CLAfICLe: Cross Lingual Adaptation for In-Context Learning

Giulio Starace
University of Amsterdam / Amsterdam, The Netherlands
`giulio.starace@gmail.com`

Abstract

As the field of natural language processing (NLP) is enveloped with pre-trained large language models (LLMs), it becomes more and more dependent on data and compute. In parallel, fine-tuning paradigms such as in-context learning (ICL) have emerged to address these requirements. Unfortunately, most of this research has only been conducted in English and is either prohibitively expensive or impossible to repeat in other languages. Multilingual LLMs have been proposed to address this issue, but have been shown to be outperformed by their monolingual counterparts. While research on the language adaptation of monolingual models shows promising results, this typically focuses on encoder-only transformers and in the decoder-only setting is limited to intrinsic evaluation. In this work, we tackle the problem of the cross-lingual adaptation of monolingual models fine-tuned to perform ICL. We combine state-of-the-art (SOTA) language and task adaptation techniques and show that it is still difficult to outperform a simple baseline consisting of sandwiching the model between translation API calls. Finally, we introduce a novel technique for post-hoc disentanglement, PHoDiVA, and propose directions for future research. We release our code and models to the public[1].

1 Introduction

Pre-trained LLMs are dominating NLP research for tackling downstream tasks (Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020). These models rely on the availability of vast amounts of unsupervised training data and the high usage computing resources that can be leveraged by variants of the transformer architecture (Vaswani et al., 2017). Because access to such data and compute is limited, and due to the Anglocentric nature of the field, the majority LLM research and application prioritises the English language. This leads to a large gap between what can be achieved in English and other languages. Research in multilingual LLMs attempts to address this issue, with encouraging results in many aspects (Conneau et al., 2020; BigScience Workshop, 2022). These multilingual models however have been shown to underperform against monolingual counterparts (Wu and Dredze, 2020), and datasets for more niche applications such as fine-tuning for zero-shot and in-context prompted generalisation remain almost exclusively in English (Bach et al., 2022; Mishra et al., 2022). One approach for addressing this issue is developing techniques for adapting existing English models to work in other languages. Recent research in model adaptation has shown promising results (Houlsby et al., 2019; Ainsworth et al., 2022), and there already exist some works applying these techniques directly to the problem of cross-lingual transfer (Artetxe et al., 2020). These works however mainly focus either on encoder-only transformer (EOT) variants or on performance on a few downstream tasks, through the use of additional language- and/or task-specific fine-tuning (de Vries et al., 2021; Gogoulou et al., 2022). Works that consider decoder-only transformer (DOT) variants on the other hand (de Vries and Nissim, 2021; Minixhofer et al., 2022) limit the scope to pre-trained variants and intrinsic evaluation, with little focus on how their techniques interact with *fine-tuned* models and their performance on downstream tasks.

This work instead considers techniques for the efficient cross-lingual transfer of models fine-tuned on *in-context learning* (ICL). Here, the models are fine-tuned to leverage information presented in the context window to address some downstream task, demonstrating improved performance and generalisation (Wei et al., 2021; Sanh et al., 2022; Wang et al., 2022), enabling multi-task learning and eliminating the need for task-specific fine-tuning. Scaled versions of these models (Chung et al., 2022) are now on par with the best models from the simi-

---

[1] https://github.com/thesofakillers/CLAfICLe

larly emerging paradigm of LLM training with reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). Lamentably, just like their training, the evaluation of these ICL fine-tuned models relies on instruction and prompt templates, which are mainly available only in English. This renders any cross-lingual adaptation of these models futile, as there is no way to extrinsically evaluate them in the target language. Recent work from Min et al. (2022) circumvents this requirement by directly fine-tuning a model on chains of input-output pairs from a suite of tasks, matching the performance of instruction-based ICL. We therefore focus on adapting models trained under this particular framework, and contribute the following:

1. We show that Minixhofer et al. (2022)'s WECHSEL language-adaptation technique scales, successfully adapting the large variant of GPT2 (774M) to French and German. We release our checkpoints, which previously did not exist at this parameter scale for these languages.

