Analyzing and Adapting Large Language Models for Few-Shot Multilingual NLU: Are We There Yet?

Evgeniia Razumovskaia **Ivan Vulić** **Anna Korhonen**
Language Technology Lab, University of Cambridge, UK
{er563, iv250, alk23}@cam.ac.uk

Abstract

Supervised fine-tuning (SFT), supervised instruction tuning (SIT), and in-context learning (ICL) are three alternative, *de facto* standard approaches to few-shot learning. ICL has gained popularity recently with the advent of LLMs due to its versatile simplicity and sample efficiency. Prior research has conducted only limited investigation into how these approaches work for *multilingual* few-shot learning, and the focus so far has been mostly on their performance. In this work, we present an extensive and systematic comparison of the three approaches, testing them on a variety of high- and low-resource languages over five different NLU tasks, and a myriad of language and domain setups. Importantly, performance is only one aspect of the comparison, where we also analyze and discuss the approaches through the optics of their computational, inference and financial costs. Some of the highlighted findings concern an excellent trade-off between performance and resource requirements/cost for SIT. We further analyze the impact of target language adaptation of pretrained LLMs and find that the standard adaptation approaches can (superficially) improve target language generation capabilities, but language understanding elicited through ICL does not improve accordingly and remains limited, especially for low-resource languages.

1 Introduction and Motivation

Recent advances in data-efficient, few-shot learning have been crucial for increasing and promoting language inclusiveness of NLP technology (Devlin et al., 2019; Conneau et al., 2020; ImaniGooghari et al., 2023), substantially lowering the dataset size-related 'entry point' for a new language. This was made possible by pretrained language models which can generalize to a new task or language from the knowledge stored in their parameters complemented with only a handful of in-task data.

The standard approaches for such few-shot adaptations are **1)** *Supervised Fine-Tuning (SFT)*, which also subsumes more recent **2)** *Supervised Instruction-Tuning (SIT)*, and **3)** *In-Context Learning (ICL)*. SFT and SIT use knowledge in pretrained model parameters for initialization and then adapt the parameters to a *language-task* combination via *supervised training* on available, even if scarce, resources. Importantly, they yield a model specialized for a single *language-task* combination and can get increasingly better at the task if a larger training dataset becomes available. ICL, in contrast, uses a single versatile model 'as is' to complete any task, without any parameter adaptation or fine-tuning. Instead, it is adapted via prompting: Given an explanation of a task (i.e., *instruction*) and a set of 'training' examples (i.e., annotated *demonstrations*), the model is tasked to generate the label for every input (Radford et al., 2019). Due to the model's context size, ICL performance is capped by the model's pretraining and the demonstrations that fit into the input context.

Existing generative models used for ICL (termed *Large Language Models, or LLMs* henceforth) are usually pretrained in an English-centric manner, with the vast majority of the pretraining corpus in English and only limited coverage of other languages (Sitaram et al., 2023), even with 'accidentally encountered' bilingual and translation data (Briakou et al., 2023). As a result, current LLMs are very far from serving the world's languages equally: While demonstrating impressive ICL results in English (Mishra et al., 2022), they still face difficulties when transferring to other languages (Winata et al., 2021; Tanwar et al., 2023), especially low-resource ones (Ojo et al., 2023). In contrast, a number of encoder and encoder-decoder models, such as XLM-R (Conneau et al., 2020) or mT5 (Xue et al., 2021),

Transactions of the Association for Computational Linguistics, vol. 13, pp. 1096–1120, 2025. https://doi.org/10.1162/tacl.a.33
Action Editor: Asli Celikyilmaz. Submission batch: 5/2024; Revision batch: 9/2025; Published 9/2025.

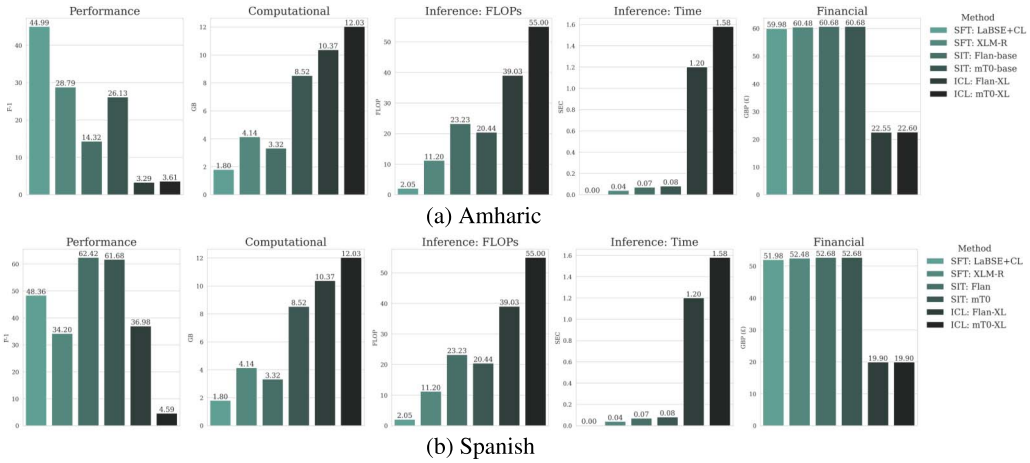


Figure 1: Comparison of practical aspects of different learning paradigms (§3.1) in the intent detection task from Multi3NLU++ (Moghe et al., 2023), with exactly the same data setup, for Amharic and Spanish. In-context learning (ICL) has low performance and high inference and computational cost while being comparatively inexpensive. Supervised fine-tuning (SFT) and supervised instruction-tuning (SIT), on the other hand, have a larger financial cost but they are much more efficient in terms of inference aspects and computational resources while also performing much better both for Amharic as a representative low-resource language (1a) and Spanish as a high-resource language (1b).

used for initialization in SFT are pretrained with much wider language coverage[1] (termed *multilingually Pretrained Language Models, or mPLMs;* Conneau et al., 2020; ImaniGooghari et al., 2023). This property enables sample-efficient transfer and adaptation of natural language understanding (NLU) models to a much larger array of languages (Ansell et al., 2021) than what is currently supported by ICL-based LLMs.

While SFT, SIT, and ICL are comparable approaches for few-shot multilingual NLU that can 'consume' the same type of task-annotated data, there has been little attention drawn to which of the techniques works better *in practice*. Therefore, this paper aims to delve deeper into understanding and analyzing a variety of factors which critically impact effective use of either from a more *practical point of view*. Our first aim is to provide answers to the following question:

(Q1) Given the same annotated examples, which of the approaches is better ***in practice****?*

In particular, the sometimes vague term 'practice' in our work comprises the following crucial aspects: **1)** sample efficiency (i.e., *'data cost'*); **2)** computational requirements (*'computational cost'*); **3)** latency (*'inference cost'*); and **4)** overall financial or *'economic' cost*. Prior work has been mainly focused on benchmarking ICL on subgroups of languages (Ojo et al., 2023) and only considered and optimized task performance of the models as the ultimate comparison criterion. In contrast, our work presents an extensive analysis studying cross-lingual capabilities of SFT and ICL both on high and low-resource languages, considering not only the task performance but also the above listed practical aspects, as illustrated in Figure 1.

Our analysis focuses on token- and sentence-level classification tasks commonly used for evaluation of NLU capabilities of the models (for details see §4.1). Classification tasks are especially demonstrative in our systematic comparison, as their well-defined ontologies and automated metrics enable direct comparison of the different modeling paradigms. Further, the classification-based tasks (or tasks which can be cast into classification) are frequently used in modern-day NLU systems (e.g., extracting named entities), which additionally motivates our choice.

Furthermore, prior work has demonstrated the effectiveness of parameter-efficient fine-tuning (PEFT) to improve the models' cross-task capabilities and to promote aspects of their generation abilities (e.g., open-domain chat; Dettmers et al., 2023). In this work, we also analyze how language

[1]For instance, XLM-R (Conneau et al., 2020), mBERT (Devlin et al., 2019), LaBSE (Feng et al., 2022), and mT5 (Xue et al., 2021) cover ∼100 languages at pretraining (albeit with different pretraining data amounts), while Glot500 (ImaniGooghari et al., 2023) covers up to 500 languages.

adaptation of LLMs 'beyond English' impacts their NLU and NLG performance in a target language. This gives rise to another core research question:

(Q2) Given the benefits of ICL as a learning paradigm (but its inferior performance in comparison to SFT), could we use the standard ***adaptation strategies*** *to improve LLMs' generation and understanding capabilities in other languages?*

To our knowledge, this is the first work analyzing how multilingual NLU capabilities of ICL with LLMs are effected by their target language adaptations, as well as studying the trade-offs for NLG.

Contributions. **1)** Related to Q1, we conduct a comprehensive analysis of ICL versus SFT and SIT paradigms in the context of multilingual few-shot adaptation, with the focus on multiple practical aspects and cost. Our analyses show that not only the SFT and SIT approaches with smaller models lead to improved task performance but also they remain more data-, computation-, inference-effective than ICL with general-purpose LLMs. **2)** Related to Q2, we investigate the effectiveness of target language adaptation, adopted from the work on mPLMs, for ICL with LLMs. The main finding is that language adaptation leads to superficially improved generation capabilities in the target language with only limited improvements on the actual tasks, calling for further research that will mitigate the large language gap in LLM development and deployment between English and other languages.

