BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer

Anonymous ACL submission

Abstract

Despite remarkable advancements in few-shot generalization in natural language processing, most models are developed and evaluated primarily in English. To establish a rigorous and equitable evaluation framework for few-shot cross-lingual transfer, we introduce a new benchmark, called BUFFET, which unifies 15 diverse tasks across 54 languages in a sequence-to-sequence format and provides a fixed set of few-shot examples and instructions. Using BUFFET, we perform thorough evaluations of ten state-of-the-art multilingual large language models with different transfer methods, namely in-context learning and fine-tuning. Our findings reveal significant room for improvement in few-shot in-context cross-lingual transfer. Strong multilingual pre-trained or instruction-tuned models such as BLOOM or ChatGPT often lag behind much smaller mT5-base models given the same number of few-shot samples, particularly in low-resource languages. Our analysis suggests avenues for future research in few-shot cross-lingual transfer.

1 Introduction

Recent advances in NLP primarily focus on English (Blasi et al., 2022). As there is a shortage of adequate training data for most languages worldwide (Yu et al., 2022), zero-shot cross-lingual transfer (Hu et al., 2020b) is an active research area. This involves training models on high-resource languages like English, and then directly applying them to new languages without any training data in the target language. This approach often results in limited success when the target language is significantly different from the source language, motivating recent efforts to adapt models to a task in a new language using a limited number of training data in the target language. Such few-shot transfer often boosts performance, especially in languages that are dissimilar to the source language (Lauscher et al., 2020; Hedderich et al., 2020).

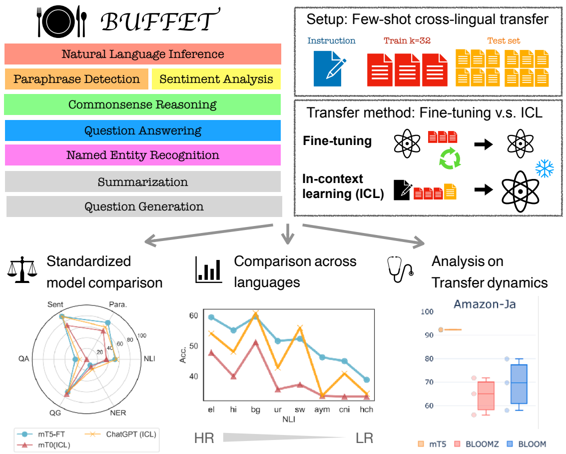

Figure 1: BUFFET includes unified diverse tasks in the same format, covering many typologically diverse languages to enable a fair comparison across different models, transfer methods, and learning setups.

Although there has been significant research on few-shot learning in English, employing techniques like in-context learning that do not necessitate parameter updates (Beltagy et al., 2022; Shin et al., 2020), few-shot cross-lingual transfer is still underexplored (Lin et al., 2021). While several recent work demonstrates the effectiveness of in-context learning in non-English languages on specific target tasks (Shi et al., 2023; Qin et al., 2023), it remains uncertain how well in-context learning performs in comparison to widely-employed fine-tuning-based transfer, particularly in a comparable setup involving diverse tasks and languages.

To comprehensively assess the capabilities of language models (LMs) for few-shot cross-lingual transfer, we introduce BUFFET: **B**enchmark of **U**nified **F**ormat **FE**w-shot **T**ransfer Evaluation (Figure 1) to enable rigorous evaluations and advance research on few-shot cross-lingual transfer. Similar to a rich buffet, BUFFET curates a diverse mix of tasks: 15 different tasks—including classification, structured prediction, and natural language generation—across 54 languages. BUFFET has several unique characteristics that are not present in prior multi-task multilingual benchmarks:

- providing a fixed set of few-shot demonstrations for training and validation for fair comparisons.
- combining diverse tasks into a unified text-to-text format with instructions.
- including datasets annotated on the target language and covering under-represented languages often missing in prior benchmarks.

On this new benchmark, we extensively evaluate the current state-of-the-art multilingual large language models (LLMs), including mT5 (Xue et al., 2021), mT0 (Muennighoff et al., 2023), BLOOM (Scao et al., 2022), BLOOMZ (Muennighoff et al., 2023), and ChatGPT (Ouyang et al., 2022), using both fine-tuning and in-context learning approaches. We also evaluate recent English-centric powerful open LMs such as Llama-2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023). In particular, BUFFET enables us to investigate the following research questions:

(RQ1) Is in-context learning competitive with fine-tuning in few-shot cross-lingual transfer? Notably, given the same small numbers of examples in the target languages, in-context learning on LLMs often under-performs much smaller specialized mT5-base models (Figure 1 bottom left).

(RQ2) How well do different transfer methods perform across tasks and languages? The performance gap between in-context learning and fine-tuning baselines is more significant in under-represented languages (Figure 1 bottom center). However, these LLMs perform well on generative tasks where a smaller task-specific LM struggles, demonstrating their superiority in generating fluent text for across languages. Meanwhile, although recent strong open LMs such as LLama2 or Mistral demonstrate strong performance in high-resource languages, possibly benefiting from a small amount of multilingual pre-training data (Touvron et al., 2023), they often show significant drops in performance on other languages less represented in English-centric pre-training corpora.

(RQ3) How does the choice of transfer setup affect different transfer strategies? BUFFET also enables us to perform an in-depth analysis of the effects of different demonstrations and instructions on the downstream transfer quality. We find that the choice of few-shot training examples has a substantial effect on model performance, especially for in-context learning, and often shows more significant effects than varying instructions. Optimal transfer settings may differ across models: instruction-tuned models often struggle to effectively utilize few-shot samples, possibly due to overfitting on their instruction-tuned training schemes. This highlights the need for a standardized benchmark like BUFFET to facilitate fair comparisons and further studies assessing these transfer dynamics in non-English data to improve few-shot cross-lingual transfer methodologies for many world languages.[1]

2 Background and Related Work

While few-shot cross-lingual transfer methods such as fine-tuning and in-context learning have been investigated (Section 2.1), limited research explores different methods *under comparable conditions*. We introduce BUFFET as a benchmark (Section 2.2) to facilitate fair comparisons between models and learning methods.

2.1 Methods for Cross-lingual Transfer

Fine-tuning for cross-lingual transfer. Prior work has shown that multilingual pre-trained models (Devlin et al., 2019; Xue et al., 2021; Conneau et al., 2020a), once trained on task data in resource-rich languages (e.g., English) have the ability to adapt to new languages with no training instances in a target language (Conneau et al., 2020b; Hu et al., 2020b; Wu and Dredze, 2019). However, such zero-shot transfer often struggles in languages that are distant from the source languages (Hedderich et al., 2020). Lauscher et al. (2020) shows that further fine-tuning models on few-shot samples in target languages give large performance improvements from zero-shot transfer approaches.

Cross-lingual in-context learning. In-context learning (Brown et al., 2020) aims to teach LMs new tasks by conditioning on a task description (instruction) and training examples (demonstrations). Despite active research on in-context learning (Schick and Schütze, 2021; Min et al., 2022b), most prior work focuses on English. Lin et al. (2021); Muennighoff et al. (2023) introduces pretrained LMs trained on more multilingual pretrained corpora or translated datasets and shows improved results. More recently, some concurrent work evaluates the effectiveness of proprietary LLMs e.g., ChatGPT on multilingual setup (Bang

[1] Our data and code are available online at XXX.

et al., 2023; Ahuja et al., 2023). However, how LLMs using in-context learning compete with the aforementioned fine-tuning approaches in a *comparable* setup and at scale has yet to be investigated.

2.2 Benchmarks for Cross-lingual Transfer

To enable a scalable and rigorous evaluation across multiple tasks, prior work has proposed multi-task benchmarks that unify existing datasets. XTREME (Hu et al., 2020b), XTREME-R (Ruder et al., 2021) and XGLUE (Liang et al., 2020) focus on zero-shot transfer of models fine-tuned on English datasets. Despite English-based few-shot evaluation benchmarks, such as CrossFit (Ye et al., 2021), in few-shot cross-lingual transfer, we lack a standardized evaluation benchmark to facilitate the comparison of models and learning methods. BUFFET provides the first large-scale few-shot cross-lingual transfer suits to address the gap. Importantly, to mitigate the effects of the high-performance variance in few-shot cross-lingual transfer (Zhao et al., 2021), we curate and aggregate results from multiple fixed k-shot training instances for each task and language. Concurrent with our work, MEGA (Ahuja et al., 2023) and XTREME-UP (Ruder et al., 2023) accelerate evaluations of cross-lingual transfer. BUFFET focuses on benchmarking few-shot transfer capabilities under *comparable* setup, with an emphasis on understanding the transfer dynamics.

3 Benchmark: BUFFET

We introduce a new standardized few-shot cross-lingual evaluation benchmark: BUFFET (**B**enchmark of **U**nified **F**ormat **Fe**w-shot **T**ransfer Evaluation). BUFFET unifies diverse NLP tasks and provides fixed sets of few-shot samples per task to facilitate fair comparisons (Table 1). **BUFFET-Full** covers 15 different tasks across 54 languages, while **BUFFET-Light** enables affordable and quick evaluations on limited subsets while retaining task and language diversities.

3.1 Design Principles

To establish a rigorous and equitable evaluation framework for few-shot cross-lingual transfer, we follow these design principles.

Standardized few-shot samples. BUFFET provides three different training and validation sets of k-shots (e.g., k=32) per task for a non-classification task, or per class for a classification task. This is to prevent significant performance discrepancies among various k-shot samples, which makes comparisons of different methods difficult.

Task diversity. BUFFET encompasses a broad range of task types, such as classification, generation, extraction, and structured prediction tasks, unlike existing cross-lingual benchmarks focusing on classification or retrieval (Hu et al., 2020b; Ruder et al., 2021; Liang et al., 2020). By converting all tasks into the same text-to-text format, we eliminate the need for task-specific model modifications.

Language diversity. BUFFET covers 54 typologically diverse languages, spanning 24 language families, including under-represented languages (e.g., indigenous languages of the Americas, African languages). The 36 out of 54 languages are not Indo-European languages. A full list of languages is available in Appendix Table 5.

Beyond evaluations on translated data. Prior few- or zero-shot evaluations were often conducted on datasets translated from English (e.g., XNLI; Conneau et al. 2018, XCOPA; Ponti et al. 2020). Those datasets might exhibit undesired biases, such as translation artifacts or unnatural topic distributions (Clark et al., 2020; Artetxe et al., 2020; Asai et al., 2021). BUFFET includes both translation-based datasets and datasets that are annotated directly in each language (Table 1, Data curation).

3.2 BUFFET Construction Process

Following Ye et al. (2021), we unify all datasets listed in Table 1 into the same text-to-text format, where a model is expected to directly generate the desired outputs given diverse inputs (Raffel et al., 2020). A task has *instructions*, k-shot training and validation examples, as well as test examples, each of which consists of input and output.

3.2.1 Unification Process

Instance selection. By default, we use all languages included in the original datasets.[2] For each language in each dataset, we use the original test or validation datasets as test instances (if the test set is not publicly available), and we randomly sample three sets of k-shot examples (*demonstrations*) for training and validation from the original training dataset, using the same random seeds.[3]

[2]For XLSUM and WikiANN, we sample languages target languages as discussed in Appendix Section A.

[3]We use 100, 13, and 21 as seed numbers.

Tasks	Dataset	Output	$\|L\|$	k	Metric	Domain	Data curation
Summarization	XLSUM	summary	12	1	ROUGE	News	aligned
Question Generation	TyDi QA-QG	question	8	8	BLEU	Wikipedia	in-language
NLI	XNLI	3-way class	14	16	acc.	misc.	translation
	Americas NLI	3-way class	10	16	acc.	misc.	translation
	Parsi NLU	3-way class	1	16	acc.	misc.	in-language
	OCNLI	3-way class	1	16	acc.	misc.	in-language
	KLUE-NLI	3-way class	1	16	acc.	misc.	in-language
Paraphrase Detection	PAWS-X	2-way class	6	7	acc.	Wikipedia	aligned
Sentiment	Indic-NLU-sent.	2-way class	14	16	acc.	e-commerce	translation
Analysis	Amazon Review	2-way class	5	16	acc.	e-commerce	in-language
Commonsense	XCOPA	multi-choice	11	16	acc.	misc.	translation
Reasoning	XWinograd	multi-choice	4	8	acc.	misc.	translation
QA	TyDiQA	span	8	8	F1	Wikipedia	in-language
Named Entity	WikiANN	names & tags	33	32	F1	Wikipedia	aligned
Recognition	MasakhaNER	names & tags	9	32	F1	News	in-language

Table 1: **The eight target tasks built upon 15 existing datasets in BUFFET.** $|L|$ indicates the number of languages, and k indicates the total number of training instances. We include datasets that are curated by translation, in-language annotation (in-language) and automatically aligned (aligned) following Yu et al. (2022).

Instruction selection. We use English instructions from SuperNaturalInstructions (Wang et al., 2022b) and PromptSource (Bach et al., 2022). Among multiple instructions, we sample the first instruction for a similar task that suits our scheme. The full list of the instructions is in Appendix Table 6.

Instruction translation. The availability of cross-lingual instruction is still largely limited, and prior work often translates instructions for target tasks (Lin et al., 2021; Shi et al., 2023). We provide translated instructions in 54 target languages, translated by NLLB (Costa-jussà et al., 2022), and manually translate the instructions into five languages.[4]

3.2.2 Tasks and Dataset Curation

Unlike in English, the availability of high-quality labeled datasets is largely limited, especially in generations or reasoning tasks, and the few available datasets are often translated from English. We select eight popular NLP tasks and identify available datasets for each task following the survey of multilingual datasets by Yu et al. (2022). Appendix Table 6 shows examples, and Appendix Section A.1 discusses the dataset choices.

Summarization. The task is to generate a summary given an article. We use the XLSum (Hasan et al., 2021) dataset of news article summarization.

