tasks, from sentiment analysis for text classification to story generation, provided only a small context (Brown et al., 2020).

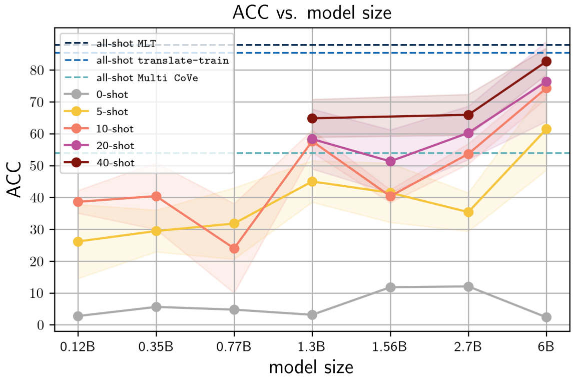

Figure 1: Accuracy vs. model size on English-Spanish MNLU dataset. Cross-lingual in-context learning with LMs (i.e., context with few English examples tested on Spanish sentences) performs as well as models trained in cross-lingual setting (Liu et al., 2020b) and translation baselines.

The idea of few-shot learning is also relevant to address the low-resource issue in non-English languages. Few-shot learning has been applied to NLP tasks (Brown et al., 2020; Madotto et al., 2020b; Lu et al., 2021; Perez et al., 2021; Liu et al., 2021a,b; Cahyawijaya et al., 2021a). Common approaches to solve the low-resource issue are to pre-train models with self-supervised learning using unlabelled monolingual text data collected from various resources available online (Wilie et al., 2020; Le et al., 2020; Martin et al., 2020; Eddine et al., 2020; Nguyen and Nguyen, 2020; Scheible et al., 2020; Bhattacharjee et al., 2021; Lee et al., 2020; Cahyawijaya et al., 2021b; Park et al., 2021) and then apply pre-training on the source language and fine-tune on the target languages (Schuster et al., 2019; Lin et al., 2019; Winata et al., 2019, 2021; Pfeiffer et al., 2020; Zheng et al., 2021; Lin et al., 2021b). Conversely, the few-shot learning does not need any training from the source and target languages. Figure 1 shows how it is possible to utilize pre-trained models on non-English languages, such as Spanish, as the performance is not random,

[*] Equal contribution

Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 1–15
November 11, 2021. ©2021 Association for Computational Linguistics

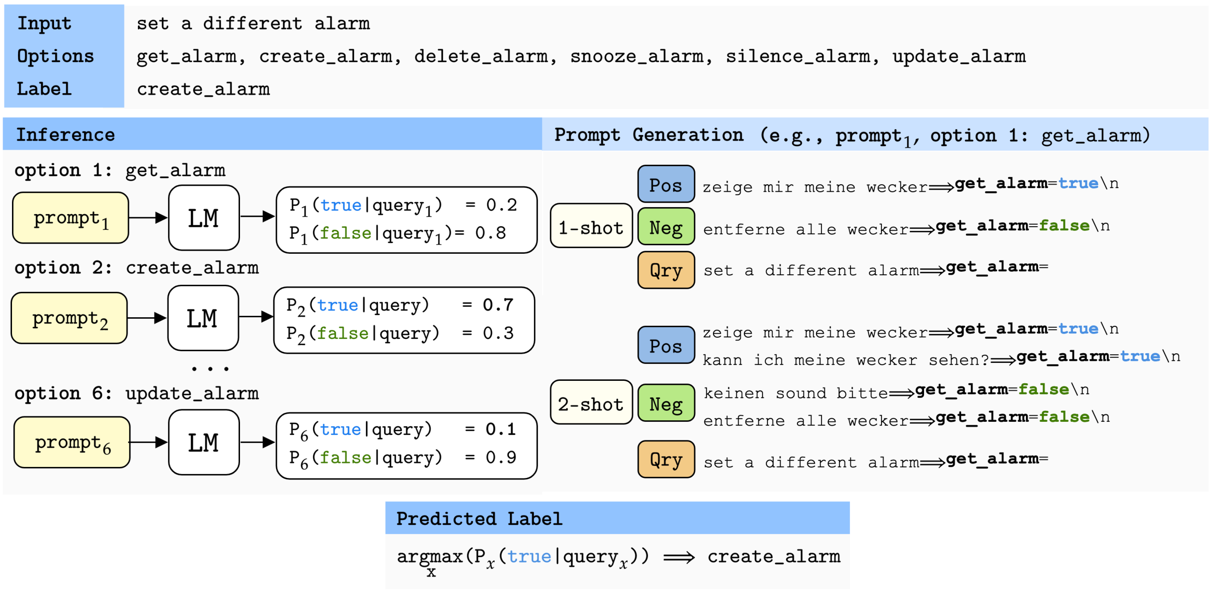
Input
set a different alarm
Options
get_alarm, create_alarm, delete_alarm, snooze_alarm, silence_alarm, update_alarm
Label
create_alarm
Inference
option 1: get_alarm
prompt$_1$
LM
P$_1$(true|query$_1$) = 0.2
P$_1$(false|query$_1$)= 0.8
option 2: create_alarm
prompt$_2$
LM
P$_2$(true|query) = 0.7
P$_2$(false|query) = 0.3
...
option 6: update_alarm
prompt$_6$
LM
P$_6$(true|query) = 0.1
P$_6$(false|query) = 0.9
Prompt Generation (e.g., prompt$_1$, option 1: get_alarm)
1-shot
Pos
zeige mir meine wecker⟹get_alarm=true\n
Neg
entferne alle wecker⟹get_alarm=false\n
Qry
set a different alarm⟹get_alarm=
2-shot
Pos
zeige mir meine wecker⟹get_alarm=true\n
kann ich meine wecker sehen?⟹get_alarm=true\n
Neg
keinen sound bitte⟹get_alarm=false\n
entferne alle wecker⟹get_alarm=false\n
Qry
set a different alarm⟹get_alarm=
Predicted Label
$\underset{x}{\text{argmax}}(P_x(\text{true}|\text{query}_x)) \Longrightarrow$ create_alarm

Figure 2: Example of the inference and query generation on the few-shot learning, where the source language and target language are German and English, respectively.

and the performance increases as the models are given more samples. We conjecture that pre-trained models may be able to adapt to languages that are similar to English. However, for many language tasks, it is difficult to collect a large supervised training dataset as language experts (e.g., linguists or native speakers) are required to annotate the data.