2 Related Work

Instruction-Tuning LLMs aim to increase their cross-task generalization capabilities. Instruction tuning is in essence an SFT technique where the input includes textual description of the task, demonstrations, and user input queries while the output is the desirable model output for a given task in text form. Through inclusion of task descriptions into the input, at inference time the model becomes capable of completing tasks unseen during training when provided with task description (Sanh et al., 2022; Chung et al., 2022, interalia). Instruction tuning has become a standard approach to turn an LLM into a model with general capabilities to perform any task, given the instructions, off-the-shelf (Wei et al., 2022a; Mishra et al., 2022).

Extending LLMs to Other Languages. Although there is a growing trend to make NLP systems more linguistically inclusive (Bender, 2011; Doddapaneni et al., 2021), widely used generative LLMs remain predominantly English. For instance, pretraining data of Llama-2 and PaLM consists of 90% and 82% English text, respectively (Touvron et al., 2023; Sitaram et al., 2023), which substantially hinders their capabilities in languages other than English (Ojo et al., 2023). In an attempt to equate the models' performance across languages, there is an increasing interest in extending their multilingual capabilities. A wide range of techniques including continued pretraining (Cui et al., 2023), or using self-instruction (Wei et al., 2023) or vocabulary extension (Zhao et al., 2024), have been applied. Due to wide adoption of ICL, another line of work focuses on improving cross-lingual instruction following capabilities via parameter-efficient multilingual instruction tuning (Li et al., 2023b), multilingual pretraining (Shliazhko et al., 2022), and injection of several multilingual examples in fine-tuning (Shaham et al., 2024). The methods show gains in various aspects of a model's target language generation capabilities, while providing no systematic empirical comparisons to prove that improved NLG necessarily correlates with stronger NLU performance via ICL.

These works provide initial insights into LLM processing for languages other than English. Interestingly, the success of mPLMs in cross-lingual transfer has always been attributed to their massively multilingual pretraining while, at first sight, LLMs seem to operate differently: They perform surprisingly well while having only a small percentage of multilingual text in their pretraining corpora (Blevins and Zettlemoyer, 2022). At the same time, little to no work has studied multilingual performance of these models in direct comparison with standard mPLMs, and even more so the practical aspects such as memory requirements or latency.

3 On Learning and Practical Aspects

We analyze three established *learning paradigms* for few-shot learning in monolingual and multilingual setups, which are compared across four *practical aspects*: data cost, computational cost,

inference cost, and financial cost. We now outline each learning paradigm and practical aspect.

3.1 A Brief Overview of Learning Paradigms

Let $\mathcal{D} = (x_1, y_1), \ldots, (x_N, y_N)$ denote a training dataset where x_i is the model input, y_i is the label annotation, and N is the number of training examples, and let $\mathcal{M}$ refer to a pretrained language model (LLM or mPLM).

Supervised Fine-Tuning (SFT). $\mathcal{M}$ is adapted to a task or a language (or both) by fine-tuning its parameters on $\mathcal{D}$ and minimising a loss function. Note that here we use SFT in its narrower sense,[2] to refer to 'standard' fine-tuning where an encoder-based model (such as mBERT) or encoder-decoder model (e.g., mT5) is tuned directly for the target task (Devlin et al., 2019; Wei et al., 2022a). At inference, the fine-tuned model $\mathcal{M}'$ is then used.

In-Context Learning (ICL). Unlike with SFT, the parameters of $\mathcal{M}$ stay fixed and the model is treated as a 'black box'. ICL relies on generative capabilities of general-purpose LLMs (Brown et al., 2020; Han et al., 2023). The model is adapted to a task by conditioning it on task instructions and in-context examples (*demonstrations*). Each demonstration included into a prompt consists of an input x and ground-truth annotated label y. In other words, the demonstrations are an alternative way to use the data available in $\mathcal{D}$. Then, $\mathcal{M}$ is expected to generate the label for the test input usually included at the end of the prompt. While in SFT the model parameters are adapted to a target task, with ICL the model is expected to learn the task by analogy, via the provided task description combined with demonstrations. ICL is an extremely popular, versatile approach since a single LLM can serve multiple tasks (or in theory, any task) without any further fine-tuning, but in this work we put its multilingual NLU ability and corresponding practical value to scrutiny.

Supervised Instruction-style Tuning (SIT). To unlock the full potential of ICL, sufficiently large language models need to be used (Wei et al., 2022b), drastically raising the computational overhead at inference in comparison with SFT. SIT thus presents the middle ground between the two. Here, one fine-tunes small(er) instruction-based models to specific tasks. While SFT fine-tunes the model directly on annotated data $\mathcal{D}$, SIT extends each input in $\mathcal{D}$ with task-specific instructions leveraging model's instruction-following capabilities (Wei et al., 2022a) obtained during pretraining. SIT typically does not include demonstrations into input, although including them there is also possible (Min et al., 2022; Chen et al., 2022), typically with small to negligible performance gains in few-shot setups but increased computational cost (Li et al., 2023a). For simplicity, we experiment only with the SIT variant *without* any demonstrations.

3.2 Practical Aspects

We consider practical costs of the *'full cycle'* of model development—from data collection costs to inference cost, and aim to associate those costs with the learning paradigms described in §3.1.

Data Cost. One key limiting factor for model adaptation to new task-language (or even finer-grained task-language-domain) combinations is the costly and complex data collection process, especially for low-resource languages and specialized domains. Therefore, it is crucial to develop methods which can efficiently generalize from a small number of annotated examples. In §5, we analyze this data cost, that is, sample efficiency as the relationship between the number of training examples and task performance.

Computational Cost. The memory requirements of LLMs keep growing proportionally to the number of their parameters. Deploying such a model to the users means that one needs to have access to and support costly infrastructure with large vRAM (Aminabadi et al., 2022; Alizadeh et al., 2023). Here, we analyze memory requirements of each learning paradigm both for model storage and training, where applicable, and how they correlate with the target task performance.

Inference Cost. Latency, or time needed by the model to complete the prediction (Huyen, 2022, Chapter 1), has the largest impact on user-facing applications such as task-oriented dialog. To make the system usable, it is critical to strike a balance between strong performance and low latency. We thus analyze the inference cost in two ways as: **1)** wall-clock inference time, aiming to directly

[2]This is in contrast to SFT as used in the context of Large Language Models where SFT refers to fine-tuning a generalist LLM as part of its post-training procedure to improve its performance on a specific task.

approximate (relative) latency of different models; and **2)** inference FLOPs, a hardware-independent metric to compare inference complexity.

Financial Cost. Each of the aspects above contributes to the overall cost of each model's life cycle. As financial resources are usually limited, we also aim to (roughly) estimate the overall financial expenditure needed for each learning paradigm, including data collection, GPU, and inference costs.

4 Experimental Setup

We focus on the comparison between SFT, SIT, and ICL in few-shot multilingual and cross-lingual setups, aiming to make the comparison as fair as possible across languages, learning paradigms, and models, and targeting the following setups:

In-Language Generalization. We evaluate the model's ability to generalize on new examples in the same language in which fine-tuning examples or demonstrations were provided to the model.

Cross-Language Generalization. We use a model trained in one language to perform the task in another one, where the transfer typically proceeds from a high-resource language to a low-resource one. In our experiments, we assume the typical transfer direction with English as the (high-resource) source language.

In-Domain and Cross-Domain Generalization. For many NLU tasks (e.g., for task-oriented dialog) it is common to consider transferring the systems between different domains, e.g., from the *flight booking* to the *restaurant booking* domain. If a model can be transferred across domains, it means that it has in-depth understanding of the classes used in the respective domain definitions/ontologies.

4.1 Evaluation Tasks and Datasets

The main focus of the analyses, revolving around Q1 and Q2 from §1, is on NLU tasks for task-oriented dialog as one widely used and established practical application of NLP technology, due to multiple reasons. **1)** Dialog is a user-facing application where computational and memory requirements, data collection cost, inference latency and other practical concerns of the model development cycle are of ultimate importance. **2)** Dialog NLU tasks provide well-defined ontologies and evaluation setups, with evaluation benchmarks that comprise comparable and semantically aligned training and test data across multiple languages, including high- and low-resource ones (Moghe et al., 2023; Hu et al., 2023a), and multiple domains. **3)** In contrast to standard 'non-dialog' NLU tasks, dialog NLU datasets are unlikely to have been seen and 'absorbed' by LLMs during their pretraining, which avoids test data leakage (Balloccu et al., 2024; Sainz et al., 2023).