Question generation. The task is to generate a question according to a given input passage and a corresponding answer (Xiao et al., 2021). We convert the TyDiQA (Clark et al., 2020) dataset into a question generation task, which we refer to TyDiQA-QG.

Natural language inference (NLI). NLI involves determining the logical relationship (`entailment`, `contradiction`, `neutral`) between two text fragments, i.e., a premise and a hypothesis. We include five datasets covering typologically-diverse languages

Paraphrase detection. The task is to identify whether two sentences have/do not have the same meaning (`duplicate` or `not duplicated`). We adopt PAWS-X (Yang et al., 2019).

Sentiment analysis. Binary sentiment analysis identifies whether a text (e.g., a product review from Amazon) expresses `positive` or `negative` sentiment towards a topic. We use the Multilingual Amazon Review dataset (Keung et al., 2020) and IndicNLU-Sentiment (Aggarwal et al., 2022), and convert the former to a binary classification task (see Appendix Section A.1).

Commonsense reasoning. For a sentence and two options, the task is to select one of the option labels, `(A)` or `(B)`, based on which is better suited to the given context. We use two commonsense reasoning datasets, XCOPA (Ponti et al., 2020) and XWinograd (Muennighoff et al., 2023).

Question answering (QA). The task is to answer a question given a paragraph, where the answer is a sub-span of the paragraph. We use TyDiQA-GoldP (Clark et al., 2020), which we refer to as TyDiQA for simplicity.

Named entity recognition. The task is representative of sequence labeling to detect and

[4]Manual translations are performed by volunteers.

classify named entities in an input sentence. We adopt WIKIANN (Pan et al., 2017) and MASAKHANER (Adelani et al., 2021). We convert the task into a text-to-text format, where a model extracts all named entities with named entity tags:[5] <location>, <person>, and <organization>.[6]

3.3 BUFFET Evaluation

Evaluation metrics. In Table 1, we list metrics for each task. To mitigate the variance from different few-shot samples, for each language included in a given task, we average the model's performance over three different sets of k-shot instances. Subsequently, each dataset score is calculated as a macro-average of the per-language score (Clark et al., 2020). Finally, following Liang et al. (2020), we take two separate average scores: (a) **Avg. class** score of all classification and QA tasks, and (b) **Avg. generation** score of all generation tasks.

BUFFET-Light. Conducting a comprehensive evaluation covering a wide range of languages and tasks in BUFFET is valuable but computationally expensive, especially when we use external APIs or large model sizes (e.g., more than ten billion). BUFFET-Light is a representative subset of languages and tasks for resource-limited scenarios. We select languages and datasets to ensure that we cover diverse languages and output formats, discussed in detail in Section A.3.

4 Benchmarking LMs on BUFFET

4.1 Transfer Methods

We investigate various transfer methods with and without parameter updates, summarized in Table 2. To assess the benefit of k-shot training examples in the target language, we also conduct experiments on zero-shot transfer methods. We assume that the model can optionally use instructions in the target language or another language, or full training sets in a high-resource language like English.

Fine-tuning (methods with parameter updates). We explore several transfer approaches that require parameter updates: **Target fine-tuning (TARGET FT)** that trains models on few-shot samples for each language, **English fine-tuning (ENGLISH FT)** that trains models on a source language (i.e., English) only and uses no target language data, and **English+Target fine-tuning (ENG.+TGT. FT)** further fine-tunes the ENGLISH FT models on few-shot samples of target languages.

Transfer	Training Demos EN	Training Demos Target	Instructions EN	Instructions Target
TARGET FT		k		
ENGLISH FT	N			
ENG.+TGT. FT	N	k		
ENGLISH ICL		k	✓	
TARGET ICL		k		✓
Z-EICL			✓	

Transfer	Pretraining	LMs
FINE-TUNING	Unlabeled	mT5-base
ICL	Unlabeled	BLOOM, mT5-xxl
ICL	+ Instruction Tuning	BLOOMZ-7B, mT0-xxl ChatGPT

Table 2: **Comparison of transfer methods, based on the resources they use, and LMs.** The top section requires parameter updates via fine-tuning (FT), and the bottom uses ICL with no updates. k = k-shot examples; N = full training data; ✓= instruction language. The bottom half lists the models evaluated in this work. The blue-colored rows are instruction-tuned models.

In-context learning (methods without updates). We explore several in-context learning methods. **English in-context learning (ENGLISH ICL)** uses English instructions and demonstrations in the target languages, while **Target In-context learning (TARGET ICL)** uses both instructions and demonstrations in the target language. Finally, **Zero-shot English In-context learning (Z-EICL)** uses only English instructions without demonstrations (neither in English nor in the target language), as in zero-shot transfer. Unlike in English, where abundant instructions and instance annotations are available, for many languages we lack annotated instructions (Wang et al., 2022b). We use machine-translated instructions in BUFFET.

4.2 Language Models

We evaluate six diverse LM (Table 2 bottom), including pretrained vanilla LMs as well as instruction-tuned LMs, which have been trained on a massive number of tasks with instructions.

Models for fine-tuning. Due to the high costs of fine-tuning for every k-shot setting, we experiment

[5]This is more challenging than the standard sequence labeling setup since the model must reproduce the entity spans and generate appropriate tags. For example, the output for "Obama served as the 44th president of the United States." would be "Obama <person> United States <location>."

[6]Although MASAKHANER supports other named entity tags and distinguishes the beginning and middle/end of the named entities, we discard named entity categories beyond the three types and merge the beginning and middle/end entity tags to make the task formulation consistent with WIKIANN.

	Output	Classification			Multi-choice		Span	Str.	Generation		Avg.	
	Tasks	NLI	Sent.	PWX	XCP	XWG	TyDi	NER	QG	Summ.	class	gen
TGT. FT	mT5	35.0	67.2	47.7	44.1	48.8	5.2	33.4	3.2	2.5	40.7	2.9
ENG. FT	mT5	49.9	89.8	77.5	49.6	50.0	66.8	39.0	3.8	6.2	60.7	5.0
ENG.+TGT.	mT5	**51.8**	**91.0**	**77.8**	49.5	48.5	**69.5**	**47.8**	12.5	**11.8**	**61.2**	**12.2**
ENG. ICL	BLOOM	32.1	81.7	42.2	50.2	51.0	54.7	24.2	9.3	3.4	45.0	6.4
	mT5	35.7	50.0	42.2	50.4	47.5	0.2	0.0	0.0	0.4	31.7	0.2
	BLOOMZ	31.5	86.3*	48.5*	50.4	54.2	65.8*	25.5	13.5	8.3*	47.5	10.9
	mT0	32.6	80.4*	60.5*	52.9	57.8	74.5*	6.9	15.3	2.7*	52.2	9.7
	ChatGPT	**54.5**	91.1	68.6	**76.7**	73.3	68.1	45.4	**21.2**	5.4	**64.6**	13.3
TGT. ICL	BLOOM	27.9	80.5	46.5	49.9	51.8	11.8	23.4	11.2	3.6	40.4	7.4
	mT5	35.7	50.0	42.2	49.8	45.2	0.2	0.0	0.0	0.4	31.5	0.2
	BLOOMZ	32.0	61.7*	52.5*	49.7	55.5	63.1*	23.4	9.1	8.0*	43.4	8.5
	mT0	36.2	72.1*	60.6*	50.5	60.3	73.6*	7.9	**16.1**	3.4*	46.3	9.7
	ChatGPT	48.2	**91.5**	68.2	74.3	**73.4**	68.0	44.8	21.1	11.4	62.7	**16.3**
Z-EICL	BLOOM	33.3	37.2	42.3	50.0	47.1	4.3	0.0	14.0	6.3	29.2	10.1
	mT5	35.1	49.8	42.2	50.7	55.5	2.2	0.0	0.1	4.8	32.5	0.6
	BLOOMZ	33.5	51.6*	57.8*	51.8	51.0	83.2*	11.2	9.5	4.3*	41.9	6.9
	mT0	48.5	90.0*	90.6*	**63.8**	**61.0**	80.1*	0.0	10.2	12.0*	56.4	11.1

Table 3: **Overall experiment results in BUFFET**. Note that to enable comparison between ChatGPT (only tested on BUFFET-Light) and other methods, we present BUFFET-Light results, and the overall results on BUFFET are available in Table 10. The blue-colored rows are instruction-tuned models. We added * symbols next to the scores for the tasks on which the models have been trained. **Bold** fonts indicate the best results for each task, among the models that are not directly trained on the task. When ChatGPT achieves the best results, we note the second-best number from the models not trained on the task, as ChatGPT may have been trained on a similar task.

with an efficient yet competitive mT5-base with 580 million parameters (Xue et al., 2021).

Models for in-context learning. We experiment with BLOOM-7B (7 billion parameters; Scao et al., 2022) and mT5-xxl (13 billion parameters; Xue et al., 2021). We also experiment with their instruction-tuned variants: BLOOMZ-7B and mT0-xxl (Muennighoff et al., 2023), as well as the current state-of-the-art ChatGPT (`gpt-3.5-turbo-0301`; Ouyang et al. 2022). Note that these models may be trained on some datasets included in BUFFET. Due to the high API costs, we conduct ChatGPT evaluations on BUFFET-Light data only with the two few-shot transfer methods. While our main experiments focus on multilingual pre-trained models, in Section 5.2 we further evaluate four English-centric LMs on BUFFET-Light.

4.3 Experiment Details

Fine-tuning. For ENG.+TGT. FT and ENGLISH FT, we train on representative English datasets following Hu et al. (2020b) for three epochs and five for smaller COPA and Winograd datasets. The source English datasets are listed in the appendix. We fine-tune on k-shot samples for 300 epochs (TARGET FT) and 200 epochs (ENG.+TGT. FT).

In-context learning. We prompt LLMs with instructions and k-shot demonstrations available in BUFFET. Our preliminary experiments reveal mT0 performs significantly better when zero or very few few-shot samples are used, so we use 4-shots for mT0 ENGLISH ICL and TARGET ICL by default, while for other models we use all demonstrations unless they exceed max context length. We use greedy decoding for predictions. For tasks with a fixed set of pre-specified answer candidates, we compute the probability of option tokens by iterating options except for ChatGPT without access to token probabilities. Due to the high inference costs, we evaluate ChatGPT only on BUFFET-Light.

5 Results and Analysis

5.1 Main Results

Table 3 shows aggregated results of fine-tuned and in-context learning-based LMs on BUFFET-Light for fair comparisons between ChatGPT and other models. Full experiment results including BUFFET-Full results on each task are in the Appendix. Below, we summarize the key findings.

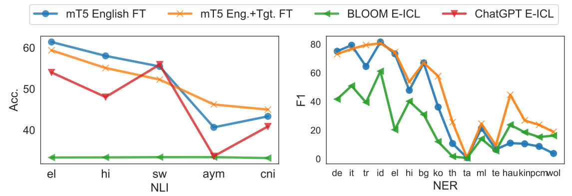

Figure 2: **Model performance on NLI and NER, displayed for various languages.** The languages are sorted based on token availability in mC4, with the left side representing high-resource ones. ChatGPT results are not shown on the NER chart as some languages are not included in BUFFET-Light.

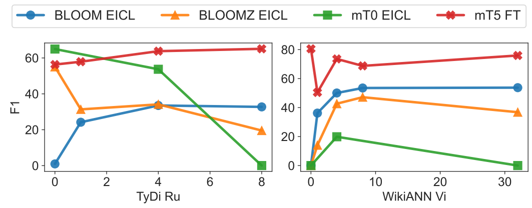

Figure 3: **Model performance across different numbers of k-shots**. mT5 FT denotes mT5 ENG.+TGT. FT. More results are in Appendix.

LLMs with in-context learning often lag behind much smaller fine-tuned models. Our comparison shows that few-shot cross-lingual transfer via in-context learning remains challenging; ENGLISH ICL using BLOOM, BLOOMZ (7B) and mT0 (13B) often under-performs mt5-base (580M) fine-tuned on English datasets (ENGLISH FT or ENG.+TGT. FT). Even the current state-of-the-art ChatGPT underperforms mT5-base ENG.+TGT. FT in simple discriminative tasks (e.g., PAWS-X) or structured prediction tasks (NER). However, ICL baselines constantly outperform mT5 (TARGET FT) across tasks and ENG.+TGT. FT on XCOPA and XWINOGRAD with limited scarce English task data. This implies that when lacking task-specific training data even in English, prompting LLMs can be more effective, while otherwise training a specialized model on English data and then fine-tuning few-shot instances is still effective in discriminative tasks.

Zero- and few-shot transfer remains challenging in under-represented languages. Figure 2 illustrates model performance on NER (WIKIANN and MASAKHANER) and NLI (XNLI, AMERICASNLI) across different languages.[7] The languages are sorted based on the token availability in the mC4 corpus,[8] with high-resource languages positioned on the left side. In general, models such as mT5 ENGLISH FT or ChatGPT ENGLISH ICL exhibit strong performance in high-resource languages, but their effectiveness diminishes in underrepresented languages (right side, Figure 2). For instance, on NLI in Aymara (`aym`), ChatGPT achieves slightly higher performance than a random baseline. We also find that fine-tuning with k in-language examples is very effective for less-represented languages: mT5 ENG.+TGT. FT significantly outperforms mT5 ENGLISH FT in lower-resource languages (e.g., 30% improvements in Hausa on MasakhaNER).

Instruction-tuning helps zero-shot ICL but may not generalize well to few-shot settings. The zero-shot performance of instruction-tuned models is significantly higher than the zero-shot performance of non-instruction-tuned models (Table 3: mT0-xxl and BLOOMZ-7B Z-EICL v.s. mT5-xxl and BLOOM-7B Z-EICL). However, instruction-tuned models show surprising performance deterioration in few-shot settings: across tasks, mT0 performs worse in few-shot settings than in zero-shot settings (ENGLISH ICL v.s. Z EICL). we hypothesize that since these models are optimized to execute a new task solely based on an instruction, with no prior demonstrations (Muennighoff et al., 2023), they struggle to learn in context from few-shot demonstrations. We conduct controlled experiments in Section 5.2 for further analysis.