Another line of work is to apply cross-lingual transfer on English with the same task as the target languages (Ponti et al., 2018; Artetxe and Schwenk, 2019; Liu et al., 2019b; Lauscher et al., 2020; Liu et al., 2020b, 2021c; Chen et al., 2021). However, such methods still need to apply a fine-tuning step to update the model for fast adaptation, which can be challenging for large pre-trained models – some models require substantial memory capacity – since the models have to be trained on high-performing machines. Different from the aforementioned method, in-context learning using an LM does not allow any parameter updates. Thus, the process does not need to compute and store the gradients for backward propagation.

In this work, we investigate the practicality of applying few-shot learning in the multilingual setting for four languages, English, French, German, and Spanish, on natural language understanding intent prediction tasks using publicly available LMs that are mainly trained on English data. We show that, given a few English examples as context, pre-trained LMs can predict not only English test samples, but also non-English ones (Figure 2). To the best of our knowledge, no existing works have studied these tasks in multilingual settings. We conjecture that the English LMs can still produce good results on languages that are closely related to English. We construct the inference for the multi-class prediction setup by extending the idea from Madotto et al. (2020b) of applying multiple binary predictions on each class. Instead of guiding the model to generate *true* or *false* like in their work, which is not consistent and sometimes generates other words –, we introduce *maximum confidence prediction*. This method considers the confidence of predicting a certain label to provide a prediction. We design this as a multiple-choice task in which the confidence of the prediction for all possible classes is compared. Each class's confidence score is computed by normalizing the logits of generating the *next boolean token* given the prompt as the context. This method is considered to be more scalable than the simple k-way few-shot learning, where we need to put all data in a single prompt, since we only have a fixed maximum sequence length and, in the deployment, each forward step can be run in parallel to speed up the process. To increase the difficulty of the challenge, we also propose a cross-lingual task, where the context and query are in different languages.

Overall, we find that conditional generative LMs, such as the GPT-2 (Radford et al., 2019), GPT$_{\text{NEO}}$ models (Gao et al., 2020a), and T5 models (Raffel et al., 2020) have the capability to predict non-English languages, and adding more shots and using larger models achieves a substantial increment in performance, making it significantly better than random, which indicates the models are able to understand the prompt. We only focus on GPT and T5 models. T5 models do not perform as well as GPT models, which might be caused by the pre-training strategy. Experimental results in the cross-lingual setting demonstrate that pre-trained LMs make correct predictions. To summarize, our contributions are as follows:

- We study few-shot learning in the multilingual setting on four languages without any gradient updates. We use the publicly available GPT and T5 LMs, and compare the results to those from the zero-shot and fine-tuning approaches.
- We propose a simple and straightforward approach to perform few-shot learning on multi-class classification by applying binary prediction and considering the confidence of predicting the boolean tokens.
- We display the zero-shot, one-shot, and many-shot proficiency of the LMs in the cross-lingual setting when the language of the prompt is different from the target language.

2 Few-shot Multilingual Learners

First, we briefly define the notation of the input and output of the task, and then we introduce our method to design prompts for few-shot in-context learning. [1]

2.1 Notation and Tasks

Let us define D as the distribution over the dataset and P as the prompt that we use as the input of the LM θ. The prompt $P = [D_{pos}, D_{neg}, Q]$ is a concatenation of few-shot samples: positive samples D_{pos}, negative samples D_{neg}, and the query Q, where $D_{pos}, D_{neg} \sim D$. D_{pos} is a sample with a label that is the same as the query, and D_{neg} is a sample that is taken from the dataset D with a label other than the query. θ takes P as the input of the model, and the LM generates a word y. We define the task $T_{s \rightarrow t}$, where s is the source language and t is the target language.

In this paper, we focus on the intent detection task in the monolingual and cross-lingual settings. In the monolingual setting, the source language is the same as the target language, and in the cross-lingual setting, we take the source language as different from the target language ($s \neq t$). We design our task as a multiple-choice problem, in which each sample has a label $l \in L$, where L is the set of possible labels. We predict the boolean (true or false) for each sample and take the highest prediction confidence.

2.2 Prompt Generation

We define the task by designing prompts to perform few-shot learning. We design our task as a binary classification for multi-class prediction by following Madotto et al. (2020b). The idea is to guide the model to predict the boolean tokens, *true* and *false*. We examine the usage of two types of LMs, GPT and T5 models, and we construct prompts specific to each model. We use a specific way to probe the LMs to perform the few-shot prediction since they are trained with different learning objectives. Table 1 shows the format of the prefix we use for the GPT and T5 models. X_i is one of the few-shot samples, and X_i^* is the sample from other classes. For the GPT models, we only input the prefix by concatenating positive and negative samples with the query. Specifically for the T5 models, we add an additional token after the query and let the model predict that particular token during the generation step.

Model	Prompt
GPT	[SAMPLES] $Q \rightarrow$
T5	[SAMPLES] $Q \rightarrow$ [MASK]

[SAMPLES]	
Format	Example
$X_1 \rightarrow$ true\n	zeige mir meine wecker=>get_alarm=true\n
$X_1^* \rightarrow$ false\n	entferne alle wecker=>get_alarm=false\n
...	...
$X_k \rightarrow$ true\n	kann ich meine wecker sehen?=>get_alarm=true\n
$X_k^* \rightarrow$ false\n	keinen sound bitte=>get_alarm=false\n

Table 1: Prompt format given a few German examples as context.

Figure 2 shows an example of how we generate the prompt in k-shot settings. We create L prompts and apply L forward steps for each sample. For each prompt, k positive and negative samples are randomly drawn from the dataset. It is worthwhile to note that the sampling method is similar to k-way few-shot learning, but the samples are not merged

[1] The code is released at https://github.com/gentaiscool/few-shot-lm.

into a single prompt. We do this because we want to give more shots as the prompt to the LMs as they have a limitation on the number of tokens they can accept as input (1,024 tokens in GPT-2$_{\text{XL}}$ and 2,048 tokens in GPT$_{\text{NEO}}$). We add a special token `\n` as a separator between each sample, as shown in Table 1.

2.3 Maximum Confidence Prediction

To get the final prediction of each sample, first, we compute the score of predicting the next boolean (true or false [2]) given the prompt X_i for label i: $P_\theta(y = \texttt{true}|X_i)$ and $P_\theta(y = \texttt{false}|X_i)$ from the prediction distribution. Then, we normalize the score to get the probability of generating the *true* token to measure how much confidence the LM has to predict label i. We collect all the confidence scores over all label options and choose the highest confidence score among them, as follows:

$$\text{MC}(X, L) = \underset{i \in L}{\text{argmax}} \frac{P_\theta(y = \texttt{true}|X_i)}{\sum_b P_\theta(y = b|X_i)}, \quad (1)$$

where $b \in \{\texttt{true}, \texttt{false}\}$. We take the label with the highest confidence score as MC(X, L).