Dialog-oriented evaluation is conducted on the tasks of intent detection (ID) and value extraction (VE). ID aims to classify user's utterance into a set of intent classes predefined in the domain ontology. The aim of VE is to identify the presence of ontology-related domain-specific slot-value pairs in a given sentence. Here, we use the Multi3NLU++ dataset (Moghe et al., 2023), covering English (Casanueva et al., 2022) and the following 4 languages: Amharic (AM), Marathi (MR), Spanish (ES), and Turkish (TR). The datas et also spans two different domains: BANKING and HOTELS, with a partial overlap in intent classes and slots. For both tasks we report micro-F1 scores.[3]

To verify that our findings extend beyond only dialog-related NLU tasks, we also evaluate on three other standard NLU classification tasks: NLI, NER, and reasoning. We evaluate NLI performance on the XNLI dataset (Conneau et al., 2018) which provides training and evaluation data in 14 languages, while we focus on a subset of 3: ES, TR, and RU, and report accuracy as the evaluation metric. NER is evaluated on MasakhaNER (Adelani et al., 2021), a high-quality NER dataset providing training and evaluation data in 10 African languages of which we focus on 5: AM, IG, PCM, SW, and YO. Lastly, we evaluate the reasoning capabilities of the models using the XCOPA dataset (Ponti et al., 2020) which asks the model to determine which of two provided sentences causally follows from a premise. XCOPA provides translation of English COPA (Roemmele et al., 2011) into 11 geographically and linguistically diverse languages, of which we focus of

[3] We also note that **1)** each utterance in Multi3NLU++ may have multiple intents; ID is thus a multi-label classification task. **2)** Further, for VE, we consider the slot value as correctly labeled only if it exactly matches the gold value. Finally, **3)** as in prior work (Casanueva et al., 2022; Moghe et al., 2023), the cross-domain performance for the two tasks is evaluated only on the intents and slots shared across domains. We refer to the original Multi3NLU++ work for further details.

Dataset		LANGS	# TEST EX.	# CLASSES
Multi3NLU++	ID	AM, EN, ES, MR, TR	300	62
	VE			17
XNLI	NLI	EN, RU, TR, ES	5,010	3
MasakhaNER	NER	AM, IG, PCM, SW, YO	500	4
XCOPA	COPA	ET, ID, IT, TA, TR	500	2

Table 1: Summary of evaluation datasets and tasks.

five: ET, ID, IT, TA, and TR.[4] Additional information on the evaluation datasets is provided in Table 1. Notably, all three of the tasks are included in the XTREME-R benchmark, a standard benchmark to evaluate multilingual language understanding capabilities of the models (Ruder et al., 2021).

We chose to focus on classification-based NLU tasks so that SFT with mPLMs (such as XLM-R) could be fairly compared with SIT and ICL with instruction-tuned models (such as Flan) and LLMs (such as GPT-3.5).

Cross-Language Parallel Few-Shot Setup. To ensure fair comparisons of all learning paradigms across languages, we make use of the multi-parallel nature of the datasets. For each *language-domain* combination in Multi3NLU++ (or just *language* in XNLI and MasakhaNER) we sample 300 test examples.[5] We also randomly sample training sets consisting of {30, 50, 100, 500, 1,000} examples which are kept exactly the same across all languages to ensure the content in training data does not coincidentally favor one of the languages.[6]

SFT Evaluation. We use two standard models, *XLM-R-Base* (Conneau et al., 2020) and *LaBSE* (Feng et al., 2022). Following prior work (Moghe et al., 2023), we train only task-specific classifiers on top of the fixed LaBSE sentence encoder. We refer to this approach, applied only to sentence-level tasks (ID and NLI), as *LaBSE+CL*.[7]

ICL Evaluation. We evaluate the following models: open-sourced Flan-T5-XL, mT0-XL, Llama-2-7B, and closed-source GPT-3.5.[8] We include further details on model openness in Appendix I. Flan-T5-XL (3B parameters; Chung et al., 2022), mT0-XL (3.7B; Muennighoff et al., 2022), and GPT-3.5 (Achiam et al., 2023) are massively instruction-tuned models. While Flan-T5 was pretrained mostly in English and several high-resource languages, mT0-XL offers a more comprehensive and balanced multilingual pretraining set.

The inputs for ICL were designed in a cross-lingual manner, where the task descriptions and context were in English while the few-shot examples and the sentence to be analyzed were provided in the target language. This follows the recommendations from prior work where it was empirically verified that English instructions led to stronger results than in-language instructions (Shi et al., 2022; Lin et al., 2022). For each task we design the instructions (i) to match the instructions in pretraining as closely as possible, while (ii) yielding reasonable output when tested on several validation examples.[9] Note that, given a fixed input context of each model, the number of demonstrations to be used for ICL is limited: for all the models in our comparison it is less than 30, and 30 is the lowest number of training samples we use in SFT.

SIT Evaluation. Here, we include individual per-class questions into instructions: This design (i) was previously shown to result in much improved SIT performance (Fuisz et al., 2022; Razumovskaia et al., 2023), while (ii) it also fits into the model input context for tasks with a large number of classes. We rely on the same instructions as with ICL. We experiment with

[4]As XCOPA only contains validation and test splits, we use subsamples of validation splits for training in each language and test set for evaluation

[5]We conduct the sampling step due to a large number of experimental runs; preliminary experiments with full test sets indicated exactly the same relative trends, but with much increased computational cost and time overheads. While sampling, we ensured that each intent and slot in the domain ontology occurred at least twice in the test set.

[6]To ensure reproducibility, unique ids of the examples in the training and test splits are available at: `https://github.com/evgeniiaraz/few_shots_supplementary.git`.

[7]For NLI as a single-label classification task, the *softmax* output activation function is used. In contrast, for ID we use *sigmoid* and consider all intents where the *sigmoid* activation is larger than a predefined threshold value θ. Following prior work, we set $\theta = 0.3$.

[8]We use *GPT-3.5-turbo-instruct* due to its instruction-following capabilities proven in prior work (Ye et al., 2023).

[9]For reproducibility, we share the full templates for the instructions for all languages and tasks in the Github repository.

two models: (i) (mostly English pretrained) *Flan-T5-Base* (250M parameters) and multilingually oriented *mT0-Base* (580M).[10]

On the Fairness of Comparisons. The aim of this study is to compare three standard paradigms for supervised few-shot learning. A limitation of the comparison is that, given the chosen setup, the base models in each paradigm have to be different. SFT assumes that an encoder model is fine-tuned on a task while SIT and ICL require a decoder model with language generation capabilities. Further, given the computational constraints, smaller generative models have to be used for SIT than for ICL, as SIT (in contrast to ICL) requires fine-tuning. Thus, using exactly the same models across the paradigms is impossible, especially given the focus on comparison of the approaches *in practice*. At the same time, the practical aspects of the comparison should remain comparable.

5 (Q1) Results and Discussion: Learning Paradigms and Practical Aspects

We first delve into comparisons revolving around Q1 (§1). The main results across different setups, models, training data sizes, and learning paradigms are summarized in Figure 2, while full (numerical) results are provided in Appendix C. We now zoom into discussions originating from the results.

Data Efficiency. One of the core reasons to use ICL is its inherent data/sample efficiency. Comparing the supervised methods against ICL, we observe that the former reach or overcome the performance of ICL with all tested open-source models (Flan-T5-XL, mT0-XL and Llama-2-7B), even when fine-tuned with mere 30 in-task examples, while they outperform GPT-3.5 with 50 or 100 in-task examples. These findings hold across all evaluation tasks and setups.

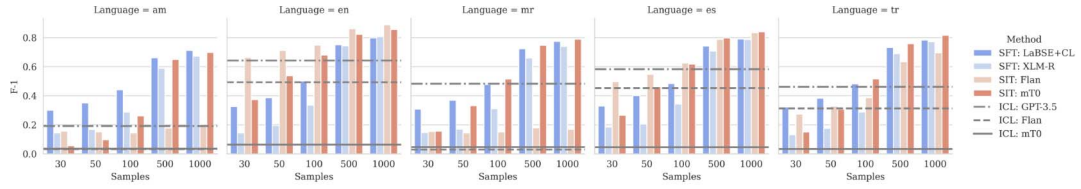
At the same time, the results also reveal several key differences between tasks and languages. Comparing the trends for ID and VE (cf. Figures 2a and 2d): For sentence classification tasks where the outputs are language-independent, the improvements of SFT over ICL are less pronounced. For instance, the gains over GPT-3.5 with 100 training examples for MR and ES are 3.15 and 4.17 F1 points,[11] respectively. In contrast, the gains over ICL for value extraction as a more language-specific task are very large, 19.8 and 25.28 for MR and ES, respectively, when comparing the best-performing SFT method with ICL. Moreover, for VE, the gaps with ICL are considerable even with 30 training examples are used (3.4 and 7.5 F_1 points for MR and ES).

For AM, a supervised model surpasses GPT-3.5 performance already with 30 training examples, while for ES 50 or 100 training examples are required, depending on the setup. For high-resource languages (EN, ES) SIT-based Flan-Base with 30 examples performs consistently better than ICL with GPT-3.5 for ID and VE across the setups. We hypothesize that high performance of SIT-based Flan is caused by i) its instruction-following capabilities, and ii) large-scale English pretraining which is helpful for both few-shot in-language generalization and cross-lingual transfer from English to Spanish, a linguistically close language. For low-resource languages (AM, MR), we notice a different tendency across the domain setups: LaBSE+CL and SIT-based mT0 show the highest performance for ID and VE, respectively. Finally, for the token-level NER task the trend is slightly different: SFT-based XLM-R outperforms all ICL-based models by a large margin. This again proves the importance of multilingual pretraining for the model to generalize to unseen or 'less seen' languages.