5.2 Analysis

Effect of varying number of k. Figure 3 demonstrates the impact of increasing the number of few-shot samples for in-context learning and fine-tuning, on two tasks: TYDIQA, and WIKIANN. We vary the number of few-shot demonstrations, including 0, 1, 4, and 8 (for the tasks with more than 8 shots). Full results on more tasks and languages are in Appendix D.3. Increasing the number of few-shot examples has a notable positive impact on fine-tuning (mT5 FT). Similarly, non-instruction-tuned BLOOM benefits from the inclusion of few-shot samples on most of the tasks. However, for instruction-tuned models, namely BLOOMZ and mT0, which were exclusively trained with instructions rather than demonstrations, we observe a sig-

[7] Several languages in MASAKHANER or AMERICAS NLI are not part of the pretraining process.

[8] We use token count statistics at https://github.com/allenai/allennlp/discussions/5265. Languages not seen in pretraining are sorted alphabetically.

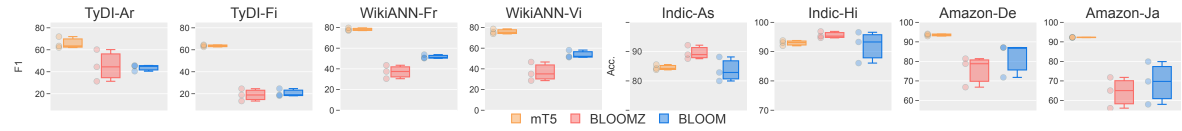

Figure 4: **Model performance across different k-shot demonstrations for TYDIQA, WIKIANN, INDICSENTIMENT and AMAZONREVIEW**. Each circle indicates performance given different k-shot demonstrations.

nificant decline in performance when additional demonstrations are added, possibly due to the overfit to the zero-shot ICL scenario, even on previously unseen tasks such as WIKIANN. Prior work on English instruction-tuning has demonstrated that training an LM on diverse setups (few-shot, zero-shot, using both demonstrations and instructions) is effective in alleviating such sensitivity of instruction-tuned models to diverse evaluation setups (Longpre et al., 2023). It is important to develop multilingual instruction-following models capable of effectively utilizing both instructions and demonstrations.

Effect of different k shots. Figure 4 shows model performance across the three different sets of k examples. We observe the significant variance in fine-tuning-based transfer across different demonstrations, confirming Zhao et al. (2021). Importantly, we show that in-context learning is *even more sensitive* to demonstration choice than few-shot fine-tuning, further emphasizing the importance of standardized k-shots for a fair transfer evaluation. For instance, the standard deviation on AMAZON REVIEW for BLOOM ENGLISH ICL and mT5 ENG.+TGT. FT is 2.2 and 0.2, respectively. We also found that in 49.7% of the cases, the optimal k-shot demonstrations for BLOOM and BLOOMZ ENGLISH ICL differ.

Effect of model scaling. Appendix Figure 12 shows the performance of BLOOM-560 million, 1 billion, and 7 billion with few-shot ENGLISH ICL on a subset of the tasks. Overall performance significantly improves across different model sizes, indicating cross-lingual transfer performance via ICL improves with scale; this is consistent with findings in Lin et al. (2021) on classification tasks.

Effect of prompt templates. We investigate the effectiveness of different English instructions on TYDIQA-QG in four-shot settings using mT0 and BLOOM as base models in Appendix Table 24. We compare four relevant instructions and one irrelevant instruction (an instruction for AMAZON REVIEW) and find that the performance sharply decreases with irrelevant instructions on the instruction-tuned model (7.1 → 0.4 BLEU). However, among relevant instructions, the performance gap on BLOOM is limited compared to the large variance observed across different demonstration sets. The larger performance gap for instruction-tuned mT0 likely indicates that instruction-tuned models are more sensitive to diverse prompts.

Evaluating English-centric LMs. BUFFET-Light enables easy and quick evaluations of LMs. We conduct BUFFET-Light evaluations on four recently released LMs (7B) primarily trained in English: LLama1 (Touvron et al., 2023), Llama2, Llama2-chat (Touvron et al., 2023) and Mistral (Jiang et al., 2023). Full results are in Table 26: on average, LLama1, LLama2, LLama2-chat, and Mistral get 28.1, 41.6, 44.1, and 45.2 on classification tasks, and 4.3, 6.4, 6.4, and 7.4 on generation tasks, respectively. Except for LLama1 which explicitly filters out text in non-alphabetic languages, other English-centric LMs match or exceed multilingual BLOOM and BLOOMZ. This result suggests even small amounts of multilingual data in pre-training help LLMs acquire multilingual abilities, corroborating Blevins and Zettlemoyer (2022a). Yet, they often struggle with many other languages (e.g., AMERICASNLI or INDIC SENTIMENT), and it remains unclear how much target language data is necessary for this to occur.

6 Conclusion and Discussion

We introduce BUFFET, a few-shot cross-lingual transfer benchmark that encompasses a diverse range of discriminative and generative tasks across many typologically diverse languages. While LLMs utilizing in-context learning excel in generation tasks, they are often surpassed by smaller fine-tuned models specifically trained for target tasks. Our analysis sheds light on several important open questions for better multilingual instruction-tuning, and more balanced multilingual pre-training, and suggests the necessity of evaluating across languages and tasks under comparable settings.[9]

[9] We provide detailed discussions in Appendix Section E.

Limitations

Selection of tasks. As the first step toward standardized evaluation for few-shot cross-lingual transfer, BUFFET focuses on popular discriminative tasks and some generative tasks, with well-studied evaluation protocols and rich annotated resources. Due to the lack of high-quality non-English annotated data, BUFFET does not include many datasets that require complex reasoning tasks. Further exploration can expand these evaluations to more diverse and complex tasks, such as MTOP (Li et al., 2021) or MGMS8K (Shi et al., 2023), or knowledge-intensive tasks (Asai et al., 2021; Ogundepo et al., 2023). Yet, it should be noted that high-quality generation or reasoning task data are often only available handful of resource-rich languages, which makes BUFFET-style comprehensive comparisons across languages difficult. We encourage the community to work towards diverse high-quality evaluation datasets in more world languages.

Hyper-parameter search or prompting. Since our main focus is to benchmark different LMs and learning methods in a comparable format, we do not explore sophisticated prompting methods or conduct task- or language-dependent hyper-parameter searches. We anticipate that BUFFET will encourage the LLM community to explore new methods to further improve in-context learning beyond English.

Translated instructions. We use instructions translated by the NLLB (Costa-jussà et al., 2022) for TARGET ICL; such machine-translated instructions are prone to errors, especially in less-represented languages, that can affect the final performance.

Lack of underrepresented variants, dialects Typologically distinct and low-resource languages are often excluded from the cross-lingual benchmarks used to assess cross-lingual transfer capabilities in LLMs. Our evaluation with BUFFET demonstrates that even the most powerful LLMs still perform poorly on less-represented languages, by evaluating them on more than 50 languages. However, we do not specifically focus on finer-grained language varieties and dialects that are commonly spoken by underrepresented populations. We advocate for conducting more studies that include under-represented languages and their dialects, as emphasized in previous works (Aji et al., 2022; Kakwani et al., 2020), particularly when evaluating massively multilingual models.

Ethics Statement

While there has been significant research on in-context learning with LLMs, most of the focus has been on the English language. This raises questions about the applicability of findings from English few-shot NLP to few-shot cross-lingual transfer scenarios. To address this gap, BUFFET aims to provide a comprehensive and less biased evaluation framework. However, it is important to note that our benchmark dataset currently covers only 54 out of the approximately 6,000 world languages. In light of these limitations, we encourage future research to explore the effectiveness and limitations of widely used transfer methods in a more diverse range of languages. This will help us gain a deeper understanding of the generalizability of transfer learning techniques across different linguistic contexts. We curate existing open-licensed datasets as source datasets of BUFFET, and manually assessed sampled questions to see the quality of data as well as potential privacy risks.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. In *Transactions of the Association for Computational Linguistics*.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of Empirical Methods in Natural Language Processing*.

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2022. Qameleon: Multilingual qa with only 5 examples. *arXiv preprint*.

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

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the Association for Computational Linguistics*.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the Association for Computational Linguistics*.

Venkat Arun and Hari Balakrishnan. 2018. Copa: Practical Delay-Based congestion control for the internet. In *USENIX Symposium on Networked Systems Design and Implementation*.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint*.

Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, and Sameer Singh. 2022. Zero- and few-shot NLP with pretrained language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the Association for Computational Linguistics*.

Terra Blevins and Luke Zettlemoyer. 2022a. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574.

Terra Blevins and Luke Zettlemoyer. 2022b. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *ACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

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

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of Empirical Methods in Natural Language Processing*.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the Association for Computational Linguistics*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint*.

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 Conference of the North American Chapter of the Association for Computational Linguistics*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the Association for Computational Linguistics*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics*.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of Empirical Methods in Natural Language Processing*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020a. OCNLI: Original Chinese Natural Language Inference. In *Proceedings of Findings of Empirical Methods in Natural Language Processing*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020b. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the International Conference of Machine Learning*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Proceedings of Findings of Empirical Methods in Natural Language Processing*.

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

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the 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 Empirical Methods in Natural Language Processing*.

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 European Chapter of the Association for Computational Linguistics*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *arXiv preprint*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the Association for Computational Linguistics*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. MetaICL: Learning to learn in context. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of Empirical Methods in Natural Language Processing*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the Association for Computational Linguistics*.

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *arXiv preprint arXiv:2305.06897*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of Advances in Neural Information Processing Systems*.

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 Association for Computational Linguistics*.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of Empirical Methods in Natural Language Processing*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint*.

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. In *Journal of Machine Learning Research*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, et al. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint*.

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 Conference of the North American Chapter of the Association for Computational Linguistics*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *Proceedings of the International Conference on Learning Representations*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of Empirical Methods in Natural Language Processing*.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *International Joint Conference on Artificial Intelligence*.

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 Conference of the North American Chapter of the 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 Empirical Methods in Natural Language Processing*.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of Empirical Methods in Natural Language Processing*.

Xinyan Velocity Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Proceedings of Findings of Empirical Methods in Natural Language Processing*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint*.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the Association for Computational Linguistics*.

Appendix

A Benchmark Details

BUFFET unifies diverse tasks and languages to enable a comparable and equitable evaluation for few-shot cross-lingual transfer. We provide a comparison with other multi-task benchmarks in Table 4. In this section, we present technical dataset details.

A.1 Task-specific Details

Natural language inference. In addition to the widely used XNLI (Conneau et al., 2018), we gather NLI datasets that are annotated in each language or designed to cover under-represented languages: AMERICASNLI (Ebrahimi et al., 2022), PARSINLU-ENTAILMENT (Khashabi et al., 2021), KLUE-NLI (Park et al., 2021), and OCNLI (Hu et al., 2020a). We use the same target labels, `entailment`, `contradiction`, `neutral` across different datasets. We use 16 examples for each class.

Paraphrase detection. We adopt PAWS-X (Yang et al., 2019) and include 16 shots for each class as few-shot training and validation data.

Sentiment analysis. We use the MULTILINGUAL AMAZON REVIEW DATASET (Keung et al., 2020) and INDICNLU-SENTIMENT (Aggarwal et al., 2022). INDICNLU-SENTIMENT is created by translating English sentiment analysis data into diverse Indic languages. For the former, we discard the neutral class (the reviews with a score of 3) and assign reviews with scores of 4 and 5 to the positive class and reviews with scores of 1 and 2 to the negative class. For both datasets, we sample 16 demonstrations per class.

Commonsense reasoning. We use two commonsense reasoning datasets, XCOPA (Ponti et al., 2020) and XWINOGRAD (Muennighoff et al., 2023). Due to the smaller scale of the datasets, we sample 16 and 8 training instances in total for XCOPA and XWINOGRAD, respectively.

Question answering. We use TYDIQA-GOLDP (Clark et al., 2020) for QA, as the data is annotated in each language, better reflecting native speakers' interests and linguistic phenomenon. Due to the longer average input length, we limit the number of exemplars to 8.

Named entity recognition. We adopt WIKIANN (Pan et al., 2017) and

	Multi-ling.	Few-S	Gen.	Low-R
XTREME	✓			
XTREME-R	✓			
XGLUE	✓		✓	
CrossFit		✓	✓	
MEGA*	✓	✓		
XTREME-UP*	✓			✓
BUFFET	✓	✓	✓	✓

Table 4: Comparison of the existing benchmarks based on their multilinguality (Multi-ling.), few-shot task formulation (Few-S), availability of generative tasks (Gen.), and coverage of low-resource languages (Low-R). * indicates concurrent work.

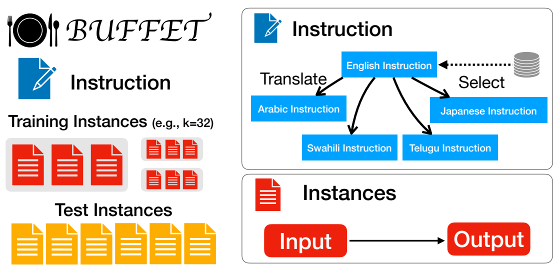

Figure 5: BUFFET includes 15 datasets, which are unified into the same single text-to-text format.

MASAKHANER (Adelani et al., 2021). WIKIANN is automatically curated and exhibit alignment errors (Yu et al., 2022). We sample languages on WIKIANN as discussed in Section A.2. We use 32 instances overall for few-shot transfer.

Summarization. We use the XLSUM (Hasan et al., 2021) dataset to benchmark models' ability to generate a summary given a news article. Due to the context window limit, we use only 1 shot for training in this task.

Question generation. We convert the TYDIQA-GOLDP dataset into a question generation task, which we refer to TYDIQA-QG. Given the gold paragraph and an answer, the system generates the original question. We use 8 examples for few-shot training.

A.2 More Details of BUFFET

This section will provide further details of the BUFFET benchmark. Figure 5 summarizes the construction process of BUFFET.