2. We continue the evaluation of WECHSEL by assessing its robustness, applying it to a fine-tuned variant of GPT2 and measuring its performance on a number of downstream tasks, rather than just examining perplexity.

3. We share our methods and results for adapting fine-tuned models capable of ICL from English to French and German.

4. We introduce the notion of "targeted distillation", a form of post-hoc disentanglement leveraging adapters (Houlsby et al., 2019) to extract only the fine-tuned information from a fine-tuned model. We refer to our technique as PHODIVA (**P**ost **Hoc** **Di**sentanglement via **V**essel **A**dapters).

Surprisingly, we fail to match the performance of a simple baseline consisting of sandwiching the model between translation API calls, which performs almost on par with the original model. We hypothesize this may be due to under-trained models, and propose directions for future research.

2 Related Work

2.1 In-context learning

ICL in NLP is a paradigm most famously popularized by Brown et al. (2020), where the input in a transformer's context window is augmented with some additional information useful for some downstream task. The model is said to "learn" from (*sc.* leverage) this information, without the need for any parameter updates. A very basic example of ICL is "prompting" the model by prefixing the input with task-specific instructions. ICL provides a number of advantages. Because the input is entirely in natural language, it acts as an interpretable interface for human-model interaction. As mentioned, ICL does not require any parameter updates, greatly reducing the computational costs necessary and enabling the multi-task generalisation necessary for language-model-as-a-service applications (Sun et al., 2022). Considerable research attention has been devoted to the paradigm (Liu et al., 2022; Lu et al., 2022; Wu et al., 2022), with many contributions focusing on informal reasoning performance through scratchpads (Nye et al., 2021), bootstrapping (Zelikman et al., 2022), chain-of-thought prompting (Wei et al., 2022; Wang et al., 2023), learned verifiers (Cobbe et al., 2021) and selection-inference (Creswell et al., 2023) techniques among others. Dohan et al. (2022) present a framework for formalizing these techniques while Zhao et al. (2021) propose solutions to the performance sensitivity to prompt choice and ordering. Unsurprisingly, some studies find a performance boost can be achieved by fine-tuning the models directly on ICL examples (Sanh et al., 2022; Wang et al., 2022) and develop template-based datasets for this purpose (Bach et al., 2022; Mishra et al., 2022). To circumvent the variability or need for templates, Lester et al. (2021) fine-tune a "soft" prompt instead of the model, while Min et al. (2022) directly use chains of input-output pairs for fine-tuning rather than templates. All of these approaches limit themselves to English, leaving a large gap in multilingual NLP. Our work attempts to partly address this gap by exploring ways of adapting ICL-fine-tuned English models to other languages.

2.2 Multilingual NLP

Democratizing LLMs to other languages has mainly been achieved by repeating the pre-training process on massive multi-lingual corpora. While non-transformer approaches exist (Artetxe and Schwenk, 2019), most attention is devoted to transformers, resulting in encoder-only, encoder-decoder and decoder-only models such as XLM-R (Conneau and Lample, 2019), mT5 (Xue et al., 2021), XLGM (Lin et al., 2021b), mBART (Liu et al., 2020) and BLOOM (BigScience Workshop, 2022). While undoubtedly valuable, these models tend to suffer from the *curse of multilinguality* (Conneau et al., 2020), where the performance degrades as the number of languages increases. Nozza et al. (2020) and Wu and Dredze (2020) reproduce this claim, showing that monolingual models tend to outperform multilingual models on downstream tasks. To address this issue, researchers have re-trained transformers in other languages (Martin et al., 2020; de Vries et al., 2019; Chan et al., 2020; De Mattei et al., 2020) but this approach may not scale well to all 200+ languages of the world. Instead, recent work has focused on efficiently transferring monolingual English representation to other target languages. Artetxe et al. (2020) first tackle this problem with the aim of testing Pires et al. (2019) and Cao et al. (2022)'s hypothesis that joint training across multiple languages on a shared vocabulary gives rise to cross-lingual representations that generalise across languages. They propose a new method for monolingual representation transfer consisting in retraining the lexical embeddings in the target language while keeping the rest of the model frozen. de Vries et al. (2021) successfully apply this method to low-resource target languages and de Vries and Nissim (2021) extend the method to DOT, successfully adapting GPT2 to Italian and Dutch. Recently, Minixhofer et al. (2022) have optimized the method by using dictionaries of parallel static multilingual word embeddings to reinitialize the embeddings in an efficient way before re-training them on the target language. Unfortunately in the DOT setting, none of these methods consider variants larger than GPT2-small (117M parameters) and crucially focus only on intrinsic evaluation rather than performance on downstream tasks. Furthermore, it remains to be seen whether any of these methods can be applied to fine-tuned versions of GPT2, without leading to catastrophic forgetting (McCloskey and Cohen, 1989). Our work attempts to address these issues by exploring the transferability MetaICL (Min et al., 2022) from English to French and German.