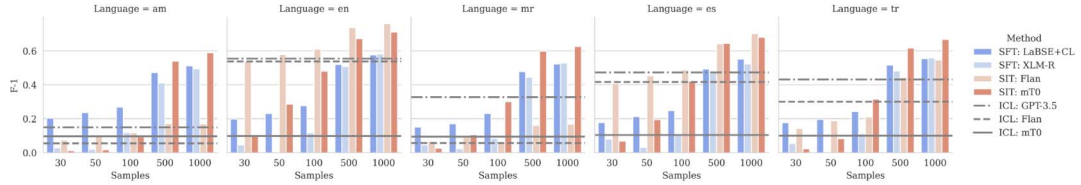
In-Domain vs Cross-Domain Evaluation. The comparison in cross-domain setups consistently shows that SIT outperforms SFT and ICL (see Figure 2b), corroborating findings from prior work on English (Razumovskaia et al., 2023). We speculate that the success of SIT in cross-domain setups stems from the model's ability to follow instructions obtained during pretraining and the ability to extract class semantics from instructions obtained during fine-tuning. The best-performing instruction-tuned LLM, however, depends on the target language: On low-resource languages multilingually pretrained models such as mT0 perform consistently better than English-pretrained models such as Flan, while we observe reversed trends for high-resource languages (ES).

[10] SFT and SIT training hyperparameters are in Appendix B.

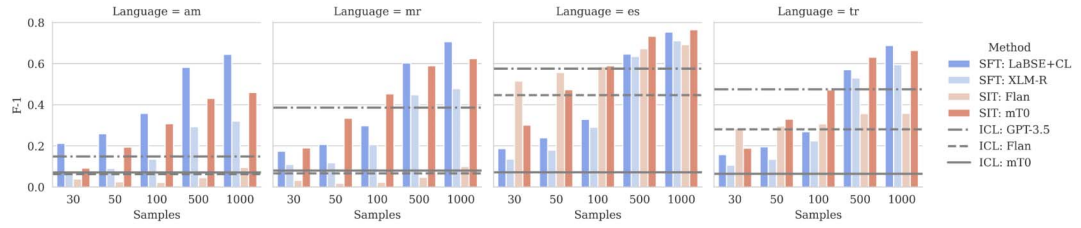
[11] The cited numbers are for in-language in-domain setups; the trends are the same in the other setups.



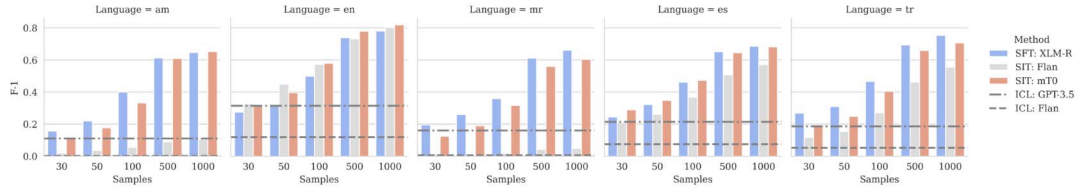
(a) ID: In-Language In-Domain



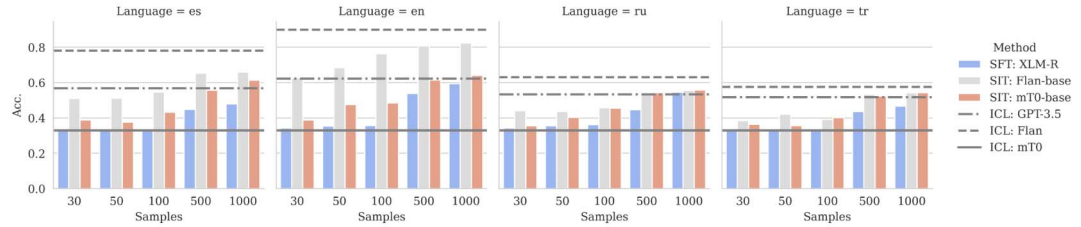
(b) ID: In-Language Cross-Domain



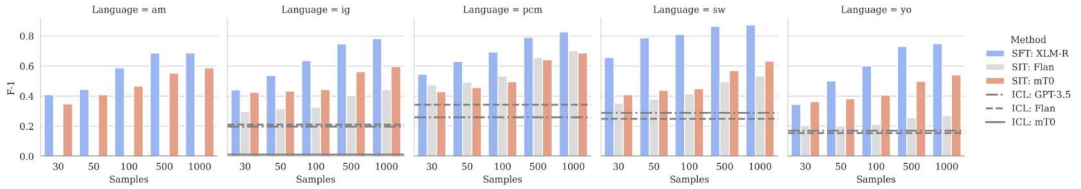
(c) ID: Cross-Lingual In-Domain



(d) VE: In-Language In-Domain



(e) XNLI



(f) NER

Figure 2: Intent detection, value extraction, NLI and NER results for the six languages in our evaluation. This performance is in line with other prior work (Hu et al., 2023b). We exclude Llama-2 results as its performance was 0.0 across all tasks. Results for VE in other setups are provided in Appendix D.

Cross-Lingual Zero-Shot Transfer. Figure 2c presents the results for zero-shot transfer in in-domain setups: Performance across languages for all approaches is substantially lower than the performance in English. Further, as expected, performance on low-resource languages is considerably lower than on high resource languages.

The results in Table 2 show that for ICL, unlike for SFT (Lauscher et al., 2020), providing the model with data examples in the target language does not always improve the final performance. Target language demonstrations seem to be helpful to the models which have strong instruction-following capabilities and are familiar

Model		AM	MR	ES	TR
GPT-3.5	ICL_t	19.19	48.28	63.25	59.27
	ICL_{en}	14.87	38.67	57.64	47.50
Flan	ICL_t	3.28	3.02	45.26	31.36
	ICL_{en}	6.33	6.68	44.70	28.04
mT0	ICL_t	3.61	4.71	4.60	3.36
	ICL_{en}	7.19	7.99	7.23	6.45

Table 2: ICL results on the ID task with English (ICL_{en}) or target language (ICL_t) demonstrations.

with the target language (e.g., ES performance of Flan and GPT-3.5).

SIT vs ICL. In general, the results indicate that SIT consistently leads to better results than ICL in few-shot setups. Smaller SIT-based models can even outperform ICL with GPT3.5 when 100+ task examples are available. Due to its sample efficiency and strong performance in cross-domain and cross-lingual setups, SIT also mitigates the issue of using a separate model for each *task-language* or *task-language-domain* combination.

Seen vs Unseen Tasks. Results for XNLI (Figure 2e) and XCOPA (Appendix E) demonstrate the effectiveness of ICL with Flan and GPT-3.5, as the tasks were 'seen' during instruction-tuning,[12] with different patterns observed for the tasks with unseen data (i.e., ID and VE). This discrepancy is especially pronounced for high-resource languages. Results in Figure 2f on MasakhaNER demonstrate a different trend: The ICL models perform considerably worse than multilingually pretrained models fine-tuned with SFT and SIT, although NER was included into pretraining of instruction-tuned models (Muennighoff et al., 2022; Chung et al., 2022). We assume that this discrepancy is due to the languages being low-resource. The model for NER is required to generate the spans in the target language for named entity-value pairs, which might not be feasible due to lack of vocabulary and lexical knowledge of the target language (cf., 0.0 ICL results for AM in Figure 2f).

[12]XNLI is based on English MultiNLI (Williams et al., 2018) while XCOPA is based on English COPA (Roemmele et al., 2011), which have been used for instruction-training of many LLMs (Muennighoff et al., 2022; Chung et al., 2022).

Paradigm: Model	MEMORY COST		INFERENCE COST	
	Max (GB)	*Storage (GB)*	*Time (s)*	*FLOPs* (10^9)
SFT: LaBSE+CL	1.80	1.80	0.004	2.05
SFT: XLM-R	4.14	1.04	0.004	11.18
SIT: Flan-Base	3.32	0.85	0.059	23.23
SIT: mT0-Base	5.82	1.45	0.081	20.44
ICL: Flan-XL	10.37	10.37	1.58	39.03
ICL: mT0-XL	12.03	12.03	1.20	55.00
ICL: GPT-3.5	–	–	2.78	–

Table 3: Memory and inference costs of SFT, SIT, and ICL, measured on the ID test dataset. MEMORY *Max* is the peak fine-tuning or storage memory cost of each approach. MEMORY *Storage* refers to storage requirements per models. For all models but GPT-3.5 the measurements were conducted on a single RTX-3090 GPU. For GPT-3.5, we report average response time per example.

5.1 Analyses of Practical Costs

Given that the results above indicate that the two supervised paradigms (SFT and SIT) substantially outscore ICL in terms of task performance in general, we now focus on comparing them in terms of practical aspects. The summary is presented in Figure 1 (see §1) and Table 3.