Instance and language sampling for XLSUM, WIKIANN and AMAZON REVIEW. For automatically aligned datasets with many test languages, such as XLSUM or WIKIANN, we filter out languages that are not included in any other BUFFET datasets following suggestions by Yu et al. (2022).[10] With large-scale automatically aligned datasets, we randomly sample 1,000 test instances in XLSUM and WIKIANN and 2,000 test instances for Amazon Review, to reduce inference time costs across many languages and multiple sets of demonstrations.

Instructions. The full list of the instructions written in English is available in Table 7. For some tasks, we modify the original instruction to make labels consistent with the names used in BUFFET or to remove task-specific dependencies in the input data field. For example, an instruction for PAWS-X says the class names are "repeated/not repeated" while in BUFFET we use "duplicated/not_duplicated" as labels, so we change the labels in the original instruction.

List of the languages. We show the list of the 54 languages included in BUFFET in Table 5. BUFFET covers 25 different language families, and also exhibits geographical diversities. Table 8 shows the full list of the datasets with language names included in BUFFET.

Examples. Table 6 shows the input and output examples in BUFFET. We reformulate all o the tasks with diverse formats into the same text-to-text format.

A.3 BUFFET-Light

Task and language decisions. The goal of building the BUFFET-Light subset is to enable quick multilingual evaluation without losing the language and task diversity in the original BUFFET. To this end, we filter BUFFET so that we evaluate between 3 and 7 languages per task, and each language is included in at most three tasks.[11] This design choice allows us to consider 31 diverse languages across all tasks in BUFFET while reducing the number of evaluation settings by 66%.

Final list of BUFFET-light. The full list of tasks and languages in BUFFET are in Table 9.

B More Experimental Details

Fine-tuning. We use the following English datasets for ENGLISH FT and ENG.+TGT. FT: SQUAD (Rajpurkar et al., 2016) for QA, MNLI (Williams et al., 2017) for NLI, PAWS (Zhang et al., 2019) for paraphrase detection, XLSUM (Hasan et al., 2021) for summarization, COPA (Arun and Balakrishnan, 2018) for XCOPA, WINOGRAD for XWINOGRAD, the AMAZON MULTILINGUAL REVIEW English set

Language name	Language family	code
Amharic	Afro-Asiatic	amh
Arabic	Afro-Asiatic	ar
Assamese	Indo-European	as
Aymara	aymaran languages	aym
Belarusian	Indo-European	be
Bengali	Indo-European	bn
Boro	Sino-Tibetan	brx
Bulgarian	Indo-European	bg
Bribri	Chibchan	bzd
Chinese	Sino-Tibetan	zh
Asháninka	Arawakan	cni
Estonian	Uralic	et
Finnish	Uralic	fi
French	Indo-European	fr
German	Indo-European	de
Guarani	Tupian	gn
Gujarati	Indo-European	gu
Haitian	French Creole	ht
Hausa	Niger–Congo	hau
Wixarika	Uto-Aztecan	hch
Hindi	Indo-European	hi
Igbo	Niger–Congo	ibo
Indonesian	Austronesian	id
Italian	Indo-European	it
Japanese	Japonic	ja
Kannada	Dravidian	kn
Kinyarwanda	Niger–Congo	kin
Korean	Koreanic	ko
Luo	Nilo Saharan	luo
Maithili	Indo-European	mai
Malayalam	Dravidian	ml
Marathi	Indo-European	mr
Modern Greek	Indo-European	el
Nahuatl	Uto-Aztecan	nah
Oriya (macrolanguage)	Indo-European	or
Otomí	Oto-Manguean	oto
Panjabi	Indo-European	pa
NigerianPidgin	English Creole	pcm
Persian	Indo-European	fa
Portuguese	Indo-European	pt
Quechua	others	qu
Russian	Indo-European	ru
Shipibo-Konibo	Panoan	shp
Spanish	Indo-European	es
Swahil	Niger–Congo	sw
Tamil	Dravidian	ta
Rarámuri	Uto-Aztecan	tar
Telugu	Dravidian	te
Thai	Kra–Dai	th
Turkish	Turkic	tr
Urdu	Indo-European	ur
Vietnamese	Austroasiatic	vi
Wolof	Niger–Congo"	wol
Yorùbá	Niger–Congo	yor

Table 5: List of all languages in BUFFET.

[10] On XLSUM, we further reduce the number of languages to reduce the inference costs while maintaining language diversities.

[11] In addition to the high-resource languages per task, we also include low-resource languages when available (i.e., for NLI) to not unfairly inflate BUFFET-Light scores.

Task	Dataset	Input	Output
NLI	AMERICAS NLI	premise: Ramonar mayamp jawsañaxanawakunalaykutix mä jiskt'aw utjitana . . . walikiwa. . . tukt'ayayita.. mä jisk't'aw utjitana kuntix lurkan ukata. [SEP] hypothesis: Janiw jayraskayat Ramonar jawsañxa. (aym)	contradiction
PARAPHRASE	PAWS-X	sentence 1: Ses parents sont Angelina Miers, une artiste de premier plan, et Don Luis Toranzos, d'Argentine. [SEP] sentence 2: Ses parents sont Angelina Miers, elle-même un artiste de premier plan, et Don Luis Toranzos d'Argentine. (fr)	duplicate
SENTIMENT	AMAZON	review title: 质量很好，空间容量大，可以装很多东西 review body: 箱子很轻盈，柔韧性不错，不易变形。外观优雅美观，出行很有范，呵呵。好评！	positive
COMMONSENSE	XCOPA	Õpetaja andis õpilastele kodutöö. (A) Õpilased saatsid kirju. (B) Õpilased ägisesid. (et)	(B)
COMMONSENSE	XWINOGRAD	フリースは綿より感触がよい。_のほうがずっと柔らかいからいだ。 (A) フリース (B) 綿	(A)
QA	TYDIQA	question: Mikä oli Stanley Kubrickin ensimmäinen elokuva? context: Lyhytelokuvien jälkeen Kubrick teki ensimmäisen pitkän elokuvansa Fear and Desire vuonna 1953 rahoittaen sen kokonaan itse ja sukulaistensa avustuksella, mikä oli tuolloin hyvin epätavallista. Kubrickin esikoiselokuva oli kuitenkin floppi, ja ohjaaja osti kaikki esityskopiot itselleen, jotta elokuvaa ei näytettäisi. Elokuva merkitsi myös hänen ensimmäisen avioliittonsa loppua, koska Kubrick tapasi kuvauksien aikana Ruth Sabotkan, jonka kanssa hän muutti yhteen avioeronsa jälkeen. Kubrick ja Sabotka menivät naimisiin vuonna 1955, ja he saivat yhdessä yhden lapsen, Katharinan (syntynyt 1953). (fi)	Fear and Desire
NER	MASAKHANER	Issachar alikuwa ametokea India akielekea Israel ambapo aliwekwa chini ya ulinzi na hakutakiwa kutoka nje ya uwanja wa ndege wa Russia .	India <organization> Israel <organization> Russia <organization>
QG	TYDIQA-QG	premise: 롯데는 이번 상반기 채용과 관련해 구직자들에게 실질적인 도움이 될 수 있도록 다양한 방법으로 정보제공 활동을 강화할 계획이다. [SEP] hypothesis: 롯데는 어떠한 정보도 제공하지 않을 계획이다.	contradiction

Table 6: The input and output examples in BUFFET. We show one example from one dataset per task. Due to the long input length, we do not include a summarization example in this table.

for sentiment analysis, and the TYDIQA-QG English set for question generation.

For ENGLISH FT, we limit the number of English training samples to 100,000 and fine-tune mT5-base (Xue et al., 2021) for 3 epochs. For the ENGLISH FT baseline, we transfer this model directly to new languages, while for ENG.+TGT. FT, we initialize the model checkpoint with the trained model weight and further fine-tune a model on few-shot samples for 300 epochs.

In-context learning. Different models have different maximum context window sizes: mT0 only accepts up to 1024 tokens, while BLOOMZ and ChatGPT accept up to 2048 and 4096, respectively. We use training instances up to the maximum context window. We set the maximum token length to 15 except for XLSUM and TYDIQA-QG. For XLSUM, we set the maximum token length to 100, and for TYDIQA-QG, we set the maximum token length to 50. We use greedy decoding throughout the experiments. For BLOOM-based model evaluations, we use a single RTX-100 GPU with 24 GB GPU memory. We use int8bit quantization to avoid GPU out-of-memory errors. To evaluate mT5 and mT0, we use TPU v3-8.

We found English-centric LMs (Llama1, Llama2, Llama2-chat, and Mistral) show strong abilities of in-context learning and often can generate output in expected formats (e.g., selecting a class label). To accelerate evaluations, we make those models directly predict outputs, rather than computing prompt token probabilities of input se-

Dataset	Instructions
NLI	Take the premise sentence as truth. Then the hypothesis is true (entailment), false (contradiction) or inconclusive (neutral)?
PAWS-X	In this task you are given a sentence pair that has high lexical overlap. If the sentences have the same meaning and are just paraphrases of each other label them as duplicate, if not label them as not_duplicate
SENTIMENT	In this task, you're given a review from Amazon. Your task is to generate a rating for the product. The rating is extremely negative, negative, neutral, positive, and really positive.
XCOPA	In this task you are given a premise and two alternatives (A) and (B). You must choose the alternative that is more plausibly the cause or effect of the situation described by the premise.
XWINOGRAD	Replace the _ in the above sentence with the correct option
QA	Read the given passage and answer a question about the information present in the passage.
NER	Given the following sentence, indicate the name entities (i.e., the real-world objects such as a person, location, organization, etc. that can be denoted with a proper name) such as "New York Times". For each word of a named-entity, indicate their type "location" or "organization" or "person".
SUMMARIZATION	In this task, you are given an article. Your task is to summarize the article in a sentence.
QG	This task is about reading the given passage and constructing a question about the information present in the passage.

Table 7: The list of English instructions for each task in BUFFET.

Task	Dataset	Languages
NLI	AMERICAS NLI	aym, bzd, cni, gn, hch, nah, too, quy, shp, tar
	KLUE NLI	ko
	OCNLI	zh
	PARSI NLU ENTAILMENT	fa
	XNLI	ar, bg, de, el, en, es, fr, hi, ru, sw, th, tr, ur, vi, zh
PARAPHRASE DETECTION	PAWS	(en,) de, es, fr, ja, ko, zh
SENTIMENT	AMAZON REVIEW	(en), de, es, fr, ja, zh
ANALYSIS	INDIC SENTIMENT	as, bn, brx, gu, hi, kn, mai, ml, mr, or, pa, ta, te, ur
COMMONSENSE	XCOPA	et, ht, it, id, qu, sw, zh, ta, th, tr, vi
COMMONSENSE	XWINOGRAD	(en,) ja, pt, ru, zh
QA	TYDIQA	(en,) ar, be, fi, id, sw, ko, ru, te
NER	WIKIANN	(en,) ta, fr, it, ja, vi, qu, be, gu, et, th, or, kn, fi, gn, ru, el, ur, es, hi, te, as, sw, pa, bg, ml, tr, fa, id, ko, mr, de, ar, bn, zh
	MASAKHANER	amh, hau, ibo, kin, luo, pcm, swa, wol, yor
SUMMARIZATION	XLSUM	(english,) ta, vi, id, tr, ja, th, bn, ar, en, es, fa, zh, sw
QG	TYDIQA-QG	(en,) ar, be, fi, id, sw, ko, ru, te

Table 8: The list of datasets with language lists in BUFFET.

Task	Dataset	Languages
NLI	AMERICAS NLI	aym, cni, hch
	KLUE NLI	ko
	PARSI NLU ENTAILMENT	fa
	XNLI	bg, el, hi, sw, ur
Paraphrase Detection	PAWS-X	de, es, ja, ko, zh
Sentiment	AMAZON REVIEW	de, fr, ja, zh
Analysis	INDIC SENTIMENT	bn, ta, ur
Commonsense	XCOPA	et, it, ta, th, tr
	XWINOGRAD	pt, ru
QA	TYDIQA	be, id, sw
NER	WIKIANN	be, bg, el, et, fi, it
	MASAKHANER	yor
Summarization	XLSUM	bn, fa, es, id, tr, vi
QG	TYDIQA-QG	ar, fi, ko, ru, te

Table 9: The subset of datasets and languages included in BUFFET-Light.

quence followed by each class token.

C Detailed BUFFET Results

This section includes the full list of the experimental results. Overall results on the full BUFFET are available in Table 10, and Figure 6 summarizes overall performance across the eight tasks, on the BUFFET-Light subset.

The overall trends on BUFFET-Light remain the same as the original BUFFET. This indicates BUFFET-Light is a reliable and more efficient alternative for holistic evaluations for few-shot cross-lingual transfer. Note that ChatGPT is only evaluated on the BUFFET-Light subsets due to the expensive API costs of experiments.

ChatGPT has strong generation capabilities but requires careful instruction design. As discussed, although ChatGPT significantly outperforms other LLMs with in-context learning, its performance often lags behind fine-tuning-based methods in some discriminative tasks, particularly in less-represented languages. ChatGPT, however, significantly outperforms fine-tuned models on tasks that require target language generations (e.g., question generation, QA) except summarization (XLSUM). On XLSUM, we found that ChatGPT often generates semantically correct summarizations in English rather than in the input article language, resulting in low ROUGE-2 scores. We do not observe that phenomenon in other LLMs (e.g., BLOOMZ); we show some ChatGPT output examples in the Appendix Table 25. Though more prompt engineering can boost ChatGPT's performance in summarization (Huang et al., 2023), we use the same prompts throughout the evaluations for a fair comparison. We also observe that when instructions are given in the target language, ChatGPT often outputs a summary in the language, as shown in improved XLSUM performance in ChatGPT TARGET ICL.

Below, we present the performance breakdown for each dataset. "–" indicates that ChatGPT is not evaluated on the subset as it is not included in the BUFFET-Light subset.