2.4 Choices of Samples

For in-context learning, choosing the order of samples is essential (Lu et al., 2021). Here, we examine the impact of the order of the samples. We construct the probing set in two ways: (1) shuffle the few-shot samples and measure the variance in performance after changing their order, and (2) arrange the positive samples before the negative samples. We find that the latter works well, specifically on the T5 models.

3 Baselines

In this work, we compare the few-shot learning performance with other common approaches: zero-shot, zero-shot cross-task, and fine-tuning.

3.1 Zero-shot Cross-Task

One way to solve zero-shot prediction is by using entailment models to calculate the entailment score between sequences and labels. Given a pre-trained LM ψ with an entailment head, a set of hypotheses H, and possible labels L, the model accepts two inputs, the hypothesis $h \in H$ and label $l \in L$, and generates the entailment score given any combinations of the hypothesis and label $P_\psi(y = \texttt{entail}|h, l)$:

$$\text{ES}(H, L) = \underset{h,l \in \{H,L\}}{\text{argmax}} P_\psi(y = \texttt{entail}|h, l). \quad (2)$$

3.2 Zero-shot In-Context Learning

This approach is very similar to our few-shot approach. It does not need any samples, and the model is only given natural language instruction. However, instead of using the prompt like in the few-shot setting, we can set up the prompt in a question-and-answer (Q&A) format as follows:

$$\texttt{Q:}\ \text{Is '}\texttt{<INTENT>}\text{' the intent of '}\texttt{<TEXT>}\text{'?}\ \texttt{A:}. \quad (3)$$

3.3 Fine-tuning

Fine-tuning is the most common approach to updating a pre-trained model's weights when training with a labeled dataset. The advantage of this approach is strong performance since we give supervised signals with the correct labels to the model. For fine-tuning, we use the same sets of few-shot samples as in the in-context learning. In Section 4.2, we provide the hyper-parameters used in the experiments.

4 Experiments

4.1 Datasets and Metrics

We use an English natural language understanding (NLU) dataset, SNIPS (Coucke et al., 2018), and two multilingual NLU datasets, MTOP (Li et al., 2021) and Multilingual NLU (MultiNLU) (Schuster et al., 2019). MTOP includes four languages, English (`en`), French (`fr`), German (`de`), and Spanish (`es`), and Multilingual NLU includes two languages, English (`en`) and Spanish (`es`). We measure the model performance by calculating the average and standard deviation of the accuracy with three runs.

4.2 Experiment Settings

We set up the experiment in two settings: monolingual and cross-lingual. In the monolingual setting, we test the ability of the model to conduct few-shot in-context learning on four languages: English (`en`), French (`fr`), German (`de`), and Spanish (`es`). In the cross-lingual setting, we test its ability to predict a query from a non-English language

[2] Notice that some tokenizers (e.g., T5) splits "true" in two sub-tokens. We compute the score of the first sub-token only since it is significantly different for the two label (i.e. "tr" and "fal").

Models	**SNIPS**	**MTOP**				**MultiNLU**	
	en	**de**	**en**	**es**	**fr**	**en**	**es**
Random	14.29	15.07	15.25	15.55	14.36	8.33	8.33
Full-training SOTA	$99.00^\ddagger$	$88.80^\dagger$	$94.00^\dagger$	$90.10^\dagger$	$89.60^\dagger$	99.11^*	98.90^*
Zero-shot Cross-Task Prediction							
BART$_{\text{LARGE}}$ `0.4B`	74.43	24.80	43.41	36.06	24.77	65.60	34.77
XLM-R$_{\text{LARGE}}$ `0.6B`	68.00	54.30	53.37	51.67	51.99	77.79	66.35
Few-shot Learning (K-shot)							
GPT-2 `0.1B`	39.33 ± 8.58	40.03 ± 6.34	35.46 ± 0.92	36.18 ± 2.12	41.16 ± 5.65	51.59 ± 12.83	37.56 ± 7.14
GPT-2$_{\text{MEDIUM}}$ `0.3B`	65.71 ± 2.80	52.94 ± 5.12	63.35 ± 3.01	54.33 ± 4.75	50.6 ± 2.44	72.21 ± 14.88	50.25 ± 4.99
GPT-2$_{\text{LARGE}}$ `0.8B`	71.43 ± 10.27	50.94 ± 6.63	59.70 ± 4.50	52.38 ± 2.65	44.75 ± 1.11	62.36 ± 13.82	58.04 ± 5.28
GPT-2$_{\text{XL}}$ `1.6B`	78.43 ± 3.16	78.43 ± 3.16	73.93 ± 1.21	56.61 ± 2.02	45.21 ± 2.54	79.04 ± 5.05	64.74 ± 7.64
GPT$_{\text{NEO}}$ `1.3B`	84.19 ± 2.78	67.17 ± 2.50	82.40 ± 1.90	73.51 ± 0.95	66.3 ± 1.29	89.70 ± 1.28	85.77 ± 2.53
GPT$_{\text{NEO}}$ `2.7B`	91.24 ± 0.68	71.57 ± 5.94	81.51 ± 0.39	76.94 ± 0.83	70.31 ± 1.99	83.76 ± 3.14	87.82 ± 1.55
GPT$_{\text{NEO-J}}$ `6B`	**93.38 ± 0.76**	**80.97 ± 3.21**	**89.66 ± 0.50**	**84.18 ± 0.32**	**85.04 ± 1.18**	**94.32 ± 1.14**	**88.54 ± 6.18**
T5$_{\text{LARGE}}$ `0.8B`	23.57 ± 8.93	41.84 ± 7.63	36.02 ± 5.26	49.49 ± 6.32	40.41 ± 5.97	37.57 ± 15.23	21.20 ± 6.51
T5$_{\text{3B}}$ `3B`	46.52 ± 6.69	50.81 ± 6.45	46.17 ± 4.06	46.45 ± 4.39	44.38 ± 0.22	31.46 ± 18.18	31.60 ± 14.90
GPT$_{\text{NEO}}$ `2.7B` (ordered)	86.71 ± 1.62	55.69 ± 3.45	55.12 ± 4.01	50.77 ± 4.41	50.70 ± 2.47	63.33 ± 7.14	61.51 ± 1.63
T5$_{\text{LARGE}}$ `0.8B` (ordered)	25.90 ± 18.51	63.06 ± 4.56	51.92 ± 3.90	62.71 ± 6.30	55.91 ± 3.82	38.97 ± 14.80	63.10 ± 4.46
T5$_{\text{3B}}$ `3B` (ordered)	**93.00 ± 3.00**	**74.11 ± 2.69**	**65.03 ± 1.87**	**66.97 ± 1.35**	**68.89 ± 2.51**	**80.12 ± 3.95**	**86.60 ± 2.40**
Fine-tuning (40-shot)							
mBERT `0.2B`	88.57 ± 3.14	25.21 ± 2.31	41.44 ± 5.59	33.82 ± 10.08	16.54 ± 5.54	84.88 ± 1.59	87.87 ± 3.29
XLM-R$_{\text{BASE}}$ `0.3B`	87.95 ± 1.39	27.47 ± 11.90	37.03 ± 5.11	27.16 ± 5.51	13.8 ± 6.50	77.06 ± 3.16	74.85 ± 1.53