Computational (Memory) Cost. Besides improved task performance, another advantage of SFT and SIT is that the underlying high-performing models are much smaller and thus have lower memory requirements. The largest memory cost for ICL is storing the model's parameters at inference time, while for SFT and SIT it is the memory requirements during fine-tuning. We rely on the HuggingFace Memory Calculator to establish vRAM needed for training and inference of every paradigm. We measure the memory requirements in full precision and using the AdamW optimizer (Loshchilov and Hutter, 2019), when applicable.[13] The results indicate that models used for ICL have more than $2\times$ higher memory needs than mT0 and Flan-T5-base used for SIT in our experiments. Another angle to memory requirements is the storage cost, i.e., how much memory is needed to store a given model ('as is' for ICL and after fine-tuning for SFT and SIT). Table 3 suggests that storage cost for models used for ICL is at least $4\times$ higher than the models used in SIT and SFT.[14]

[13]Closed-source GPT-3.5 is excluded from the comparison.

[14]We note that analyses of computational cost reduction techniques such as quantization or model pruning (Gholami et al., 2022; Liang et al., 2021) lie beyond the scope of this work, as we chose to focus on standard approaches.

Inference Cost. Beyond average wall-clock inference time per test example, we also report the number of FLOPs measured using fvcore, also averaged per test example. As expected, the inference cost scales with the size of the underlying model, with inference time of GPT-3.5 being more than $3\times$ higher than that of SIT-ed models, and inference FLOPs of open-source ICL models being $2.5\times$ higher than for their smaller SIT-ed counterparts. While SFT methods demonstrate even higher inference efficiency, we observe that SIT has the best trade-off between inference cost and performance.

Financial Cost. Having demonstrated considerably higher inference costs of ICL, we also consider overall economic costs required for SFT, SIT, and ICL. Target language data annotation accounts for the largest expenditure in the process. We calculate the annotation cost based on Moghe et al. (2023). ICL consumes up to 30 annotated examples, with total costs of £15.9 and £18.6 for high-resource and low-resource languages, respectively, where the annotations are obtained both for ID and VE. In the VE task, SIT-based methods reach or surpass the ICL performance of the strongest model (GPT-3.5) already with 20 extra examples (i.e., with 50+ training examples for supervised learning), which only adds £11 or £10 to the overall cost for low- and high-resource languages, respectively.

Given the larger inference time and computational costs of the ICL, the total ongoing costs are likely to be larger than the one-time additional annotation budget. To put the numbers in context, the inference cost of 300 test examples with GPT-3.5 was between £3 and £4 for high- and low-resource languages at the time of experimentation, respectively. Put simply, the actual cost balance should take into account also the expected, tentative number of inference calls.

Further, while increasing the input context length of LLMs is an active research area (Press et al., 2022; Rubin and Berant, 2023, *among others*), many models relying on the ICL paradigm are still constrained by input context length, and there is evidence that ICL performance even becomes quickly saturated with the addition of extra in-context examples (Chen et al., 2023; Li et al., 2023a) and that the long context is not leveraged adequately (Liu et al., 2023). On the contrary, unlike with ICL, our experiments demonstrate that performance of SFT and SIT improves with more annotated examples (both in-language and cross-lingually, see Figure 2). Data annotation of 100 training examples raises the annotation cost by an average of £37.5 while increasing the ID and VE performance by an average of 15 F-1 points over ICL with GPT-3.5.

Comparison to Multilingual LLMs. Since the focus of this work is on multilingual few-shot NLU tasks, here we further extend our study by examining ICL with two state-of-the-art models which were pretrained and developed, having multilingual capabilities in their prime focus, unlike previously discussed Llama-2 and Flan models. In particular, we study Aya-Expanse-8B (Dang et al., 2024) and Llama-3.1-8B (Dubey et al., 2024). Beyond increased coverage of languages in pretraining, Aya and Llama-3 models also have a longer context window than previously evaluated models. This entails that more training examples can be fit in-context. Our comparison will focus on the in-language in-domain intent detection.[15]

The results in Table 4 show that the multilingually pretrained and instruction-tuned models demonstrate higher average performance than the models considered previously, especially for high-resource languages (EN, ES), while still lagging behind for the low-resource ones. For instance, both models show performance lower than that of LaBSE+CL on AM. These results also come at a cost of other practical aspects. Both Aya and Llama-3.1 have higher memory requirements than any of the previously considered models. Further, on-par performance even for high-resource languages can be achieved with SIT-ed models performing inference $\approx 240\times$ faster and with $\approx 4\times$ less FLOPs. These results further stress the necessity to consider aspects beyond performance when picking the model for multilingual few-shot NLU.

Further Discussion and Summary. Beyond the quantitative comparisons, we now discuss the practical aspects from a qualitative angle. SFT and SIT might be perceived as more computationally costly than ICL as they include two stages: (i) task fine-tuning/specialization, and (ii) inference with the fine-tuned models. In contrast, ICL only

[15] The same experiments were conducted for value extraction. However, the models largely ignored the output format.

Paradigm: Model	ID					MEMORY COST		INFERENCE COST		FINANCIAL COST
	AM	EN	MR	ES	TR	*Max (GB)*	*Storage (GB)*	*Time (s)*	*FLOPs* (10^9)	GBP (£)
SFT: LaBSE+CL	29.98	32.53	30.80	32.95	32.24	1.80	1.80	0.01	2.05	56.68
SFT: XLM-R	14.34	14.35	14.57	18.57	13.17	4.14	1.04	0.01	11.18	57.14
SIT: Flan-Base	15.60	66.25	15.42	49.69	27.27	3.32	0.85	0.06	23.23	57.34
ICL: Flan-XL	3.28	49.27	3.02	45.26	31.36	10.37	10.37	1.58	39.03	57.34
ICL: GPT-3.5	19.19	64.22	48.28	58.25	46.12	–	–	2.78	–	21.48
ICL: Aya	5.74	52.28	36.29	52.03	51.92	14.95	14.95	14.42	77.87	57.34
ICL: Llama-3.1	26.57	66.57	57.03	61.68	59.49	18.45	18.45	13.89	93.81	57.34

Table 4: Performance of 'multilingual-first' LLMs (Aya and Llama-3.1) versus other models in our evaluation. We use F-1 score $\times$ 100 as a metric. For financial cost we provide an average cost per language to collect the data. We focus on non-English data collection.

involves the latter step, and does not get specialized to a particular task. At the same time, due to the model size the computational cost of each inference call for ICL is typically as much as two orders of magnitude larger than that of SFT- and SIT-specialized models; this might balance out the cost for training (Liu et al., 2022). Another trade-off to consider when choosing between ICL and supervised methods is the necessity for optimizing the 'training' recipe. Put simply, supervised methods require hyperparameter tuning which might increase the overall computational cost, while ICL requires careful prompt tuning, which in the majority of cases means increased manual labor (i.e., the so-called prompt engineering), time, and computational cost.

Lastly, ICL is often perceived as an easier and more versatile paradigm as a single model can be used for any *task-domain-language* combination. As our experiments above demonstrate, if only one or a small subset of tasks is in the focus, a highly specialized small(er) model tuned with SFT or SIT would be the best trade-off option performance-wise and cost-wise. At the same time, if the incentive is to have a general-purpose model that serves many tasks at once (but likely with reduced performance in each one of them), ICL would be a more suitable option. In sum, we believe that the needs of each application and the trade-offs between all the practical costs need to considered when selecting a method for real-life applications.

6 (Q2) Results and Discussion: Target Language Adaptation of LLMs

§5 indicates that ICL is consistently inferior to the two supervised learning paradigms, SFT and SIT, not only in terms of task performance but also concerning computational and inference costs. At the same time, as discussed previously, ICL relies on a single model and is thus appealing when extending a system to a large number of language-task combinations.

Prior work on 'decoder-only' LLMs demonstrated the effectiveness of parameter-efficient fine-tuning (PEFT) to improve their cross-task generalization capabilities (Page-Caccia et al., 2024), whereas PEFT is a standard approach for cross-lingual adaptation of 'encoder-only' and 'encoder-decoder' models such as XLM-R or mT5 (Conneau et al., 2020; Xue et al., 2021). In this work, we explore whether such language-specific PEFT-style adaptation can improve ICL and generation capabilities of LLMs in languages other than English. We focus on Llama-2-7B as our base model, as: (i) it is a 'decoder-only' model that (ii) has been trained as the 'English-first' model, with almost 90% of its pretraining data in English; and (iii) it displayed the lowest performance in our experiments in §4 while being the largest model in our evaluation.

Language Adaptation Setup. We use QLoRA (Dettmers et al., 2023) as a standard PEFT-based language adaptation technique. QLoRA performs two modifications to the base LLM. The model is first quantized to reduce the memory requirements and then a low-rank adapter (Hu et al., 2022) is trained on top of the quantized model. In our experiments the adapter is tuned on the target language data, aiming to boost the target language capabilities of the underlying LLM.

For the adaptation experiments, we focus on three languages: Spanish, Turkish, and Marathi. The adapter for each language is trained on the

respective portion of mC4 (Xue et al., 2021). Hyper-parameters are set following Dettmers et al. (2023), with exact details available in Appendix F.

6.1 Generation after Language Adaptation?

First, we assess whether target language adaptation boosts generation capabilities of the LLM in the target language. To this end, we use the Bactrian-X dataset (Li et al., 2023b), a multilingual instruction dataset containing parallel instruction-response pairs in 52 languages. For our evaluation, we use a subset of 100 randomly sampled examples ensuring the same parallel examples across the three languages in our evaluation (ES, TR, MR).