C.1 NLI

Table 11 shows the full results on AMERICASNLI. Table 12 shows the full results on XNLI. Table 13 presents the full results on the other three entailment datasets annotated in each language, KLUENLI, OCNLI, and PARSINLUENTAILMENT.

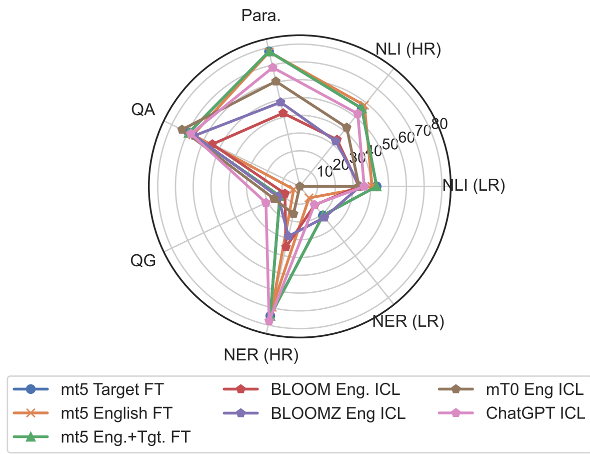

Figure 6: Overall results on BUFFET-Light.

On XNLI, ENGLISH FT (zero-shot transfer) shows strong performance and often outperforms ENG.+TGT. FT (few-shot transfer). Among ICL baselines, mT0 ZICL shows the best macro performance on XNLI. However, on AMERICASNLI, all methods struggle, while ENG.+TGT. FT shows the best macro performance on AMERICAS NLI. The performance gap between ENGLISH FT and ENG.+TGT. FT get significantly larger, with the largest gap in Aymara (5.5%). Despite its strong performance on XNLI, mT0 ZICL struggles in AMERICAS NLI (33.7% on average).

While mT0 ZICL shows robust performance across XNLI languages, ChatGPT shows a large performance gap between higher-resource languages and low-resource languages (57% in Greek v.s. 33% Urdu).

C.2 Paraphrase Detection

The results on PAWS-X results are available in Table 14. ENGLISH FT shows the best performance on this task among non-instruction-tuned models. We hypothesize that as the languages included in PAWS-X are all relatively well-represented languages and the task is relatively simple, ENGLISH FT, which is not trained in the target languages, can achieve high performance. mT0 ZICL shows quite high performance, likely because the model is trained on PAWS-X (Muennighoff et al., 2023).

C.3 Sentiment Analysis

The experimental results on AMAZON REVIEW MULTI and INDIC SENTIMENT are available in Tables 15 and 16. On both datasets, all models yield high accuracy across languages, except for mT5 ZEICL.

	Output	Classification			Multiple Choice		Span	Str.	Generation		Avg.	
	Tasks	NLI	Sent.	Para.	XCPA	XWGD	QA	NER	QG	Summ.	class	gen
Random		33.3	50.0	50.0	50.0	50.0	–	–	–	–	–	
TGT. FT	mT5	34.6	67.2	47.2	46.7	50.0	8.3	30.8	3.4	2.8	40.2	3.1
ENG. FT	mT5	46.0	89.7	78.6	49.5	48.4	62.9	30.8	4.2	4.0	57.9	4.1
ENG.+TGT.	mT5	**48.8**	**90.4**	**77.9**	49.9	49.0	**66.7**	**43.5**	12.2	**8.4**	**58.8**	**10.0**
ENG. ICL	BLOOM	33.6	85.3	42.4	50.0	50.8	39.2	25.0	11.6	2.4	44.0	7.0
	mT5	34.5	50.0	43.2	50.0	49.2	0.3	1.6	0.0	0.3	32.1	0.1
	BLOOMZ	33.0	87.2*	49.5*	50.5	52.1	44.5*	20.0	13.9	9.0*	44.3	11.4
	mT0	33.6	79.9*	61.1*	57.1	59.6	69.0*	7.9	15.3	1.5*	45.6	8.4
	ChatGPT†	54.5	91.1	68.6	76.7	73.3	68.1	45.4	21.2	5.4	64.6	13.3
TGT. ICL	BLOOM	31.7	85.3	45.9	50.1	51.7	7.0	25.2	12.8	4.7	41.2	8.7
	mT5	34.4	50.0	43.1	50.0	47.3	0.2	0.2	0.0	0.3	31.7	0.1
	BLOOMZ	32.1	64.7*	51.7*	50.5	53.1	43.7*	19.1	12.0	10.9*	40.6	11.4
	mT0	38.1	70.6*	60.9*	52.8	57.9	70.8*	8.5	14.6	1.8*	45.7	8.2
	ChatGPT†	48.2	91.5	68.2	74.3	73.4	68.0	44.8	21.1	11.4	62.7	16.3
Z-EICL	BLOOM	32.3	35.8	42.3	50.1	46.4	3.1	0.0	**16.4**	4.1	28.8	10.0
	mT5	34.2	50.0	42.4	50.1	46.4	2.0	0.0	0.1	1.3	32.5	0.7
	BLOOMZ	34.0	51.6*	58.0*	50.1	50.9	65.2*	7.6	10.2	2.9*	39.3	6.6
	mT0	49.1	90.2*	91.2*	64.1	64.5	75.2*	0.0	10.3	8.5*	56.0	9.4

Table 10: **Overall experiment results on BUFFET**. The blue-colored rows are instruction-tuned models, and we added * symbols next to the scores for the tasks on which the models have been trained. "Random" shows random baseline performance. **Bold** fonts indicate the best results for each task, among the models that are not directly trained on the task. When ChatGPT achieves the best results, we also note the second-best number from the models that are not trained on the task, acknowledging the possibility that ChatGPT may have encountered a similar task during training.

Transfer + Model	Macro	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar
Target FT	35.9	36.0	35.5	35.5	35.7	32.7	37.5	35.2	35.4	37.6	37.8
English FT	42.6	40.7	44.9	43.3	46.8	38.0	42.5	41.6	46.1	43.2	39.2
English Target FT	45.1	46.2	48.6	45.0	49.7	38.8	46.8	44.2	46.4	42.5	43.0
EICL BLOOM	33.7	33.4	34.6	33.2	34.1	33.3	33.5	33.4	34.3	34.0	33.6
EICL mT5	33.3	33.3	32.8	33.3	33.3	33.2	33.2	33.2	33.3	33.3	33.3
EICL BLOOMZ	33.3	33.1	33.5	33.7	33.3	33.3	33.8	32.0	33.3	33.3	33.3
EICL mT0	33.3	33.3	33.2	33.3	33.3	33.4	33.3	33.3	33.4	33.3	32.9
EICL ChatGPT	36.3	33.6	–	40.9	–	34.3	–	–	–	–	–
TICL BLOOM	33.7	33.5	34.6	33.2	33.6	33.3	33.5	33.3	34.3	34.0	33.6
TICL mT5	33.3	33.3	32.8	33.3	33.6	33.2	33.2	33.3	33.3	33.3	33.3
TICL BLOOMZ	33.4	33.3	33.5	33.7	33.3	33.3	33.8	33.4	33.3	33.3	33.3
TICL mT0	33.4	33.6	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
TICL ChatGPT	34.7	33.6	–	36.7	–	33.9	–	–	–	–	–
ZICL BLOOM	33.5	33.7	32.0	33.7	32.5	34.7	31.6	33.8	34.7	34.7	33.9
ZICL mT5	34.0	36.3	34.4	32.9	32.8	36.4	33.6	33.7	32.9	33.3	34.1
ZICL BLOOMZ	34.3	36.3	33.5	33.7	33.3	36.4	33.6	33.7	32.9	33.3	34.1
ZICL mT0	33.7	33.5	33.5	33.3	33.7	33.3	34.1	33.2	35.3	33.1	33.5

Table 11: Model performance on AMERICASNLI. We report the average of the three few-shot samples.

C.4 Commonsense

XCOPA. The experimental results on XCOPA are available in Table 17. On XCOPA, ChatGPT and mT0 (Z EICL) yield high performance across languages. ChatGPT achieves particularly notable performance in Italian (91.2%). On the other hand, all of the fine-tuning-based methods struggle, as the small size of the source datasets in English. This result indicates that for a task that lacks a large-scale training dataset even in high-resource languages, LLMs using in-context learning may often result in higher performance. We observed that mT0 ENGLISH FT faces difficulties when applied

Transfer + Model	Macro	ar	bg	de	el	es
Target FT	36.4	35.8	37.8	37.3	37.4	37.0
English FT	59.4	59.2	62.9	61.5	61.4	63.7
English Target FT	57.3	57.7	59.5	59.0	59.4	62.7
EICL BLOOM	33.7	34.0	33.9	33.4	33.3	34.2
EICL mT5	33.3	33.3	33.3	33.3	33.3	33.3
EICL BLOOMZ	33.1	34.1	33.6	33.7	27.9	34.2
EICL mT0	36.3	37.8	36.3	35.3	33.4	33.7
EICL ChatGPT	50.3	–	60.7	–	54.0	–
TICL BLOOM	33.4	33.6	32.7	33.2	33.7	32.9
TICL mT5	33.3	33.3	33.3	33.3	33.2	33.3
TICL BLOOMZ	33.4	33.3	33.7	33.3	34.4	33.3
TICL mT0	40.4	38.8	51.2	41.8	47.8	43.1
TICL ChatGPT	50.5	–	52.4	–	56.9	–
ZICL BLOOM	33.6	33.7	34.1	34.3	33.7	33.7
ZICL mT5	32.3	32.8	32.1	32.5	32.3	30.6
ZICL BLOOMZ	32.1	–	–	–	–	–
ZICL mT0	56.2	56.1	58.4	58.7	57.5	58.0

Transfer + Model	fr	hi	ru	sw	th	tr	ur	vi	zh
Target FT	37.4	35.7	36.0	35.1	36.7	36.8	34.2	36.3	35.5
English FT	62.1	58.0	59.8	55.5	57.4	58.4	54.0	57.1	60.4
English Target FT	59.0	55.1	60.1	52.3	56.4	56.1	51.6	55.8	58.3
EICL BLOOM	36.2	33.4	33.6	33.4	33.3	33.3	33.3	33.3	33.4
EICL mT5	33.4	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
EICL BLOOMZ	35.1	33.4	32.1	33.9	33.0	32.1	33.1	33.2	33.8
EICL mT0	47.3	36.3	34.9	35.8	33.4	38.1	34.9	37.9	33.7
EICL ChatGPT	–	48.0	–	55.9	–	–	33.1	–	–
TICL BLOOM	33.3	33.3	33.2	34.3	34.8	33.8	33.6	32.5	33.0
TICL mT5	33.3	33.2	33.3	33.3	33.5	33.3	33.3	33.3	33.3
TICL BLOOMZ	32.9	33.2	34.0	33.6	33.7	32.9	33.1	32.8	33.3
TICL mT0	39.7	39.9	47.7	37.3	37.4	33.5	35.7	35.3	36.8
TICL ChatGPT	–	51.8	–	47.3	–	–	44.2	–	–
ZICL BLOOM	34.0	33.4	33.5	33.9	33.3	33.1	34.7	33.3	32.3
ZICL mT5	29.6	33.3	32.3	32.7	33.1	34.7	32.8	32.4	31.1
ZICL BLOOMZ	–	–	–	–	–	–	32.8	32.4	31.1
ZICL mT0	58.7	55.3	57.0	53.7	51.6	56.1	54.5	57.3	54.5

Table 12: Model performance on XNLI. We report the average of the three few-shot samples.

Transfer + Model	KLUENLI	ParsinluEntailment	OCNLI
Target FT	34.0	34.6	34.0
English FT	57.9	51.1	32.5
English Target FT	61.1	50.5	38.6
EICL BLOOM	33.8	28.9	38.9
EICL mT5	33.3	40.4	31.0
EICL BLOOMZ	31.9	28.8	38.2
EICL mT0	34.3	30.0	36.7
EICL ChatGPT	64.8	62.3	–
TICL BLOOM	33.4	28.8	38.2
TICL mT5	33.3	40.4	30.5
TICL BLOOMZ	33.8	29.0	32.1
TICL mT0	43.1	37.4	38.6
TICL ChatGPT	56.5	50.2	–
ZICL BLOOM	33.8	37.4	32.0
ZICL mT5	32.4	31.9	37.6
ZICL BLOOMZ	32.4	31.9	37.6
ZICL mT0	56.9	55.2	50.6

Table 13: Model performance on KLUENLI, OCNLI and ParsiNLUEntailment. We report the average of the three few-shot samples.

to XCOPA. This could be attributed to the limited size of the XCOPA English set, which might not provide enough data for a smaller mT5-base model to acquire comprehensive task knowledge.

Transfer + Model	Macro	de	es	fr	ja	ko	zh
Target FT	47.2	47.5	48.8	47.1	48.1	44.2	47.3
English FT	78.6	79.9	83.5	84.0	74.5	74.3	75.5
English Target FT	77.9	79.9	82.6	81.0	73.1	73.9	77.0
EICL BLOOM	42.4	41.5	42.3	43.0	42.7	42.0	42.8
EICL mT5	43.2	41.5	42.4	47.7	42.7	42.0	42.6
EICL BLOOMZ	49.5	58.9	58.9	57.7	34.5	29.5	57.8
EICL mT0	61.1	78.7	57.6	57.8	57.3	58.0	57.4
EICL ChatGPT	68.6	73.5	72.0	–	67.4	60.1	69.8
TICL BLOOM	45.9	49.3	42.3	42.4	42.9	54.9	43.0
TICL mT5	43.1	41.5	46.4	43.0	42.7	42.0	42.6
TICL BLOOMZ	51.7	47.4	56.4	51.3	48.8	55.6	50.4
TICL mT0	60.9	67.9	68.1	57.0	57.3	58.0	57.4
TICL ChatGPT	68.5	71.9	71.5	–	67.0	62.8	69.1
ZICL BLOOM	42.4	41.6	42.4	42.9	43.0	42.0	42.7
ZICL mT5	58.0	58.0	57.8	58.6	57.7	58.1	57.5
ZICL BLOOMZ	58.0	58.0	57.8	58.6	57.7	58.1	57.5
ZICL mT0	91.2	91.5	95.5	94.3	87.5	87.9	90.8

Table 14: Model performance on PAWSX. We report the average of the three few-shot samples.