Table 2: Zero-shot and few-shot results in the monolingual setting. The SOTA results are taken from †Li et al. (2021), ‡Qin et al. (2019), and *Schuster et al. (2019).

with the English context (`en`$\rightarrow$`XX`). In the few-shot in-context learning, we use k-way-few-shot classification, taking k samples. For each model, we take $k \in [0, 5, K]$, where $K \leq 40$ is the largest number of few-shot samples that can be passed to the model as input and is divisible by 10 without exceeding the maximum input token limit. We utilize an NVIDIA Tesla V100 16GB GPU to run the inference so that the model is ensured to fit in a single GPU, and we use 16-bit precision.

Model details We run experiments on a variety of publicly available models:[3] four sizes of GPT-2 models (0.1B, 0.3B, 0.8B and 1.6B), three sizes of GPT$_{\text{NEO}}$ models (1.3B, 2.7B, and 6B), and two sizes of T5 models (0.8B and 3B). Table 3 shows the details of each pre-trained model.

Baselines We use the same sets of few-shot samples for the baselines. We run fine-tuning on the pre-trained models mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), and also compare our models with the zero-shot cross-task models using pre-trained models XLM-R, fine-tuned on XNLI (Conneau et al., 2018), and BART, fine-tuned on MNLI (Williams et al., 2018);[4] a random baseline; and state-of-the-art results reported on each dataset. For the finetuning, we use a learning rate of 5e-5 with a decay of 0.9 for every epoch, and a batch size of 32. We apply an early stopping after 5 epochs without any improvement on the validation set.

Model Name	n_{params}	n_{layers}	n_{hidden}	n_{ffn}
GPT-2	0.1B	12	768	
GPT-2$_{\text{MEDIUM}}$	0.3B	24	768	-
GPT-2$_{\text{LARGE}}$	0.8B	36	1,280	-
GPT-2$_{\text{XL}}$	1.6B	48	1,600	-
GPT$_{\text{NEO}}$	1.3B	24	2,048	-
GPT$_{\text{NEO}}$	2.7B	32	2,560	-
GPT$_{\text{NEO-J}}$	6B	28	4096	16,384
T5$_{\text{LARGE}}$	0.8B	24	1,024	4,096
T5$_{\text{3B}}$	3B	24	1,024	16,384

Table 3: Model architecture.

[3] The models except GPT$_{\text{NEO-J}}$ are taken from `https://huggingface.co/`. The GPT$_{\text{NEO-J}}$ model is taken from `https://github.com/kingoflolz/mesh-transformer-jax/`

[4] The XLM-R model fine-tuned with XNLI data can be accessed at `https://huggingface.co/joeddav/xlm-roberta-large-xnli`. The BART model fine-tuned with MNLI data can be accessed at `https://huggingface.co/facebook/bart-large-mnli`

Models	MTOP			MultiNLU
	en→de	en→es	en→fr	en→es
Fine-tuning (all-shot on source language, zero-shot on target language)				
Seq2Seq w/ CRISS (Li et al., 2021)	36.10	48.60	46.60	-
Seq2Seq w/ XLM-R (Li et al., 2021)	42.30	50.30	43.90	-
NLM (Liu et al., 2021d)	54.91	59.99	58.16	-
X2Parser (Liu et al., 2021d)	**56.16**	**60.30**	**58.34**	-
Multi CoVe (Schuster et al., 2019)	-	-	-	53.89
Translate-Train (Liu et al., 2020b)	-	-	-	85.39
MTL (Liu et al., 2020b)	-	-	-	**87.88**
Few-shot Learning (K-shot)				
GPT-2 0.1B	23.89 ± 1.52	27.10 ± 3.19	26.14 ± 0.54	38.60 ± 3.54
GPT-2$_{\text{MEDIUM}}$ 0.3B	39.61 ± 5.42	41.81 ± 4.66	42.40 ± 3.84	40.40 ± 10.48
GPT-2$_{\text{LARGE}}$ 0.8B	30.94 ± 4.45	34.69 ± 6.50	33.04 ± 4.56	23.99 ± 14.02
GPT-2$_{\text{XL}}$ 1.6B	42.88 ± 4.94	48.43 ± 4.42	50.67 ± 4.50	51.31 ± 9.87
GPT$_{\text{NEO}}$ 1.3B	56.14 ± 2.75	63.14 ± 2.52	60.25 ± 3.32	64.82 ± 5.94
GPT$_{\text{NEO}}$ 2.7B	58.27 ± 1.28	64.79 ± 1.69	62.30 ± 1.60	65.91 ± 6.42
GPT$_{\text{NEO-J}}$ 6B	**79.41 ± 1.18**	**81.57 ± 0.83**	**77.85 ± 1.63**	**82.66 ± 4.19**
T5$_{\text{LARGE}}$ 0.8B	37.14 ± 5.44	38.14 ± 3.20	33.53 ± 4.85	14.95 ± 16.34
T5$_{\text{3B}}$ 3B	35.35 ± 7.07	34.64 ± 6.21	37.26 ± 8.68	14.11 ± 14.01
GPT$_{\text{NEO}}$ 2.7B (ordered) 0.8B	42.23 ± 3.24	48.62 ± 2.60	46.30 ± 3.02	47.83 ± 5.73
T5$_{\text{3B}}$ (ordered) 3B	**52.23 ± 4.29**	**52.74 ± 3.20**	**49.72 ± 5.37**	**50.42 ± 6.01**

Table 4: Few-shot results in the cross-lingual setting on MTOP and MultiNLU datasets.