Generation Evaluation: True or Superficial Improvements? We focus on the three aspects of generation capabilities: (i) whether the model outputs text in the same language as expected by the input (i.e., *I/O language agreement*, similarly to Kew et al., 2023); (ii) *naturalness* of the generated text; and (iii) *lexical overlap* between gold responses and generation outputs. I/O language agreement involves doing automated language identification of the generated text and establishing whether it corresponds to the input text. For this purpose, we use the current state-of-the-art language identification model, GlotLID-500 (Kargaran et al., 2023). We evaluate naturalness via MAUVE (Pillutla et al., 2021) which measures the distributional gap between human written and generated texts. For lexical overlap, we report ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002).

Figure 3 shows consistent gains of generation capabilities over the three evaluation aspects after target language adaptation, with especially large improvements for Marathi as the lowest-resource language. The I/O agreement scores suggest that through language adaptation LLM's abilities to generate text in the target languages are reinforced.

However, those standard metrics still do not fully capture the potential usefulness of generated output and (improved) generation capabilities. We thus also conduct human-based evaluation for Spanish for *naturalness* and *usefulness*. Each output is evaluated on a simple 3-point Likert-like scale.[16] Interestingly, the average naturalness score raises from 1.4 to 2.2 after language adaptation while usefulness only increases from 1.4 to

[16]Annotation instructions are provided in Appendix H.

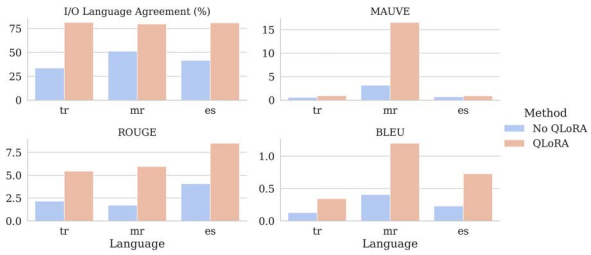


Figure 3: Generation evaluation after target language adaptation (Llama-2).

1.6. In practice, this means that even after language adaptation the model is still far from being useful for the target language speakers. This finding corroborates preliminary observations of Kew et al. (2023) that the English-centric models can learn to generate text in a target language comparatively easily, but useful instruction-following capabilities still remain largely out of reach. Put simply, while generated text in the target language becomes more fluent, its coherence and relevance remain limited.

6.2 NLU after Language Adaptation?

Given only superficial improvements in generation capabilities, we now assess whether the ICL capabilities improve for NLU tasks. For brevity, we focus on XNLI as the least complex NLU task. Even for XNLI, we observe only a negligible non-significant improvement from the average accuracy score of 30.5 to 30.7.[17] Performance is in fact below random (33%), supporting the observations from §5 that resource-efficient ICL requires both multilingual pretraining and instruction tuning. From qualitative assessment of the outputs, we notice that the models struggle to follow the task description and instructions, and often do not adhere to output formatting requirements.

Massively Multilingually Adapted LLMs in NLU tasks. The results above suggest that *direct* target language adaptation of 'English-first' LLMs such as Llama-2 does not yield any benefits to ICL performance in NLU tasks. Next, we study whether *massively multilingual* adaptation of 'English-first' models, as done in very

[17]Per-language scores are provided in Appendix G. Similar relative trends have been observed in preliminary experiments on another, more complex NLU task: Belebele (Bandarkar et al., 2023), where the results are on-par or lower than the random choice baseline, as well as in more complex NLU tasks from our evaluation in §5.

Model	AM	EN	MR	ES	TR
ICL: Llama-2	0.0	0.0	0.0	0.0	0.0
ICL: MaLA-500	1.0	3.01	0.0	1.0	3.01
ICL: GPT-3.5	19.19	64.22	48.28	58.25	46.12
SIT: mT0	26.13	68.00	51.44	61.69	51.60
Tr-Test + ICL: GPT-3.5	32.25	–	48.49	47.95	48.60

Table 5: ICL results on the ID task in the in-language in-domain setup.

recent work, can improve their ICL capabilities in different languages. We evaluate the MaLa-500 model (Lin et al., 2024) which was adapted for 534 languages using the Glot-500-c dataset (ImaniGooghari et al., 2023), and is also based on Llama-2.[18] We focus on the ID task to evaluate MaLa's ICL performance in the (easiest) in-language in-domain setup, with results summarized in Table 5. They reveal that, while adaptation gives marginal improvements for ICL, the performance is still extremely low and lags substantially behind GPT-3.5 performance and SIT with mT0-Base with 100 training examples. A comparison with the *translate-test* baseline shows that while translate-test benefits low-resource AM, it is still outperformed by SIT on all other languages. Overall, the scores suggest that more work is needed on multilingual adaptation of LLMs to unlock their ICL capabilities in other languages, and current adaptation strategies do not yield models with competitive (nor even useful at all) NLU.

7 Conclusions and Future Work

This work has provided a series of in-depth analyses of and has discussed multilingual capabilities of three learning paradigms, two supervised ones versus in-context learning (ICL), with the focus on few-shot learning and NLU tasks. Besides task performance, the focus of the analyses has also been on multiple practical aspects (e.g., data efficiency, memory requirements, inference latency).

In general, our work has affirmed the importance of multilingual pretraining and the potential of supervised training on top of smaller LLMs in multilingual setups. Future work should invest more effort into the creation of massively multilingual- and multitask-pretrained LLMs with higher language coverage. Further, our analysis in §6 calls for new and improved language adaptation methods atop the LLMs. Finally, we hope that our work will also steer researchers and practitioners in multilingual NLP towards a more holistic view of model properties that combines task performance with practical aspects during the model development cycle (e.g., size, latency, memory, data collection cost). While such practical aspects and model capabilities are constant moving targets and change over time, our work stresses the importance of multi-dimensional analyses reaching beyond task performance for the adoption of particular learning paradigms in particular multilingual NLU tasks.

[18]MaLA-500 was adapted using: (i) LoRA-based parameter adaptation; (ii) vocabulary extension to accommodate for languages that do not use the Latin script.

Acknowledgments

We thank the TACL reviewers and action editor for their helpful feedback. The work has been in part supported by a Huawei research donation to the Language Technology Lab at the University of Cambridge. It has also been supported by the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 EQUATE (the UK government's funding guarantee for ERC advanced grants) awarded to Anna Korhonen at the University of Cambridge. The work of Ivan Vulić has been supported in part by a Royal Society University Research Fellowship *'Inclusive and Sustainable Language Technology for a Truly Multilingual World'* (no. 221137).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131. https://doi.org/10.1162/tacl_a_00416

Keivan Alizadeh, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C. Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. 2023. Llm in a flash: Efficient large language model inference with limited memory. *ArXiv preprint*, abs /2312.11514.

Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. 2022. Deepspeed-inference: Enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. IEEE. https://doi.org/10.1109/SC41404.2022.00051

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.410

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *ArXiv preprint*, abs/2402.03927. https://doi.org/10.18653/v1/2024.eacl-long.5

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The Belebele Benchmark: A parallel reading comprehension dataset in 122 language variants. *ArXiv preprint*, abs/2308.16884. https://doi.org/10.18653/v1/2024.acl-long.44

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6. https://doi.org/10.33011/lilt.v6i.1239

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explain the cross-lingual capabilities of english pretrained models. *ArXiv preprint*, abs/2204.08110. https://doi.org/10.18653/v1/2022.emnlp-main.233

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.524

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.154

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159. Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.745

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via

language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.53

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1269

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *ArXiv preprint*, abs/2304.08177.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv preprint*, abs/2305.14314.

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

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models. *ArXiv preprint*, abs/2107.00676.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz,

Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, and Junteng Jia. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Gabor Fuisz, Ivan Vulic, Samuel Gibbons, Iñigo Casanueva, and Paweł Budzianowski. 2022. Improved and efficient conversational slot labeling through question answering. *ArXiv preprint*, abs/2204.02123.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2022. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC. https://doi.org/10.1201/9781003162810-13

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.708

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In the *Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023a. Multi 3 woz: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415. https://doi.org/10.1162/tacl_a_00609

Songbo Hu, Han Zhou, Moy Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen, and Ivan Vulić. 2023b. A systematic study of performance disparities in multilingual task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6851, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.422

Chip Huyen. 2022. *Designing Machine Learning Systems*. O'Reilly Media, Inc.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.61

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational*

Linguistics: EMNLP 2023, pages 6155–6218, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.410

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? *ArXiv preprint*, abs/2312.12683. https://doi.org/10.18653/v1/2024.findings-emnlp.766

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.363

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vuli. 2023a. On task performance and model calibration with supervised and self-ensembled in-context learning.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *ArXiv preprint*, abs/2305.15011.

Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. 2021. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461:370–403. https://doi.org/10.1016/j.neucom.2021.07.045

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *ArXiv preprint*, abs/2401.13303.

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

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *ArXiv preprint*, abs/2307.03172.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.201

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.244

Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen,

and Alexandra Birch. 2023. Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.230

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *ArXiv preprint*, abs/2211.01786. https://doi.org/10.18653/v1/2023.acl-long.891

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I. Adelani. 2023. How good are large language models on african languages? *ArXiv preprint*, abs/2311.07978.