Transfer + Model	Macro	de	zh	es	fr	ja
Target FT	76.3	72.9	77.1	76.1	82.3	73.1
English FT	91.9	94.2	84.5	93.8	95.1	91.8
English Target FT	92.4	93.6	87.6	93.4	94.9	92.3
EICL BLOOM	83.4	82.0	84.9	92.8	88.0	69.2
EICL mT5	50.2	49.4	50.6	50.9	50.6	49.8
EICL BLOOMZ	81.5	75.7	80.2	93.8	93.5	64.3
EICL mT0	79.8	88.7	70.6	81.8	89.6	68.5
EICL ChatGPT	85.8	94.3	87.5	–	96.1	65.0
TICL BLOOM	84.2	87.3	85.7	92.8	84.2	70.9
TICL mT5	50.2	49.4	50.6	50.9	50.6	49.8
TICL BLOOMZ	64.9	57.1	71.2	79.2	61.5	55.5
TICL mT0	72.2	88.9	51.3	58.9	85.1	76.8
TICL ChatGPT	89.7	94.4	85.5	–	95.6	83.2
ZICL BLOOM	50.3	49.4	50.6	50.9	50.7	49.8
ZICL mT5	45.1	48.5	49.6	39.9	37.0	50.4
ZICL BLOOMZ	15.6	23.9	18.4	6.0	9.6	19.8
ZICL mT0	87.3	90.5	72.7	90.8	93.0	89.5

Table 15: Model performance on AMAZON REVIEWS MULTI. We report the average of the three few-shot samples.

XWINOGRAD. The experimental results on XWINOGRAD are available in Table 18. Similar to XCOPA, on XWINOGRAD, fine-tuning-based methods often struggle, while in-context learning with competitive LLMs yields strong performance.

C.5 Question Answering

TYDIQA experimental results are available in Table 19. Both the fine-tuning and ICL methods exhibit commendable performance on this particular task. It is intriguing to note that both mT0 and BLOOMZ demonstrate relatively lower efficacy in Korean, Finnish, and Russian. This can be attributed to the fact that these languages were not included during the pretraining phase.

C.6 Named Entity Recognition

WIKIANN. Table 20 contains the results for WIKIANN. We specifically present the few-shot results since we discovered that zero-shot baselines consistently exhibit extremely poor performance, often close to zero, primarily because generating the answer in the precise output format proves to be challenging.

It's important to acknowledge that the BUFFET-Light WIKIANN subset comprises languages that are relatively high-resource, which could potentially lead to an overestimation of ChatGPT's performance. When comparing the best fine-tuning method with ChatGPT in the BUFFET-light languages, they generally perform competitively, with the exception of Finnish.

Transfer + Model	Macro	as	bn	brx	gu	hi
Target FT	58.2	61.4	55.8	62.6	56.7	64.1
English FT	87.4	85.0	87.4	89.4	88.4	91.6
English Target FT	88.4	84.6	90.2	90.6	89.7	93.0
EICL BLOOM	87.2	83.7	87.6	91.2	86.1	92.0
EICL mT5	49.8	49.8	49.8	49.8	49.8	49.8
EICL BLOOMZ	93.0	89.6	94.2	94.9	93.1	95.6
EICL mT0	79.9	73.6	88.4	81.3	80.2	81.1
EICL ChatGPT	89.3	–	91.8	–	–	–
TICL BLOOM	86.5	83.1	86.7	91.2	84.1	92.6
TICL mT5	49.8	49.8	49.8	49.8	49.8	49.8
TICL BLOOMZ	64.5	67.0	61.2	94.9	52.8	56.5
TICL mT0	69.0	87.4	82.9	50.1	78.2	68.3
TICL ChatGPT	89.7	–	92.6	–	–	–
ZICL BLOOM	49.7	49.8	49.8	49.8	49.8	49.8
ZICL mT5	26.5	24.4	24.4	24.8	26.0	26.1
ZICL BLOOMZ	64.5	67.0	61.2	94.9	52.8	56.5
ZICL mT0	93.2	90.5	93.7	94.3	92.2	95.3

Transfer + Model	kn	mai	ml	mr	or	pa	ta	te	ur
Target FT	59.5	62.6	45.8	60.4	62.7	48.9	57.8	55.0	60.8
English FT	88.4	89.4	86.9	86.1	77.2	90.4	87.0	86.7	90.3
English Target FT	89.6	90.6	86.4	86.2	77.9	91.6	87.4	88.5	91.1
EICL BLOOM	83.0	91.2	85.8	88.9	85.8	89.0	85.0	86.0	85.1
EICL mT5	49.8	49.8	49.8	49.8	49.8	49.8	49.8	49.8	49.8
EICL BLOOMZ	92.7	94.9	91.8	92.4	93.8	94.2	90.6	90.5	93.5
EICL mT0	74.8	71.6	83.2	81.6	78.3	88.1	86.7	78.0	71.7
EICL ChatGPT	–	–	–	–	–	–	82.3	–	93.9
TICL BLOOM	81.8	91.2	84.0	88.2	85.0	88.2	85.3	85.1	84.1
TICL mT5	49.8	49.8	49.8	49.8	49.8	49.8	49.8	49.8	49.8
TICL BLOOMZ	49.7	94.9	66.3	58.3	59.2	57.3	68.2	50.3	66.9
TICL mT0	72.1	49.7	84.4	79.7	66.1	68.8	55.3	58.7	64.9
TICL ChatGPT	–	–	–	–	–	–	83.9	–	92.4
ZICL BLOOM	49.8	49.8	49.3	49.8	49.8	49.8	49.6	49.8	48.7
ZICL mT5	26.8	24.8	29.0	20.7	22.4	32.4	25.4	28.9	34.5
ZICL BLOOMZ	26.8	24.8	29.0	20.7	22.4	32.4	25.4	28.9	34.5
ZICL mT0	93.5	94.3	92.0	92.8	91.2	95.2	92.3	92.9	94.6

Table 16: Model performance on INDIC SENTIMENT. We report the average of the three few-shot samples.

Transfer + Model	Macro	et	ht	it	id	qu	sw	zh	ta	th	tr	vi
Target FT	46.7	50.0	50.1	48.3	50.5	50.4	32.5	49.8	49.3	49.4	33.9	50.0
English FT	48.4	49.8	50.2	49.6	51.0	48.6	48.8	49.0	50.8	48.0	49.6	49.2
English Target FT	49.9	50.3	49.9	49.6	49.2	50.5	50.4	50.4	49.2	50.7	49.5	49.4
EICL BLOOM	50.0	51.5	49.0	49.9	50.0	50.6	50.0	50.1	49.5	50.0	49.9	50.0
EICL mT5	50.0	50.0	49.9	50.7	50.0	49.5	49.8	49.9	50.7	50.0	50.0	50.0
EICL BLOOMZ	50.5	50.7	51.2	50.9	50.0	52.7	49.9	50.0	50.1	49.8	49.8	50.0
EICL mT0	57.1	60.7	60.6	53.4	59.8	50.0	61.6	64.1	51.9	54.1	54.1	58.1
EICL ChatGPT	76.7	87.6	–	91.2	–	–	–	–	54.6	62.6	87.4	–
TICL BLOOM	50.1	49.8	50.4	50.4	49.0	48.8	52.2	50.6	49.6	50.0	49.8	50.2
TICL mT5	50.0	49.9	50.0	49.9	50.0	50.0	49.9	50.0	50.0	50.0	49.5	50.9
TICL BLOOMZ	50.5	45.6	50.8	50.4	53.4	47.4	49.8	51.8	53.2	50.0	49.4	53.4
TICL mT0	52.8	50.4	51.9	51.0	51.9	50.6	53.7	50.5	50.1	50.6	54.3	65.5
TICL ChatGPT	74.4	89.2	–	91.6	–	–	–	–	49.5	55.7	86.2	–
ZICL BLOOM	50.9	51.8	48.8	51.2	51.4	50.6	51.2	53.6	52.4	48.2	49.8	50.6
ZICL mT5	50.1	49.8	50.4	50.4	49.0	48.8	52.2	50.6	49.6	50.0	49.8	50.2
ZICL BLOOMZ	50.1	48.6	50.2	52.4	47.4	50.8	45.2	46.8	54.8	50.6	52.8	51.0
ZICL mT0	64.1	64.0	62.2	66.2	70.0	48.8	66.2	71.8	61.0	63.0	65.0	67.2

Table 17: Model performance on XCOPA. We report the average of the three few-shot samples.

MASAKHANER. Results on MASAKHANER are available at Table 21. In this benchmark, all ICL methods, including ChatGPT, encounter difficulties, whereas TARGET FT and ENG.+TGT. FT consistently demonstrates strong performance across various languages. Notably, by incorporating an

Transfer + Model	Macro	jp	pt	ru	zh
Target FT	50.0	48.4	50.3	49.9	51.4
English FT	48.4	52.2	52.2	45.4	51.2
English Target FT	49.0	48.4	48.4	48.8	50.6
EICL BLOOM	50.8	49.6	48.0	54.0	51.5
EICL mT5	49.2	48.4	49.5	47.4	51.3
EICL BLOOMZ	52.1	52.6	50.3	55.3	50.1
EICL mT0	59.6	62.2	57.7	53.2	65.2
EICL ChatGPT	73.3	–	74.1	72.5	–
TICL BLOOM	51.7	52.2	50.2	54.3	50.1
TICL mT5	47.3	48.4	46.2	44.4	50.3
TICL BLOOMZ	53.1	52.7	54.5	55.3	50.0
TICL mT0	57.9	54.9	57.2	62.9	56.5
TICL ChatGPT	71.6	–	70.4	72.8	–
ZICL BLOOM	53.7	51.9	54.4	56.7	51.9
ZICL mT5	46.4	47.4	48.5	45.7	44.2
ZICL BLOOMZ	50.9	51.9	51.9	50.2	49.6
ZICL mT0	64.5	68.7	59.8	62.2	67.3

Table 18: Model performance on XWinograd We report the average of the three few-shot samples.

Transfer + Model	Macro	ar	be	fi	id	sw	ko	ru	te
Target FT	8.3	8.1	6.1	9.1	6.4	5.5	7.5	9.2	14.7
English FT	62.9	61.0	63.2	65.3	69.2	67.9	57.1	56.3	63.5
English Target FT	66.7	65.9	68.0	63.6	70.0	69.3	60.6	65.1	70.7
EICL BLOOM	39.2	43.8	58.2	20.6	47.0	57.5	23.2	32.7	30.4
EICL mT5	0.3	0.7	0.1	0.4	0.2	0.3	0.0	0.3	0.0
EICL BLOOMZ	44.5	45.3	67.7	18.9	61.0	73.7	12.4	19.6	57.6
EICL mT0	69.0	54.0	75.8	68.9	68.8	75.5	68.1	53.7	86.7
EICL ChatGPT	70.8	–	58.9	–	76.5	77.0	–	–	–
TICL BLOOM	7.0	13.2	11.9	1.7	19.1	4.5	0.7	1.3	3.7
TICL mT5	0.2	0.4	0.1	0.2	0.6	0.2	–	0.3	–
TICL BLOOMZ	43.7	44.7	63.7	17.5	60.3	71.5	12.1	20.3	59.3
TICL mT0	70.8	58.7	75.8	66.9	72.1	78.3	72.1	65.9	76.6
TICL ChatGPT	66.7	–	46.0	–	76.7	77.4	–	–	–
ZICL BLOOM	2.0	2.2	1.1	3.1	3.2	2.3	1.0	1.5	1.7
ZICL mT5	65.2	80.0	86.3	7.3	81.3	82.0	44.7	55.0	85.1
ZICL BLOOMZ	65.2	80.0	86.3	7.3	81.3	82.0	44.7	55.0	85.1
ZICL mT0	75.2	71.8	84.4	67.3	77.3	78.6	68.3	65.0	88.9

Table 19: Model performance on TyDiQA. We report the average of the three few-shot samples.

additional 32 training examples, Eng.+Tgt. FT achieves a significant 34% improvement in performance for Hausa. These remarkable enhancements underscore the effectiveness of fine-tuning a specialized model on a limited set of training samples in target languages.

C.7 Generation

TyDiQA-QG. The experimental results for TyDiQA-QG are available in Table 22. On this task, ChatGPT and mT0 English ICL show superior performance than smaller fine-tuned models, demonstrating their competitiveness in generating fluent text in target languages.

XLSum. XLSum results are available in Table 23. Despite strong generation capability, ChatGPT English ICL performance remains low. We found that when instructed in English, ChatGPT often generates summaries in English, not in the article language. We haven't observed such behaviors on other tasks or other LLMs. ChatGPT Target ICL shows large improvements from English ICL, which has not been observed in other tasks. When instructions in the target language are given, ChatGPT almost always generates a summary in the target language.

Among non-instruction-tuned models, Eng.+Tgt. FT yields the highest average performance. It should be noted that mT0 and BLOOMZ are trained on XLSum. Nevertheless, their performance in some languages remains low.