5 Results and Analysis

5.1 Model Performance

Tables 2 and 4 show the results in the monolingual and cross-lingual settings, respectively. The tables show that the performance improvement is highly related to the size of the pre-trained model, and the performance gap between the fully trained state-of-the-art model and the few-shot learning models decreases when we use larger models, indicating the usefulness of utilizing models of bigger sizes. The performance of the models with few-shot learning is considered promising as they are not trained at all, and the best model's performance gap with the fine-tuned model is less than 10%.

Few-shot vs. Fine-tuning. Comparing the performance of generative models to fine-tuning, it is clear that we can achieve higher accuracy without any training. However, in this experiment, we acknowledge GPT and T5 models we use for in-context learning are larger than the models we fine-tune, and few-shot learning is much more efficient since the models are not required to store the intermediate memory. In terms of inference speed, the few-shot models require more time to run an inference step, which may cause a bottleneck when the number of few-shot samples is relatively large. This is the limitation of this method, and reducing the inference time is an open research area to improve the efficiency of in-context learning.

Zero-shot cross-task baselines. Surprisingly, the zero-shot cross-task models are able to predict the samples much better than the random baseline, particularly on English tasks. Overall, the XLM-R$_{\text{LARGE}}$ model performs better than the BART$_{\text{LARGE}}$ models in all tasks except SNIPS.

GPT vs. T5 models. In general, the GPT models outperform the T5 models in all language pairs and datasets in a head-to-head comparison: Both GPT-2$_{\text{LARGE}}$ and T5$_{\text{LARGE}}$ have a similar number of parameters (0.8B), but they have a significant performance difference. A similar pattern can also be observed on larger models, such as GPT$_{\text{NEO}}$ 2.7B and T5$_{\text{3B}}$ 3B. Although the T5 models perform worse than the GPT models, they do not have a maximum token size for the input, as the GPT models do, which is one of the advantages of using them. On the other hand, we find that changing the sample order tremendously affects the performance of the T5 models. As shown in Tables 2 and 4, the performance increases substantially when we sort the few-shot samples based on their label (i.e., first all positive and then all negative examples). Conversely, the GPT models suffer loss in performance. Thus, we can make the conclusion that changing the sample order may produce high variance in the results, as also shown in (Lu et al., 2021).

Effectiveness on non-English languages. Based on the results, the performance of the models is lower in the non-English languages than in English. These results are expected since the pre-trained models are mostly trained on English data. However, the differences in performance are marginal. This finding may indicate that our few-shot learning method can be effectively utilized for languages that are in the same language family as English, such as French, German, and Spanish, but this will require further investigation in the future.

Cross-lingual results. Based on the results in Table 4, we can see that the generative models are able to use the context from English to predict the sample in non-English languages. The cross-lingual setting is considered harder than the monolingual one since the models need to contextualize and understand the source and target languages to predict the test samples correctly. In general, the trend of the results in the cross-lingual setting is similar to the monolingual setting. In the MTOP dataset, we find that the models generally achieve higher performance for **en**→**es** than for the other two target languages (**de** and **fr**). In MultiNLU, our $\text{GPT}_{\text{NEO-J}}$ closes the gap with the existing state-of-the-art baseline with fine-tuning from Liu et al. (2020b) underperforming it only by a close margin of around 4.2%, and the $\text{GPT}_{\text{NEO-J}}$ performance is only less than 3% worse than that of the Translate-Train model. These results show a promising new direction in the zero-shot cross-lingual research that can be applied to other datasets and language pairs.

5.2 Ablation Study

To further understand how much data we need for the in-context learning, we conduct experiments with different numbers of few-shot samples, including zero-shot experiments on the MTOP and MultiNLU datasets.

MTOP dataset. Figures 3, 4, 5, and 6 illustrate the results with different numbers of samples on the MTOP dataset in the monolingual setting. We show a different set of k-shot results for each model according to the maximum samples that can be used in the model as input. The results consistently improved as the number of shots increases. Interestingly, the QA style's zero-shot strategy can outperform random prediction only on two or three models in each language, and the others are worse. The fine-tuning results on MTOP are thus far worse than those of few-shot learning.

MultiNLU dataset. Figures 7 and 8 illustrate the results with different numbers of samples on the MultiNLU dataset in the monolingual setting. The results on MultiNLU for the models with fine-tuning are closer to those of few-shot learning than those on the MTOP dataset. The reason may be the number of labels that the MTOP dataset has compared to MultiNLU. As a result, the zero-shot performance on the GPT models is sometimes worse than that of the random baseline.

6 Related Work

6.1 Few-shot In-Context Learning

Recent work on few-shot in-context learning uses LMs to solve NLP tasks (Petroni et al., 2019; Brown et al., 2020; Gao et al., 2020b; Madotto et al., 2020b; Zhao et al., 2021; Schick and Schütze, 2021; Lin et al., 2021a). In this approach, we select the appropriate prompts to trigger the LMs to behave so that they can predict the desired output (Liu et al., 2021b). However, the prompts have to be engineered to allow the LM to generate a text appropriate to solve the task. Learning to calibrate the few-shot results is also essential to reduce the model's performance variance (Zhao et al., 2021), and the selection criteria in choosing the prompts are also important (Perez et al., 2021). In another stream of work, Shin et al. (2020); Li and Liang (2021) proposed an automated method to create prompts for a diverse set of tasks by gradient-based tuning instead of manually searching for a good prompt. Using such a method, may allow us to find an optimal prompt easier, it is very difficult to discover the optimal prompts for complicated natural language processing tasks, such as semantic parsing (Liu et al., 2021b).

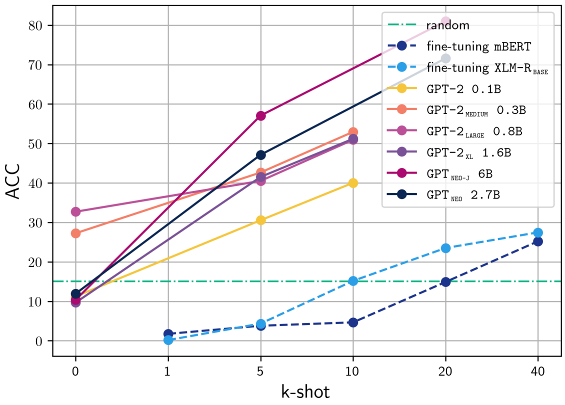

Figure 3: The results on German (de) MTOP dataset with GPT models.