Lucas Page-Caccia, Edoardo Maria Ponti, Zhan Su, Matheus Pereira, Nicolas Le Roux, and Alessandro Sordoni. 2024. Multi-head adapter routing for cross-task generalization. *Advances in Neural Information Processing Systems*, 36.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, pages 4816–4828.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.185

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net.

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

Evgeniia Razumovskaia, Goran Glavaš, Anna Korhonen, and Ivan Vulić. 2023. Sqatin: Supervised instruction tuning meets question answering for improved dialogue nlu. *ArXiv preprint*, abs/2311.09502.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Ohad Rubin and Jonathan Berant. 2023. Long-range language modeling with self-retrieval. *ArXiv preprint*, abs/2306.13421.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.125

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.802

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787. https://doi.org/10.18653/v1/2023.findings-emnlp.722

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

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *ArXiv preprint*, abs/2401.01854. https://doi.org/10.18653/v1/2024.findings-acl.136

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. 2022. Language models are multilingual chain-of-thought reasoners. In the *Eleventh International Conference on Learning Representations*.

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

Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja, and Kalika Bali. 2023. Everything you need to know about multilingual llms: Towards fair, performant and reliable models for languages of the world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 21–26. https://doi.org/10.18653/v1/2023.acl-tutorials.3

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.346

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan

Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *ArXiv preprint*, abs/2307.06018.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.1

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

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *ArXiv preprint*, abs/2303.10420.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *ArXiv preprint*, abs/2401.01055.

A Instructions Used for Different Tasks

Task	Instruction Text
ID	The aim is to understand user's intent from the utterance. Include all applicable options exactly as they are provided. Separate the classes by hyphen. If no options are applicable, return an empty string. Options: - to deny something - to ask about savings account <list of all options applicable in the domain> Utterance: {demonstration1} Intents: {intent1}-{intent2}-{intent3} <all in-context demonstrations> Utterance: {test example} Intents:
VE	The aim is to extract slot values from the user utterance. Use $$ as delimiter between slot-value pairs. The slot values should be tagged as: - amount_of_money: specific amount of money - adults: number of adults <list of all slot classes applicable in a given domain> Utterance: {demonstration1} Values: {slot_class1}:{value1}$${slot_class2}:{value2} <all in-context demonstrations> Utterance: {test example} Values:
NLI	The aim is to determine whether the premise entails, contradicts or is neutral with respect to the hypothesis. Only output the label. Premise: {premise-demonstration1} Hypothesis: {hypothesis-demonstration2} Does the premise entail, contradict, is neutral to the hypothesis? Answer: {label1} <all in-context demonstrations> Premise: {premise-test} Hypothesis: {hypothesis-test} Does the premise entail, contradict or is neutral to the hypothesis? Answer:

Table 6: Text of instructions used in ICL. ID and NLI: instructions were adapted from Flan (Chung et al., 2022) with intent descriptions from the ontology provided with NLU++ (Casanueva et al., 2022). VE: instructions were adapted from XTREME-UP (Ruder et al., 2023) and Ojo et al. (2023).

B Finetuning Hyperparameters

Hyperparameter	Value
LaBSE+CL: *dim*	512
LaBSE+CL: *non-linearity*	tanh
Batch size	32
Learning rate	2e-5
Weight Decay	0.1
Evaluation Frequency	500 steps
Max Epochs	500
Optimiser	AdamW

Table 7: Fine-tuning hyperparameters used across supervised training experiments. The rest of the parameters were set to the default values in HuggingFace Transformers.

C Full Experimental Results

Full numerical results for the experiments are provided in Tables 8–12.

		In-domain results					Cross-domain results				
	Samples	AM	EN	MR	ES	TR	AM	EN	MR	ES	TR
SFT: LaBSE+CL	30	0.2998	0.3253	0.308	0.3295	0.3224	0.2041	0.1978	0.1507	0.178	0.1773
	50	0.3502	0.3863	0.3679	0.4014	0.3826	0.2362	0.2305	0.1700	0.2123	0.1962
	100	0.4409	0.5007	0.4773	0.4836	0.4815	0.2688	0.2774	0.2304	0.2485	0.2432
	500	0.6606	0.7509	0.7235	0.7412	0.7328	0.4728	0.5204	0.4780	0.4936	0.5169
	1000	0.7116	0.7978	0.7736	0.7900	0.7825	0.5119	0.5759	0.5225	0.5516	0.5539
SFT: XLM-R	30	0.1434	0.1435	0.1457	0.1857	0.1317	0.0284	0.0456	0.0463	0.0799	0.0544
	50	0.1676	0.1946	0.1696	0.2060	0.1750	0.0200	0.0042	0.0226	0.0318	0.0021
	100	0.2879	0.3363	0.3115	0.3421	0.288	0.1196	0.1176	0.0848	0.1009	0.1123
	500	0.5882	0.742	0.6592	0.7075	0.6898	0.4107	0.5076	0.4441	0.4694	0.4806
	1000	0.6715	0.8066	0.7391	0.7862	0.7721	0.4943	0.5822	0.5282	0.5223	0.5584
SIT: Flan-T5-Base	30	0.156	0.6625	0.1542	0.4969	0.2727	0.0750	0.5369	0.0677	0.4081	0.1419
	50	0.1520	0.7110	0.1434	0.5468	0.3282	0.0999	0.5794	0.0904	0.4535	0.1888
	100	0.1432	0.7483	0.1501	0.6242	0.3865	0.1168	0.6103	0.0638	0.4882	0.2094
	500	0.1769	0.8601	0.1780	0.7873	0.6341	0.1716	0.7355	0.1620	0.6421	0.4434
	1000	0.2040	0.8877	0.1680	0.8333	0.6957	0.1699	0.7602	0.1679	0.7017	0.5453
SIT: mT0-Base	30	0.0560	0.3735	0.1558	0.2646	0.1505	0.0125	0.097	0.0269	0.0684	0.0228
	50	0.0962	0.5375	0.3309	0.4614	0.3068	0.0169	0.2868	0.1059	0.1957	0.0825
	100	0.2613	0.68	0.5142	0.6169	0.516	0.0936	0.4795	0.3009	0.4209	0.3151
	500	0.6488	0.8222	0.7466	0.7978	0.7579	0.5394	0.6711	0.5985	0.6441	0.6163
	1000	0.6980	0.8559	0.7889	0.8393	0.8157	0.5892	0.7113	0.6264	0.6798	0.6681
ICL: Flan-T5-XL		0.0328	0.4927	0.0302	0.4526	0.3136	0.0554	0.5375	0.0581	0.4176	0.3012
ICL: mT0-XL		0.0361	0.064	0.0471	0.0460	0.0336	0.0969	0.0989	0.0947	0.1049	0.1006
ICL: GPT-3.5		0.1919	0.6422	0.4828	0.5825	0.4612	0.1501	0.5552	0.3283	0.4728	0.4320

Table 8: Per-language intent detection results for in-domain and cross-domain setups.

		In-domain results					Cross-domain results				
	Samples	AM	EN	MR	ES	TR	AM	EN	MR	ES	TR
SFT: XLM-R	30	0.1566	0.2748	0.1953	0.2444	0.2681	0.0275	0.0336	0.0369	0.03	0.0232
	50	0.2199	0.3234	0.2603	0.3221	0.3098	0.049	0.0543	0.0329	0.0462	0.031
	100	0.4003	0.4991	0.3598	0.4615	0.4665	0.0362	0.0279	0.0229	0.0469	0.0072
	500	0.6130	0.7392	0.6118	0.6508	0.6937	0.036	0.06	0.05	0.103	0.098
	1000	0.6468	0.7801	0.6614	0.6855	0.7539	0.05	0.087	0.083	0.137	0.117
SIT: Flan-T5-Base	30	0.0191	0.327	0.0156	0.2091	0.1174	0.0019	0.2514	0.0018	0.07	0.0511
	50	0.0362	0.4486	0.0083	0.2627	0.1537	0.009	0.3006	0.0031	0.1089	0.0887
	100	0.0555	0.5728	0.0198	0.3678	0.2705	0.0103	0.4043	0.005	0.1987	0.1395
	500	0.0896	0.7314	0.042	0.5073	0.4615	0.0313	0.5956	0.0141	0.3689	0.3272
	1000	0.1055	0.8041	0.0484	0.5707	0.5552	0.0445	0.6244	0.014	0.3975	0.3577
SIT: mT0-Base	30	0.1193	0.3182	0.1246	0.2893	0.1886	0.0433	0.1615	0.0688	0.1162	0.1011
	50	0.1774	0.3954	0.1899	0.347	0.2488	0.0511	0.2153	0.1118	0.1679	0.1371
	100	0.3313	0.58	0.3167	0.4723	0.4055	0.0972	0.3345	0.14	0.212	0.1927
	500	0.6093	0.779	0.5596	0.6458	0.6596	0.3864	0.5838	0.3562	0.4535	0.4417
	1000	0.6521	0.8193	0.6036	0.6814	0.7065	0.4304	0.6528	0.4337	0.4572	0.5097
ICL: Flan-T5-XL		0.0022	0.1187	0.0063	0.0756	0.0524	0	0.04	0.02	0.006	0.02
ICL: mT0-XL		0	0	0	0	0	0	0	0	0	0
ICL: GPT-3.5		0.1105	0.3148	0.1611	0.2142	0.1862	–	–	–	–	–

Table 9: Per-language value extraction results for in-domain and cross-domain setups.