Transfer + Model	Macro	ta	fr	it	ja	vi	be	gu	et	th
Target FT	43.7	0.2	59.0	55.5	43.9	58.3	63.5	26.0	54.4	23.7
English FT	52.2	0.8	78.2	79.4	56.1	80.5	73.9	24.0	60.5	10.7
English Target FT	56.6	0.8	78.1	76.8	55.7	75.9	76.8	37.0	76.0	25.6
EICL BLOOM	32.8	0.6	51.6	51.0	22.1	53.8	25.6	22.3	37.0	1.7
EICL mT5	1.6	0.0	0.0	0.0	0.0	0.0	3.3	0.3	0.0	0.0
EICL BLOOMZ	22.4	0.5	37.1	43.4	15.6	36.8	15.4	13.0	29.6	0.3
EICL mT0	15.8	0.1	13.8	13.0	9.1	22.9	11.0	6.0	24.1	1.4
EICL ChatGPT	77.6	–	–	81.8	–	–	78.2	–	78.2	–
TICL BLOOM	32.8	0.7	52.5	50.2	20.8	53.5	24.4	24.0	34.0	1.0
TICL mT5	0.3	0.0	0.0	0.1	0.0	0.1	0.2	1.3	0.0	1.7
TICL BLOOMZ	20.7	0.6	37.3	39.8	15.0	32.1	13.5	8.7	25.1	0.2
TICL mT0	15.8	0.1	13.8	13.0	9.1	22.9	11.0	6.0	24.1	1.4
TICL ChatGPT	76.8	–	–	82.3	–	–	78.4	–	76.9	–

Transfer + Model	or	kn	fi	gn	ru	el	ur	es	hi	te	as
Target FT	36.5	12.5	55.5	60.3	50.1	59.0	68.4	54.9	42.4	7.0	25.3
English FT	35.5	11.0	64.2	71.0	60.4	73.4	79.6	75.7	47.9	6.6	26.0
English Target FT	40.0	22.5	74.8	68.0	67.8	74.4	79.1	78.3	53.7	9.5	28.3
EICL BLOOM	22.0	6.0	39.5	47.3	26.1	20.4	70.7	55.2	40.2	5.6	22.7
EICL mT5	0.0	1.3	0.0	0.0	0.0	0.0	10.1	0.0	10.0	0.0	0.7
EICL BLOOMZ	10.0	5.7	31.8	28.0	19.7	15.8	41.7	37.5	30.9	4.2	16.0
EICL mT0	16.3	3.3	15.2	24.3	15.1	12.8	47.1	20.3	18.7	3.3	10.0
EICL ChatGPT	–	–	81.5	–	–	72.4	–	–	–	–	–
TICL BLOOM	25.3	6.7	37.6	49.0	26.2	19.7	71.7	55.6	39.9	5.3	24.0
TICL mT5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	0.0	1.0
TICL BLOOMZ	6.5	4.0	26.5	24.7	17.4	13.0	47.3	41.1	26.5	3.8	13.0
TICL mT0	16.3	3.3	15.2	24.3	15.1	12.8	47.1	20.3	18.7	3.3	10.0
TICL ChatGPT	–	–	81.9	–	–	69.3	–	–	–	–	–

Transfer + Model	sw	pa	bg	ml	tr	fa	id	ko	mr	de	ar	bn	zh
Target FT	57.5	29.7	54.2	19.7	55.4	48.0	64.2	36.1	34.8	51.2	40.6	43.0	49.9
English FT	61.0	35.5	67.0	21.4	64.5	60.5	81.6	36.2	36.6	75.1	52.9	48.7	66.6
English Target FT	75.3	42.3	67.1	24.5	79.5	57.6	80.7	57.7	44.7	73.2	52.9	47.7	65.2
EICL BLOOM	60.3	26.3	30.9	14.0	39.4	28.6	61.2	12.0	28.4	41.7	43.9	34.9	38.7
EICL mT5	0.0	0.7	0.0	0.0	0.0	0.0	0.3	0.0	0.4	6.7	16.7	3.7	0.0
EICL BLOOMZ	34.9	15.0	22.7	5.0	34.6	14.7	31.7	9.8	22.6	26.4	21.0	36.0	31.3
EICL mT0	24.3	10.0	14.7	5.0	20.2	21.4	23.4	11.2	12.3	15.7	23.0	23.9	27.7
EICL ChatGPT	–	–	73.3	–	–	–	–	–	–	–	–	–	–
TICL BLOOM	58.8	26.7	29.6	14.4	39.6	27.8	61.4	10.6	27.9	43.3	44.6	36.8	38.3
TICL mT5	0.4								0.5	0.1	0.4	0.3	
TICL BLOOMZ	26.8	14.0	19.7	4.2	31.3	14.7	35.2	8.1	20.4	22.4	23.6	36.2	31.0
TICL mT0	24.3	10.0	14.7	5.0	20.2	21.4	23.4	11.2	12.3	15.7	23.0	23.9	27.7
TICL ChatGPT	–	–	72.0	–	–	–	–	–	–	–	–	–	–

Table 20: Model performance on WIKIANN. We report the average of the three few-shot samples.

Transfer + Model	Macro	amh	hau	ibo	kin	luo	pcm	swa	wol	yor
Target FT	17.4	13.6	31.5	28.6	12.8	14.2	11.1	26.4	8.7	9.9
English FT	9.4	6.2	11.0	14.8	10.5	10.5	8.7	10.4	3.8	8.3
English Target FT	30.5	27.0	44.7	44.3	26.8	26.0	23.7	40.6	18.8	22.4
EICL BLOOM	17.2	3.4	23.8	27.4	18.5	11.6	15.2	24.9	16.3	13.9
EICL mT5	1.5	0.0	13.3	0.0	0.0	0.4	0.0	0.0	0.0	0.0
EICL BLOOMZ	14.9	0.2	11.3	28.4	14.3	4.6	12.4	24.4	17.7	21.0
EICL mT0	1.3	0.0	1.7	0.8	4.9	1.2	0.0	2.2	0.0	0.8
EICL ChatGPT	13.2	–	–	–	–	–	–	–	–	13.2
TICL BLOOM	17.2	3.4	23.8	27.4	18.5	11.6	15.2	24.9	16.3	13.9
TICL mT5	0.2	0.0	1.6	0.0	0.0	0.4	0.0	0.0	0.0	0.0
TICL BLOOMZ	14.9	0.2	11.3	28.4	14.3	4.6	12.4	24.4	17.7	21.0
TICL mT0	1.3	0.0	1.7	0.8	4.9	1.2	0.0	2.2	0.0	0.8
TICL ChatGPT	12.8	–	–	–	–	–	–	–	–	12.8

Table 21: Model performance on MASAKHANER. We report the average of the three few-shot samples.

Transfer + Model	Macro	ar	be	fi	id	sw	ko	ru	te
Target FT	3.4	2.7	4.1	2.5	4.4	3.2	2.8	2.1	5.8
English FT	4.2	2.1	3.5	5.1	6.2	5.1	3.0	4.7	4.2
English Target FT	12.2	11.5	7.3	15.8	14.1	13.1	7.9	8.9	18.8
EICL BLOOM	11.6	18.3	10.4	10.8	16.1	15.2	1.3	3.7	17.4
EICL mT5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
EICL BLOOMZ	13.9	19.5	14.2	7.8	23.6	23.1	0.7	2.1	20.3
EICL mT0	15.3	25.8	10.3	3.7	19.6	12.3	4.1	6.2	40.1
EICL ChatGPT	17.8	30.6	–	28.2	–	–	0.7	2.6	26.9
TICL BLOOM	12.8	18.1	9.6	10.0	15.7	14.9	7.7	9.2	16.8
TICL mT5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TICL BLOOMZ	12.0	16.0	10.7	5.0	20.0	21.1	1.9	5.2	15.9
TICL mT0	14.6	17.7	9.1	6.6	18.3	12.0	5.1	8.5	39.3
TICL ChatGPT	19.2	24.0	–	27.5	–	–	14.8	17.6	12.2
ZICL BLOOM	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.0
ZICL mT5	16.5	30.6	15.5	5.2	24.5	21.8	3.0	4.6	26.8
ZICL BLOOMZ	1.7	2.4	2.1	1.7	2.5	2.2	1.0	0.9	1.2
ZICL mT0	10.3	4.9	13.7	3.5	12.3	5.4	1.9	2.0	39.1

Table 22: Model performance on TyDiQA-QG. We report the average of the three few-shot samples.

Transfer + Model	Macro	Tamil	Vietnamese	Swahili	Indonesian
Target FT	2.8	0.8	11.0	2.0	1.7
English FT	4.0	0.1	18.4	7.8	4.9
English Target FT	8.4	10.9	24.7	8.8	7.8
EICL BLOOM	2.4	0.1	9.0	4.6	3.8
EICL mT5	0.3	0.0	1.7	0.4	0.2
EICL BLOOMZ	9.0	18.6	12.3	1.6	3.3
EICL mT0	1.8	0.0	10.4	5.3	1.0
EICL ChatGPT	5.4	–	19.5	–	4.9
TICL BLOOM	4.7	13.9	10.3	4.6	3.1
TICL mT5	0.3	0.0	1.7	0.3	0.3
TICL BLOOMZ	10.9	4.6	12.9	1.2	15.7
TICL mT0	1.8	0.0	10.4	5.3	1.0
TICL ChatGPT	11.4	–	19.5	–	7.2
ZICL BLOOM	4.1	0.1	10.7	9.0	9.5
ZICL mT5	1.3	0.5	4.8	1.1	0.7
ZICL BLOOMZ	4.3	0.0	0.0	0.0	9.5
ZICL mT0	8.5	1.1	26.9	18.3	16.8

Transfer + Model	Turkish	Japanese	Thai	Bengali	Arabic	Spanish	Persian	Chinese
Target FT	1.1	6.5	6.5	0.0	0.0	1.5	0.0	2.2
English FT	8.0	0.7	0.9	0.0	0.0	5.7	0.0	1.2
English Target FT	12.1	2.8	8.5	0.0	3.3	10.7	10.0	1.5
EICL BLOOM	5.2	0.3	0.2	0.0	0.1	3.7	0.0	1.1
EICL mT5	0.4	0.0	0.0	0.0	0.0	0.4	0.0	0.0
EICL BLOOMZ	7.0	0.9	48.6	0.0	0.0	5.0	10.0	0.4
EICL mT0	1.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0
EICL ChatGPT	2.4	–	–	–	–	–	–	–
TICL BLOOM	5.2	14.1	0.5	0.0	0.0	3.6	0.0	1.2
TICL mT5	0.5	0.0	0.0	0.0	0.0	0.4	0.0	0.0
TICL BLOOMZ	3.2	37.4	48.6	0.0	0.0	5.8	0.0	1.5
TICL mT0	1.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0
TICL ChatGPT	10.0	–	–	–	–	–	20.1	–
ZICL BLOOM	4.3	0.8	0.2	0.0	0.0	3.3	10.0	1.6
ZICL mT5	1.1	2.4	1.9	0.0	0.1	0.7	0.0	1.9
ZICL BLOOMZ	0.0	0.0	0.0	0.0	0.0	7.6	0.1	0.0
ZICL mT0	15.7	3.1	2.4	0.0	0.1	12.4	0.2	4.4

Table 23: Model performance on XLSum

D More Analysis

D.1 Performance across Languages

Figure 7 shows performance across languages on the three tasks, NLI, NER, and QA, adding two more LLMs: BLOOMZ and mT0. We observe performance drops in Finnish, Korean, and Russian for BLOOM and BLOOMZ in TyDiQA. Finnish, Korean, and Russian are excluded from BLOOM pretraining,[12] which we attribute to these performance drops. Conversely, mT5 fine-tuning-based methods consistently display strong performance across languages. Interestingly, in Bengali, which is often considered less represented, BLOOM achieves performance comparable to mT5 fine-tuned models. These results suggest pretraining setup may strongly affect downstream task performance even after instruction tuning.

D.2 Variances of Different k-shots

In Section 3, we show that different sets of demonstrations can cause significant performance differences. We provide the full visualization results across different tasks.

D.3 Variances of the Varying Number of k

We provide the full experimental results with a different number of k. We evaluate BLOOM English ICL, BLOOMZ English ICL and mT5-Eng.+Tgt. Fine-tuning and mT0 English ICL experimental results on Amazon Review, TyDiQA, TyDiQA-AG, WikiANN, and in Figures 8, 9, 10 and 11, respectively.

Amazon Review. On Amazon Review, All models except for BLOOM (pretraining only) show competitive zero-shot performance. BLOOM English ICL benefits from few-shot demonstrations while mT0 English ICL exhibit performance deterioration as adding more demonstrations across languages.

TyDiQA. Among English ICL baselines, mT0 shows strong performance up to four demonstrations, although their performance gets really low once more demonstrations are added. Similar deterioration happens in BLOOMZ. On the contrary, BLOOM performance improves as more shots are added.

TyDiQA-QG. Unlike in Amazon Review or TyDiQA, BLOOMZ English ICL shows performance improvements with more demonstrations in Arabic and Bengali, reaching the highest QG performance in Bengali with four demonstrations. On the contrary, both BLOOM and BLOOMZ show poor QG performance in Korean and Russian, possibly due to the lack of those languages during pretraining.

WikiANN. On WikiANN, all of the models gain performance improvements by adding at least one demonstration, possibly due to the difficulty of learning the exact output format expected given the instruction only. As in other datasets, mT0 reaches its highest performance with four demonstrations. mT5 Eng.+Tgt. FT exhibits performance drops with one shot, possibly due to their overfit to the single example.

D.4 Variances of Different Instructions

We investigate the effectiveness of different English instructions on question generation tasks for TyDiQA in the four-shot setting using mT0 and BLOOM as base models in Table 24. We compare four relevant instructions and one irrelevant instruction (an instruction for Amazon Review).

In the four-shot setting, whether the instruction is relevant does not make a huge difference for BLOOM, and we observed that selections of different demonstrations often largely impact the performance. Yet the performances do suffer a sharp loss if we are using irrelevant instruction in the instruction-tuned model. We also discovered that different models might favor different instructions for different languages, for example, in Swahili, four-shot BLOOM favors the first instruction, while mT0 favors the fourth instruction.

D.5 Qualitative Results for Generation Tasks

Table 25 shows some qualitative results of ChatGPT English ICL and Target TCL on XLSum and TyDiQA. Given English instruction, ChatGPT often generates summaries in English, rather than in the article language. On the other hand, such cross-lingual generation behaviors don't occur in QA tasks, and the model's predictions with Target ICL and English ICL exhibit high overlap with each other. We hypothesize that ChatGPT's cross-lingual summarization behavior can be related to their private training corpus, and future work can further investigate this issue.