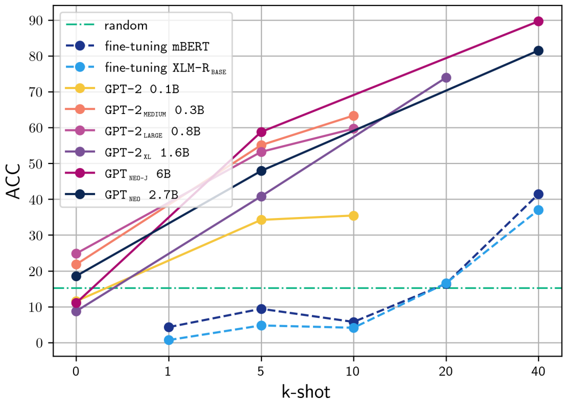

Figure 4: The results on English (en) MTOP dataset with GPT models.

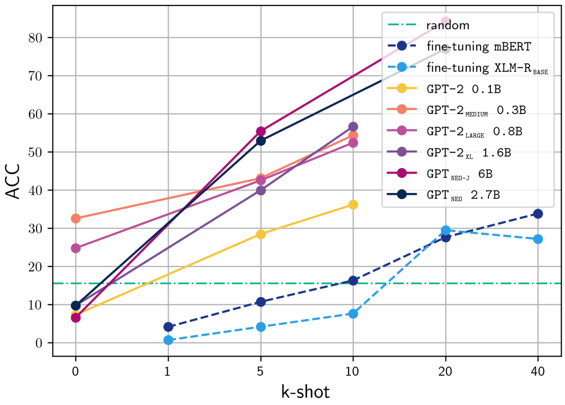

Figure 5: The results on Spanish (es) MTOP dataset with GPT models.

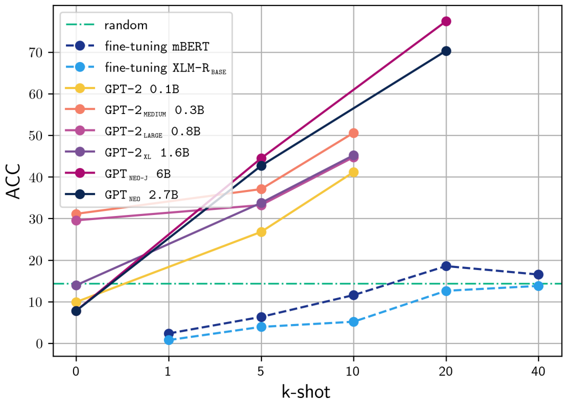

Figure 6: The results on French (fr) MTOP dataset with GPT models.

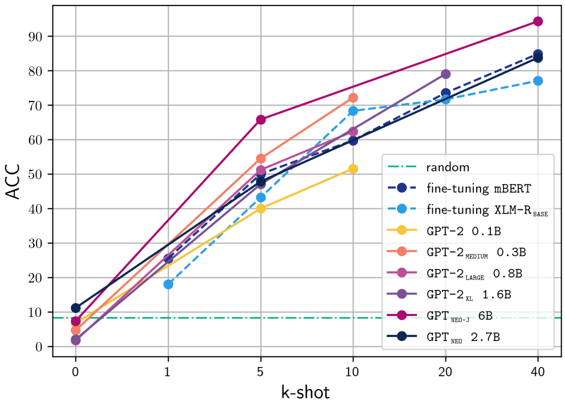

Figure 7: The results on English (en) multilingual NLU dataset with GPT models.

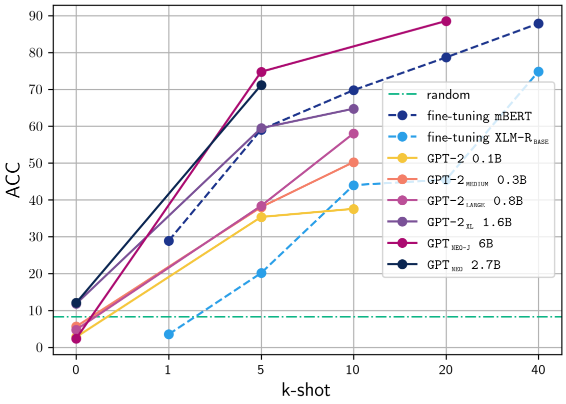

Figure 8: The results on Spanish (es) multilingual NLU dataset with GPT models.

6.2 Pre-trained Language Models

Recent advances in pre-trained LMs have been focused on building pre-trained encoders, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019a), ELMO (Peters et al., 2018), ULMFiT (Howard and Ruder, 2018), ELECTRA (Clark et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020; Goyal et al., 2021), decoder-only models, such as GPT models (Radford et al., 2019; Brown et al., 2020) and encoder-decoder models, such as T5 (Raffel et al., 2020), BART (Lewis et al., 2020), and their multilingual versions, mT5 (Xue et al., 2021) and mBART (Liu et al., 2020a).

Pre-trained encoders have been used to improve the contextualized representations of multilingual systems in various NLP tasks, for example, dialogue systems (Liu et al., 2020b, 2021d; Li et al., 2021), code-switching sequence labeling (Aguilar et al., 2020; Winata et al., 2021; Winata, 2021), and multilingual speech recognition (Datta et al., 2020; Winata et al., 2020). Meanwhile, the pre-trained encoder-decoder models, have been used for various sequence generation tasks, such as summarization (Raffel et al., 2020), conversational agents (Lin et al., 2020b,a; Madotto et al., 2020a; Wu and Xiong, 2020; Hosseini-Asl et al., 2020; Lin et al., 2021b), and knowledge grounding (Chen et al., 2020; Zhao et al., 2020).

7 Limitation and Future Work

More Languages In this paper, we explored only cross-lingual transfer learning from and to Latin-based language (e.g., English to Spanish / French / German). Extending our approach to non-Latin languages (e.g., Thai, Chinese, etc.) is challenging for two reasons: 1) we are currently using English tokenizers which are known to fails, or they assign UNK tokens when prompt with non-Latin characters, and 2) a possible little, or absent, the named entity overlap between the source and target language, which could make the English prompt completely irrelevant. The latter suggests an interesting future work, where we could study the correlation between performance and word (or token) overlapping of the source (`en`) and the target language samples.