	Samples	AM	MR	ES	TR
	30	0.2123	0.1742	0.1869	0.1574
	50	0.2595	0.2076	0.2392	0.1959
SFT: LaBSE+CL	100	0.3586	0.2977	0.3287	0.2693
	500	0.5821	0.6027	0.6465	0.5704
	1000	0.6448	0.7064	0.7528	0.6887
	30	0.0738	0.1094	0.1364	0.1068
	50	0.0905	0.1177	0.1794	0.1348
SFT: XLM-R	100	0.1355	0.2058	0.2908	0.2247
	500	0.2931	0.4485	0.6347	0.5303
	1000	0.3206	0.4778	0.7105	0.5947
	30	0.0385	0.032	0.5155	0.2849
	50	0.0253	0.0186	0.5567	0.2959
SIT: Flan-T5-Base	100	0.0218	0.023	0.5858	0.3065
	500	0.0449	0.0469	0.672	0.3575
	1000	0.0947	0.0993	0.692	0.3585
	30	0.0923	0.1901	0.3008	0.1889
	50	0.194	0.3336	0.4732	0.3298
SIT: mT0-Base	100	0.3078	0.4534	0.5898	0.4716
	500	0.4305	0.5893	0.7329	0.6316
	1000	0.4602	0.6246	0.765	0.6638
ICL: Flan-T5-XL		0.0633	0.0667	0.447	0.2804
ICL: mT0-XL		0.0719	0.0799	0.0723	0.0645
ICL: GPT-3.5		0.1487	0.3867	0.5764	0.475

Table 10: Per-language intent detection results for the cross-lingual setup for EN → TGT transfer.

	Samples	AM	MR	ES	TR
	30	0.0867	0.11	0.2004	0.1531
	50	0.1206	0.1623	0.2511	0.213
SFT: XLM-R	100	0.1748	0.2151	0.3278	0.2847
	500	0.3174	0.405	0.5335	0.4647
	1000	0.3569	0.4256	0.5769	0.5177
	30	0.0567	0.0299	0.2122	0.1032
	50	0.0481	0.0265	0.238	0.1236
SIT: Flan-T5-Base	100	0.0531	0.0316	0.2885	0.1411
	500	0.0659	0.0422	0.3549	0.1706
	1000	0.0806	0.0435	0.3837	0.1788
	30	0.1045	0.0872	0.2471	0.1728
	50	0.1123	0.0994	0.2937	0.1895
SIT: mT0-Base	100	0.1447	0.1556	0.3939	0.2382
	500	0.1855	0.2141	0.5445	0.3293
	1000	0.2049	0.2327	0.5585	0.3394
ICL: Flan-T5-XL		0	0.001	0.03	0
ICL: mT0		0	0	0	0
ICL: GPT-3.5		–	–	–	–

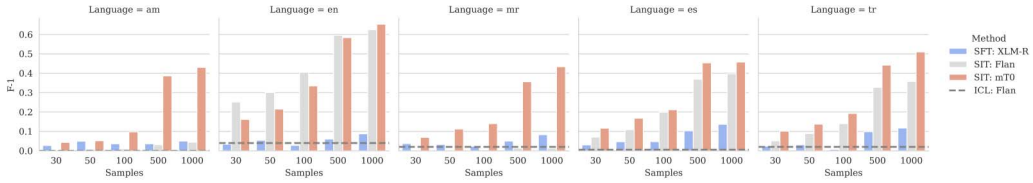
Table 11: Per-language value extraction results for the cross-lingual setup for EN → TGT transfer.

	Samples	RU	ES	EN	TR
	30	0.3426	0.333	0.3422	0.3373
	50	0.3554	0.3357	0.3538	0.3285
SFT: XLM-R	100	0.362	0.3321	0.3578	0.3345
	500	0.4458	0.4482	0.5378	0.4357
	1000	0.545	0.4795	0.5936	0.4663
	30	0.4399	0.5094	0.6255	0.385
	50	0.4355	0.5106	0.6844	0.4212
SIT: Flan-T5-Base	100	0.4577	0.5451	0.7623	0.3914
	500	0.5409	0.6537	0.807	0.5339
	1000	0.5549	0.6593	0.8238	0.5423
	30	0.3545	0.388	0.3878	0.3641
	50	0.4034	0.3756	0.4762	0.3567
SIT: mT0-Base	100	0.4551	0.4321	0.4848	0.402
	500	0.5427	0.5563	0.615	0.5238
	1000	0.5581	0.6132	0.6403	0.5425
ICL: Flan-T5-XL		0.6307	0.7808	0.8994	0.5758
ICL: mT0		0.33	0.33	0.33	0.33
ICL: GPT-3.5		0.5333	0.5683	0.6224	0.518

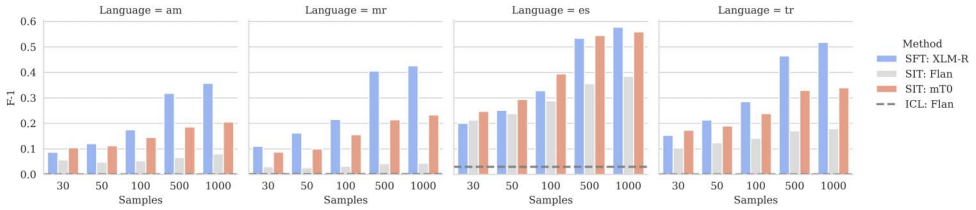
Table 12: Per-language natural language inference results on XNLI.

D Further Value Extraction Results

Further detailed results for value extraction are presented in Figure 4.



(a) VE: Cross-Domain In-Language



(b) VE: Cross-Lingual In-Domain

Figure 4: Value extraction results for Amharic (AM), English (EN), Marathi (MR), Spanish (ES), and Turkish (TR) for two setups: a) cross-domain in-language; and b) cross-lingual in-domain performance. We exclude ICL-mT0 XL from the plot, as it had 0.0 performance on VE task in these setups. Qualitative analysis of the outputs of ICL-mT0 showed that the outputs neither adhered to the slot-value pair formatting nor included the right values.

E XCOPA Results

We present the results on XCOPA in Figure 5.

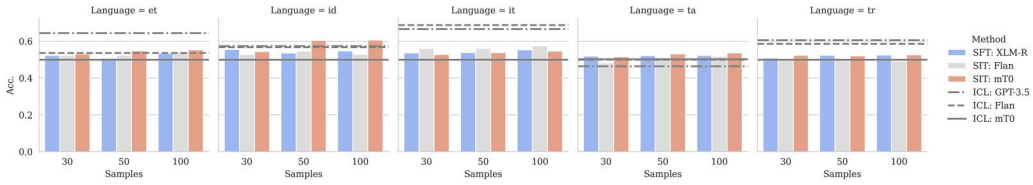


Figure 5: XCOPA results for Estonian (ET), Indonesian (ID), Italian (IT), Tamil (TA) and Turkish (TR). Importantly, XCOPA only includes validation set (of 100 examples) and test set (of 500 examples). Therefore, we resort to using validation set for training.

F QLoRA Finetuning Hyperparameters

Hyperparameter	Value
Quantisation	
Precision	4-bit
dtype	float16 + nf4
LoRA	
Rank	64
α	16
Dropout	0.1
LoRA Modules	All Layers
Finetuning	
Scheduler	Constant
Learning Rate	0.0002
Warm-Up Ratio	0.03
Batch Size	16
Weight Decay	0.0
Learning Steps	10000

Table 13: Hyperparameters for QLoRA tuning. Except for the hyperparameters provided, the rest were set to the default values in HuggingFace Transformers.

G In-Context Learning Results for XNLI

Model		ES	RU	TR	AVG
Llama-2-7B	Raw	32.3	33.1	26.2	30.5
	+QLoRA	33.1	33.4	25.8	30.7

Table 14: ICL results for XNLI before and after QLoRA language adaptation.

H Human Annotation Instructions for Naturalness and Usefulness

Score	**Instruction Text**
	Naturalness
Please read the text below and evaluate its *naturalness* using 3-point scale below:	
1	The text is not natural, you would never say anything similar as a native speaker.
2	The text is generally natural but contains some elements which feel odd.
3	The text is completely natural, it contains nothing you find odd.
	Usefulness
Please read the instruction and the text and evaluate its *usefulness* using 3-point scale below:	
1	The text does not provide any useful information to complete the task in the instruction.
2	The text contains some information which could be useful for the instructions.
3	The text contains full information and completes the task in the instructions.

Table 15: Text of instructions used in human evaluation.

I Model Openness Analysis

Model	WEIGHTS	DATA
LaBSE	✓	D,O
XLM-R	✓	D,O
FLAN-base/-XL	✓	D
mT0-XL	✓	D,O
GPT-3.5	×	×
Aya	✓	D
Llama-2/-3.1	✓	D

Table 16: Model transparency and accessibility details based on two factors: a) whether model weights are openly available; and b) whether the pretraining data of the model is documented D and open O.