D.6 Results of English-centric LMs

Table 26 shows BUFFET-Light performance on four more recent and English-centric LMs whose checkpoints are publicly available: Llama1-7B, Llama2-7B, Llama2-7B-Chat and Mistral 7B.

[12] https://huggingface.co/bigscience/bloom

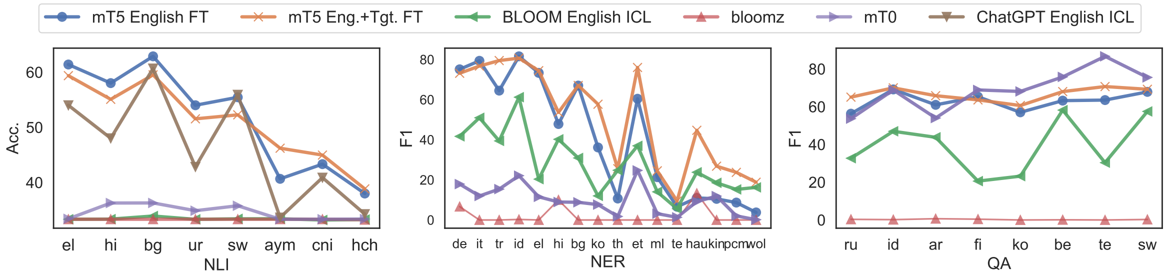

Figure 7: **Model performance across three tasks, NLI, NER, and QA, displayed for various languages.** The languages are sorted based on token availability in mC4, with the left side representing high-resource languages. All methods show performance deterioration in lower-resource languages (right side), with larger drops in ENGLISH-ICL methods. Additional fine-tuning in target languages is more effective in less-represented languages.

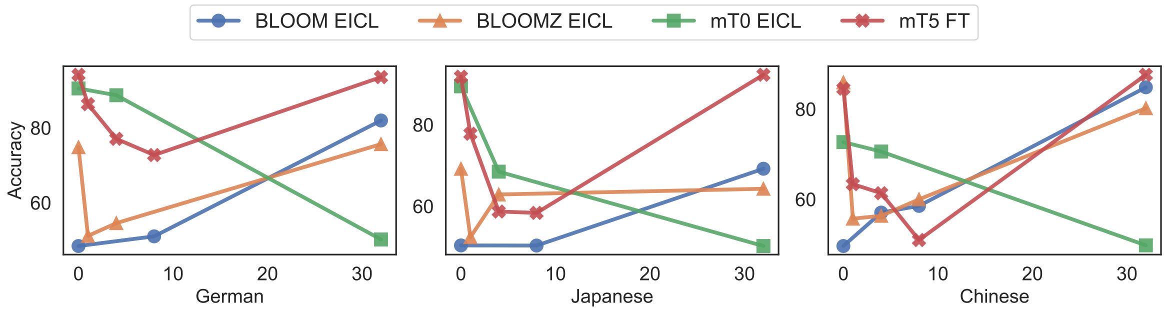

Figure 8: **Effects of demonstrations on Amazon Review.** The x-axis indicates the number of training instances used during the transfer.

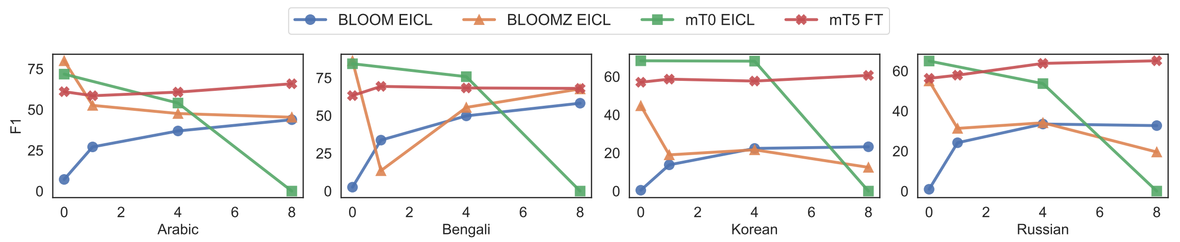

Figure 9: **Effects of demonstrations on TYDIQA.** The x-axis indicates the number of training instances used during the transfer.

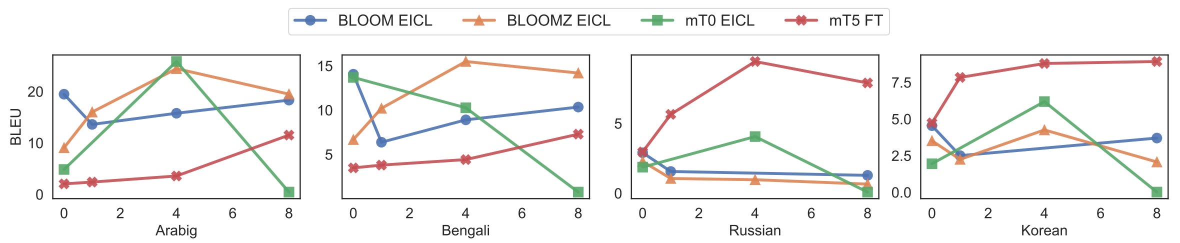

Figure 10: **Effects of demonstrations on TYDIQA-QG.** The x-axis indicates the number of training instances used during the transfer.

Despite large-scale multilingual pre-training or instruction-tuning as in prior work (Muennighoff et al., 2023), Mistral, Llama2 (pre-trained and chat) demonstrate strong performance while Llama1 performance is largely limited. Prior work has shown that a small amount of pre-training data often results in strong multilingual capabilities of LLMs that are primarily trained in English pre-training (Blevins and Zettlemoyer, 2022b; Briakou et al., 2023). On the other hand, we found that

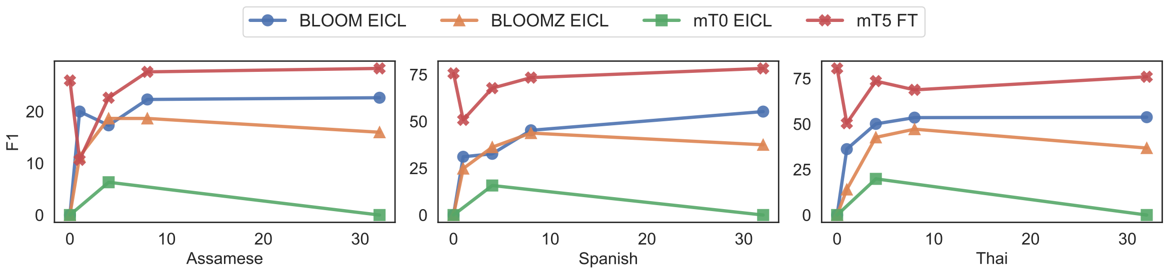

Figure 11: **Effects of demonstrations on WIKIANN.** The x-axis indicates the number of training instances used during the transfer.

Instruction	BLOOM ru	BLOOM sw	BLOOM fi	mT0 ru	mT0 sw	
This task is about reading the given passage and constructing a question about the information present in the passage. Construct a question in such a way that (i) it is unambiguous, (ii) it is answered from the passage, (iii) its answer is unique (iv) its answer is a continuous text span from the paragraph. Avoid creating questions that (i) can be answered correctly without actually understanding the paragraph and (ii) uses the same words or phrases given in the passage.	8.7	4.3	**10.8**	5.0	5.3	3.1
Could you generate a question in `lang` whose answer is as provided based on the following context?	9.1	4.3	9.5	6.5	7.5	8.7
Could you generate a `lang` question whose answer is as provided based on the following context?	9.2	4.3	9.3	6.6	7.4	8.4
Generate a `lang` question whose answer is as provided based on the following context. **9.3**	**4.4**	9.1	**7.1**	**7.7**	**9.0**	
In this task, you are given a review from Amazon. Your task is to generate a rating for the product on a scale of 1-5 based on the review. The rating means -2: extremely poor, 1: poor, 0: neutral, 1: good, 2: extremely good, 3: mixed and -3: no sentiment expressed	9.2	**4.4**	10.4	0.4	0.1	0.4

Table 24: The performance (in BLEU score) for different instructions for TYDIQA-QG. `lang` stands for the full language name, e.g., Arabic. We modify the instructions from BUFFET, PromptSource (Bach et al., 2022), and SuperNaturalInstructions (Wang et al., 2022b).

Target	Gold.	Pred.
XLSUM-Indonesian	Manajer Chelsea Jose Mourinho memberi waktu 48 jam untuk Wayne Rooney memutuskan apakah dia ingin meninggalkan Manchester United dan pindah ke Stamford Bridge. (Google Translate: Chelsea manager Jose Mourinho has given Wayne Rooney 48 hours to decide whether he wants to leave Manchester United and move to Stamford Bridge.)	ENGLISH ICL: Manchester United striker Wayne Rooney has been given 48 hours to resolve speculation over his future, with Chelsea expressing interest in the player. [en] TARGET ICL:Mourinho memberi ultimatum kepada Wayne Rooney untuk mengakhiri spekulasi kepindahannya dalam waktu 48 jam, sementara MU dikabarkan sedang mengincar Gareth Bale. [id]
XLSUM-Turkish	Suriye'nin Rakka vilayetine bağlı Tel Abyad kent merkezinde IŞİD ile YPG güçleri arasında şiddetli çatışmalar sürerken, Akçakale sınırı sığınmacılara yeniden açıldı. IŞİD'e karşı savaşan YPG güçleri de Tal Abyad'a girdi.	ENGLISH ICL: YPG forces have cut off a vital ISIS supply route from Tel Abyad to Raqqa, and have taken control of two neighborhoods in Tel Abyad. [en] TARGET ICL: YPGńin ele geçirdiği, çatışmaların devam ettiği ve sivillerin tehlikede olduğu" bilgisini verdi. [tr]
TYDIQA-Swahili	'Sehemu ya chakula pamoja na wanga, protini na vitamin	ENGLISH ICL: sehemu za chakula pamoja na wanga, protini na vitamini.[sw] TARGET ICL: Sehemu za chakula pamoja na wanga, protini na vitamini. [sw]

Table 25: ChatGPT outputs for XLSUM and TYDIQA Output. [lang] indicate the language codes of the output text languages.

model	avg. class	avg. gen	NLI	SA	PAWS	QA	QG	SUM	NER
Llama 1B (7B)	28.1	4.2	12.9	48.1	27.4	24.4	6.4	2.1	20.2
Llama 2B (7B)	41.6	6.4	32.3	67.4	44.6	36.7	9.6	3.2	26.8
Llama2 Chat (7B)	44.1	6.4	**35.0**	70.8	45.9	43.1	11.3	1.4	**28.0**
Mistral (7B)	**45.2**	**7.4**	33.3	**77.4**	**46.0**	**51.8**	**12.4**	**2.4**	24.0

Table 26: Results of Llama1, Llama2, Llama2-chat and Mistral on BUFFET-light. All models are 7 billion parameters.

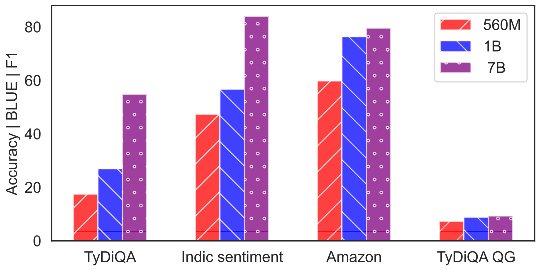

Figure 12: **Model scaling experimental results.** We conduct experiments on four sub-tasks and use three BLOOM models, BLOOM-560M, 1B, and 7B.

those models often show limited performance in languages that are less represented in such pretraining corpora (e.g., AMERICASNLI, INDIC SENTIMENT). This result suggests the importance of understanding how much multilingual training data needs to be included during pre-training to make an LM learn the target languages, which remains unclear.

D.7 Effect of Model scaling

Figure 12 shows the effects of model scaling on BLOOM.

E Discussions for Future Directions

Built upon findings from our extensive BUFFET experiments, we suggest the following opportunities for future research on few-shot cross-lingual transfer learning:

Improve multilingual instruction tuning. Instruction tuning causes certain models, such as mT0, to become overly specialized to specific ICL formats. Although these models demonstrate impressive zero-shot performance, they struggle in unfamiliar settings such as few-shot ICL or tasks in less common formats (e.g., NER). It is important to develop multilingual instruction-following models capable of effectively utilizing both instructions and demonstrations, potentially by drawing inspiration from recent work on better instruction-tuning in English (Chung et al., 2022; Min et al., 2022a).

Overcome data scarcity using LLMs. Our evaluation reveals that smaller task-specific models (with intermediate training in English) outperform ChatGPT on discriminative tasks with strict output formats. In contrast, ChatGPT outperforms fine-tuned models on generation, consistent with recent work (Goyal et al., 2022). This impressive generation capacity has prompted investigations into generating training instances from LLMs; these predominantly focus on English (Wang et al., 2022a; Honovich et al., 2022) with some preliminary work on generating multilingual task data (Agrawal et al., 2022). Further work in this direction offers a promising solution to obtaining more annotated data for under-represented languages.

Understand transfer dynamics in cross-lingual in-context learning. The impact of various instructions and demonstrations has been extensively examined in the context of English in-context learning, highlighting critical concerns (Lu et al., 2022; Min et al., 2022b) and motivating methods (Su et al., 2022). BUFFET will inspire and assist in further research into the relationship between language and instruction/demonstration for cross-lingual in-context learning.

Fairness beyond languages: underrepresented variants, dialects, and cross-cultural NLP. Typologically distinct and low-resource languages are often excluded from the cross-lingual benchmarks used to assess cross-lingual transfer capabilities in LLMs. Our evaluation with BUFFET demonstrates that even the most powerful LLMs still perform poorly on less-represented languages. The most competitive instruction-tuned models, ChatGPT or mT0, show significant performance declines when it comes to indigenous languages, reaching a level akin to a random baseline. We advocate for conducting more studies that include under-represented languages and their dialects, as emphasized in previous works (Aji et al., 2022; Kakwani et al., 2020), particularly when evaluating massively multilingual models.