More Datasets and Models Intent recognition is an important task, especially in multiple language scenarios. In future work, we plan to include the missing languages of MTOP and MultiNLU, and to add more languages from the MultiATIS++ (Xu et al., 2020) which consists of a total of 9 languages, that is, English, Spanish, German, French, Portuguese, Chinese, Japanese, Hindi, and Turkish. Moreover, to cope with the tokenization issues, we would like to explore multilingual LMs such as MT5 (Xue et al., 2021).

8 Conclusion

This paper demonstrates the multilingual skills of pre-trained LMs, GPT and T5, in conducting in-context learning without parameter updates. This work is our initial attempt to show the effectiveness of in-context learning in the multilingual and cross-lingual setting. It covers four different languages and explores the possibility of conducting efficient inference on low-resource tasks. We find that LMs can predict samples correctly, significantly better than the random prediction, in cross-lingual tasks with no training examples of the target languages. We would like to investigate further the applicability of this method to other tasks and languages in future work.

Acknowledgment

We want to thank Bryan Wilie and Samuel Cahyawijaya for their support in accessing the cloud service. We also sincerely thank Zihan Liu and ML Collective members for helping with the discussion about this project. Finally, we want to thanks the reviewer of the paper for their meaningful comments and suggestions. Given the short time for the camera-ready, we tried our best to improve the final version of the paper.

References

Gustavo Aguilar, Bryan McCann, Tong Niu, Nazneen Rajani, N. Keskar, and T. Solorio. 2020. Char2subword: Extending the subword embedding space from pre-trained models using robust character compositionality. *ArXiv*, abs/2010.12730.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, M Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *arXiv preprint arXiv:2101.00204*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Samuel Cahyawijaya, Genta Indra Winata, Holy Lovenia, Bryan Wilie, Wenliang Dai, Etsuko Ishii, and Pascale Fung. 2021a. Greenformer: Factorization toolkit for efficient deep neural networks.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Leylia Khodra, et al. 2021b. Indonlg:

Benchmark and resources for evaluating indonesian natural language generation. *arXiv preprint arXiv:2104.08200*.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. *arXiv preprint arXiv:2104.08757*.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-grounded pretraining for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

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.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Arindrima Datta, Bhuvana Ramabhadran, Jesse Emond, Anjuli Kannan, and Brian Roark. 2020. Language-agnostic multilingual modeling. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8239–8243. IEEE.

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.

Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020a. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020b. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. 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.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. Kr-bert: A small-scale korean-specific language model. *arXiv preprint arXiv:2008.03979*.

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

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul A Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021a. Leveraging slot descriptions for zero-shot cross-domain dialogue statetracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020a. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021b. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv e-prints*, pages arXiv–2106.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020b. Mintl: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

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. *arXiv preprint arXiv:2107.13586*.

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

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

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019b. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020b. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2021c. X2parser: Cross-lingual and cross-domain framework for task-oriented compositional semantic parsing. *arXiv preprint arXiv:2106.03777*.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2021d. X2Parser: Cross-lingual and cross-domain framework for task-oriented compositional semantic parsing. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 112–127, Online. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020a. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2372–2394.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020b. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing: Findings, pages 1037–1042.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *arXiv preprint arXiv:2105.11447*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. 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 2078–2087.

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

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. *Journal of Machine Learning Research*, 21:1–67.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*.

Timo Schick and Hinrich Schütze. 2021. 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.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Genta Indra Winata. 2021. Multilingual transfer learning for code-switched language and speech neural modeling. *arXiv preprint arXiv:2104.06268*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280.

Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven Hoi. 2020. Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition. *arXiv preprint arXiv:2012.01687*.

Chien-Sheng Wu and Caiming Xiong. 2020. Probing task-oriented dialogue representation from language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5036–5051.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. *arXiv preprint arXiv:2106.08226*.

A Full k-shot Results

This appendix shows the results on few-shot monolingual and cross-lingual settings on SNIPS, MTOP, and multilingual NLU datasets over a different number of samples.

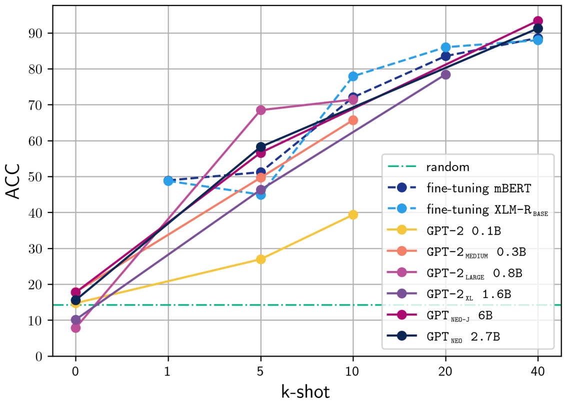

Figure 9: The acc results on English (en) SNIPS with GPT models.

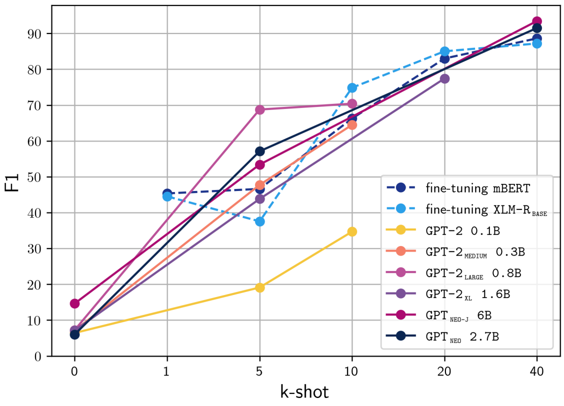

Figure 10: The f1 results on English (en) SNIPS with GPT models.

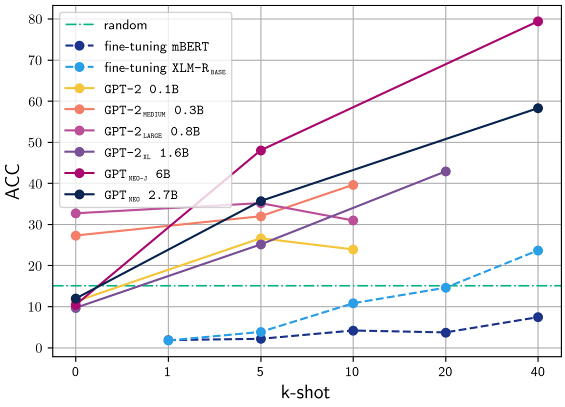

Figure 11: The acc results on the cross-lingual setting, English-German (de) MTOP dataset with GPT models.