2.3 Adaptation

Adapters (Houlsby et al., 2019) are an emerging method in NLP for the efficient fine-tuning of LLMs, operating by injecting a small number of trainable parameters into the layers of a pre-trained base. Originating from computer vision (Rebuffi et al., 2017), in NLP they have mainly been used for adapting to new tasks (Stickland and Murray, 2019) and domains (Bapna and Firat, 2019) while avoiding catastrophic forgetting. There exist a few methods employing adapters for adapting to new languages (Pfeiffer et al., 2020b; Üstün et al., 2022), but these methods generally rely on multilingual transformer bases. In our work, we explore the use of adapters for separating fine-tuning from language-adaptation, so to avoid catastrophic forgetting occurring during the latter. Contemporaneously, Marchisio et al. (2022) propose a similar method which trains a small model in parallel to the base and then operates over the smaller model for adaptation. Other non-adapter approaches include meta-learning, which has also been applied to language adaptation (Nooralahzadeh et al., 2020).

2.4 Other related work

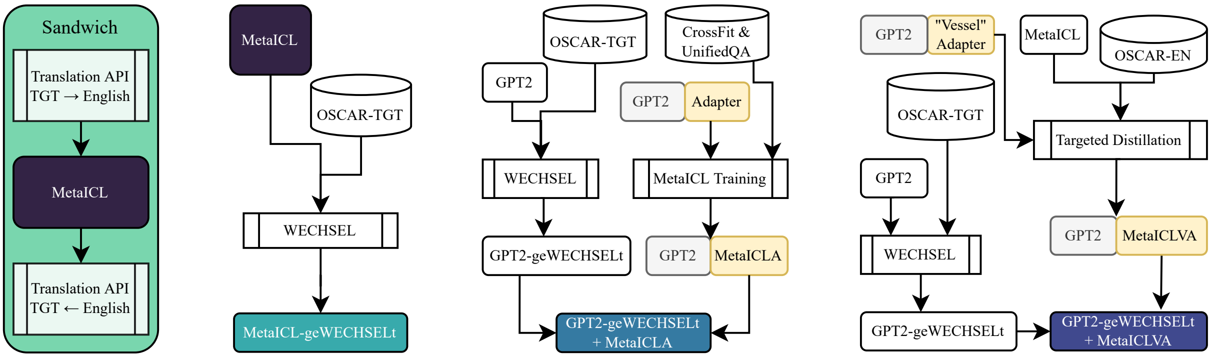
Our work makes brief references to the area of distillation (Hinton et al., 2015), which is typically defined to be the process of transferring information from a large model to a smaller one. We also briefly touch on disentanglement (Bengio et al., 2013), wherein representations are "disentangled" such that a change in one dimension corresponds to a change in one factor of variation but does not affect other factors. In our case, we are interested in a "post-hoc" disentanglement of the representations responsible for the fine-tuned capabilities of a base model, similarly to Khrulkov et al. (2021).

3 Method

We present an overview of our methods in Fig. 1.