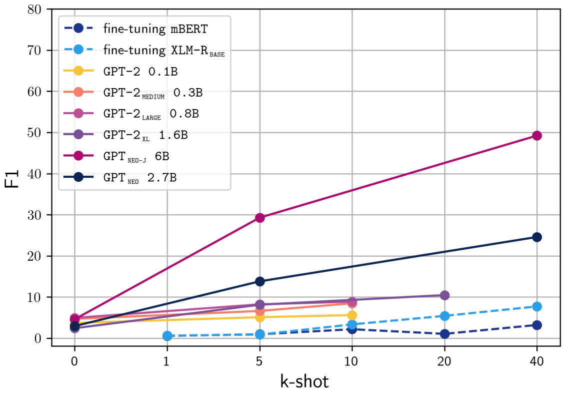

Figure 12: The f1 results on the cross-lingual setting, English-German (de) MTOP dataset with GPT models.

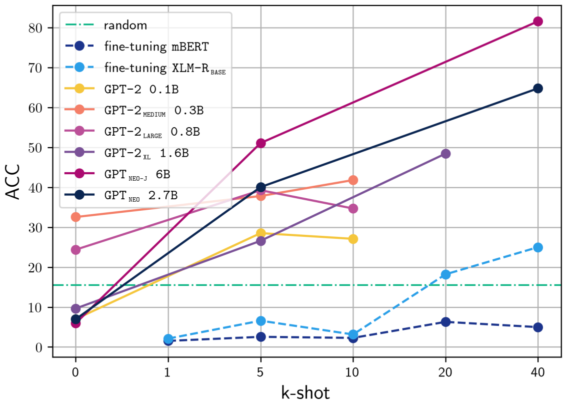

Figure 13: The acc results on the cross-lingual setting, English-Spanish (es) MTOP dataset with GPT models.

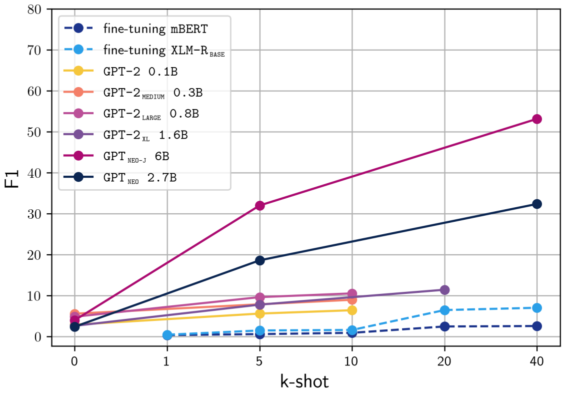

Figure 14: The f1 results on the cross-lingual setting, English-Spanish (es) MTOP dataset with GPT models.

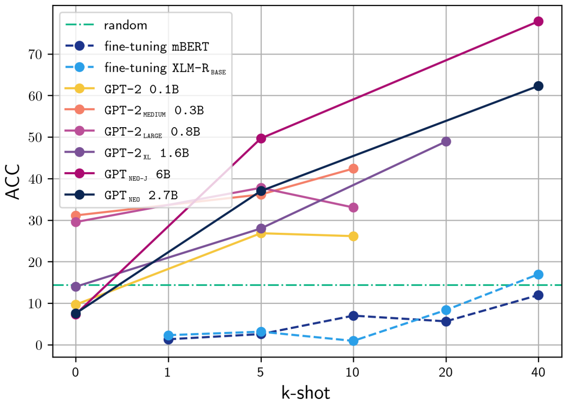

Figure 15: The acc results on the cross-lingual setting, English-French (fr) MTOP dataset with GPT models.

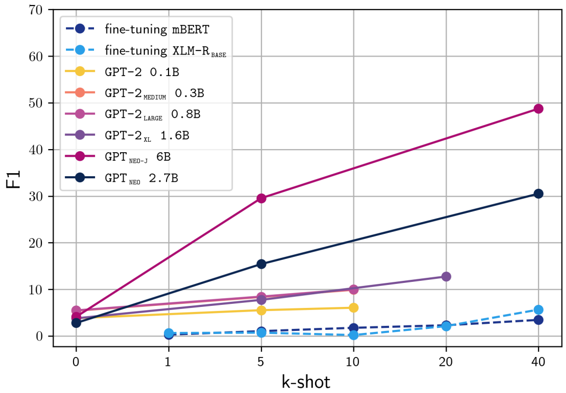

Figure 16: The f1 results on the cross-lingual setting, English-French (fr) MTOP dataset with GPT models.

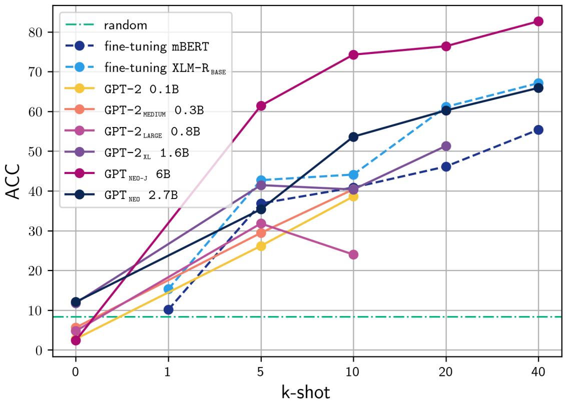

Figure 17: The acc results on the cross-lingual setting, English-Spanish (es) multilingual NLU dataset with GPT models.

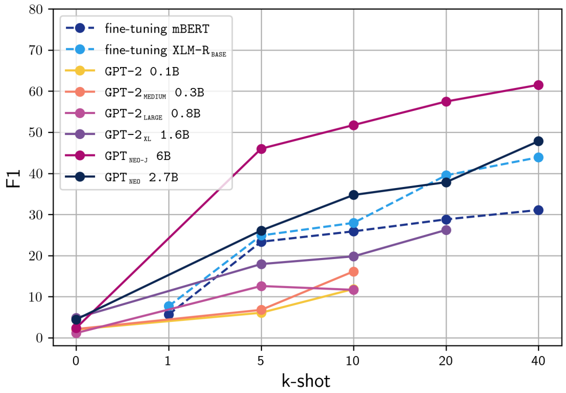

Figure 18: The f1 results on the cross-lingual setting, English-Spanish (es) multilingual NLU dataset with GPT models.