3.1 MetaICL

Due to the complete lack of prompting/instruction templates in non-English languages, we rely on MetaICL (Min et al., 2022), which circumvents the need for prompt/instruction templates at train-time and test-time. With MetaICL, a pre-trained DOT is fine-tuned by concatenating k examples of input-output pairs ("shots") from a variety of tasks and feeding this as input to the model. The final input-output pair is truncated such that only the input is shown, and the model is trained to predict the output using a negative log-likelihood objective from a number of possible options. The trained model then generalises to unseen tasks presented



(a) (b) (c) (d)

Figure 1: Overview of each of the models evaluated in one of the two target (TGT) languages (French or German). The baseline Sandwich model (a) sandwiches MetaICL (Min et al., 2022) (which we separately evaluate only in English) between two complementary translation API calls. MetaICL-geWECHSELt (b) is the result of applying WECHSEL (Minixhofer et al., 2022) to MetaICL. GPT2-geWECHSELt+MetaICLA combines MetaICLA, an adapter trained on the MetaICL dataset and objective, with a TGT-language GPT2 base obtained via WECHSEL. GPT2-geWECHSELt+MetaICLVA does the same, except MetaICLVA is trained via targeted distillation with supervision provided by MetaICL. For more details, refer to section 3.

in the same way by utilizing the k shots provided in the context. We refer to this model as *MetaICL*.

3.2 Sandwich

As a baseline, we consider the obvious solution of simply translating input in the target language to English, feeding the translation to MetaICL, and translating the output back to the target language. We refer to this model as *Sandwich*. We make use of Google's Cloud Translation AI API[2].

3.3 WECHSEL

Aside from translation API calls, to adapt a monolingual DOT from a source language to a target language we employ WECHSEL (Minixhofer et al., 2022), which has shown success in adapting the small variant of GPT2 (117M parameters) to a number of target languages. WECHSEL works by retraining the tokenizer into the target language and re-initializing the transformer embedding layers such that the target embeddings are semantically similar to the source embeddings. This is done by leveraging existing parallel multilingual static word embeddings. As done by de Vries et al. (2021), after re-initialization, additional causal language modeling (CLM) is performed in the target language to account for syntactical differences. Applying WECHSEL to MetaICL, we obtain what we refer to as *MetaICL-geWECHSELt*.

3.4 Adapters

Because we are interested in adapting a fine-tuned DOT (MetaICL), we hypothesize that the additional CLM at the end of WECHSEL can lead to catastrophic forgetting of the fine-tuning. Furthermore, we hypothesize that the fine-tuning may contain language-specific information, entangled with the task information relevant to the fine-tuning objective. To address this issue, inspired by MAD-X (Pfeiffer et al., 2020b) we train a "task adapter" on the same ICL objective and data as MetaICL with a GPT2 base, obtaining an "ICL-adapter", which we refer to as *MetaICLA*. Adapters introduce "bottleneck" dense layers at each transformer layer of their base. The adapter is trained on a particular objective while the base is kept frozen, allowing for parameter-efficient and modular fine-tuning. These dense layers consist in a down matrix $\mathbf{W}_{down}$, projecting the hidden states into a lower dimension $d_{bottleneck}$, a non-linearity f, which is applied to this projection and an up matrix $\mathbf{W}_{up}$ that projects back to the original dimension:

$$\mathbf{h} \leftarrow \mathbf{W}_{up} f(\mathbf{W}_{down}\mathbf{h}) + \mathbf{r}, \qquad (1)$$

where $\mathbf{r}$ is a residual connection. Having separated the task-specific information, we apply WECHSEL to the GPT2 base, obtaining what we refer to as *GPT2-geWECHSELt*. Adding MetaICLA to GPT2-geWECHSELt, we obtain a model theoretically capable of ICL in the target language, *GPT2-geWECHSELt+MetaICLA*.

---

[2] https://cloud.google.com/translate

3.5 PHODIVA

To address situations where repeating fine-tuning is not permissible, either because the data is not released, the process too complicated or the compute simply not available, we propose PHODIVA. Here, instead of repeating ICL fine-tuning, we leverage the fine-tuned MetaICL checkpoint, using it as a teacher in a modified student-teacher offline distillation (Hinton et al., 2015) setup. More specifically, before WECHSEL adaptation, we add a “vessel” adapter to a (frozen) GPT2 base, and then perform CLM in the source language (English). Vessel adapters are exactly the same as task adapters, except that they act as a “vessel” for distilled capabilities rather than as additional parameters for fine-tuning. Rather than predicting the actual next word, the adapter is trained to predict the next word greedily sampled from the teacher. The idea is to overfit the adapter to the teacher outputs (hence the greedy sampling). Because the GPT2 base is frozen and theoretically shares the original language modeling capabilities of the teacher, we hypothesize that this “targeted distillation” can disentangle the fine-tuned capabilities into the vessel adapter. We use the CLM objective because of the constraint to keep the distillation process as simple as possible, so to make it advantageous over repeating a potentially complex fine-tuning process. The only constraint of this method is that the adapter base is the same pre-trained base that was fine-tuned into the teacher. When using MetaICL as the teacher, we refer to the resulting vessel adapter as *MetaICLVA*. Like in section 3.4, after applying WECHSEL to a GPT2 base, we can then combine the language-adapted base and MetaICLVA to obtain *GPT2-geWECHSELt+MetaICLVA*, another model theoretically capable of ICL in the target language.

4 Experimental Setup

We use the PyTorch Lightning Python framework (Falcon and The PyTorch Lightning team, 2019) to implement our work. Because we envision the direct application of this work to be most useful to smaller companies and start-ups, we limit our compute to a single 40GB NVIDIA A100 GPU and run jobs for a maximum of 24 hours.

4.1 Models

There exist various possible adapter setups, specifying different configurations of the up and down weight matrices, non-linearity and residual connection, among other settings. For our work, we use the `pfeiffer` configuration from AdapterHub (Pfeiffer et al., 2020a).

Table 1: The datasets constituting our ICL benchmark. Most originate from pre-existing benchmarks, namely XGLUE (Liang et al., 2020) and (Lin et al., 2021a).

Dataset	(Origin)	Collection
HateCheck	(Röttger et al., 2021)	-
XNLI	(Conneau et al., 2018)	XGLUE
QAM	(Liang et al., 2020)	XGLUE
QADSM	(Liang et al., 2020)	XGLUE
PAWS-X	(Yang et al., 2019)	XGLUE
MARC	(Keung et al., 2020)	-
X-CODAH	(Lin et al., 2021a)	XCSR
X-CSQA	(Lin et al., 2021a)	XCSR
Wino-X	(Emelin and Sennrich, 2021)	-

Regarding MetaICL, Min et al. (2022) train a number of variants, releasing checkpoints however only for variants fine tuning the large version of GPT2 (774M parameters). We base the rest of our models on the same GPT2 version and use the “high resource to low resource” direct MetaICL checkpoint as we consider this to be the most realistic. We make use of the HuggingFace Transformers (Wolf et al., 2020) implementation of GPT2 throughout.

4.2 Evaluation

For assessing ICL, we reimplement the same evaluation setup as Min et al. (2022), evaluating a given model on a suite of tasks, prepending the input with $k = 16$ input-output pairs sampled randomly for each task from a separate split. For our evaluation metrics, like Min et al. (2022), we use F1 for tasks where the label options change across examples, and accuracy for tasks where the label options are always the same. Because the evaluation benchmark used by Min et al. (2022) is limited to English, we develop our own multilingual multi-task benchmark spanning 9 different tasks across 3 languages (English, German, and French). Our benchmark design is restricted to tasks that can be handled by the MetaICL framework, namely multi-class, single-label tasks. To enable complete comparisons across languages, we also restrict our benchmark to only contain language-parallel datasets. Correspondingly, we list the benchmark datasets in Table 1.

To evaluate the successful application of WECHSEL to GPT2, we use the same process as Minixhofer et al. (2022), namely measuring perplexity on a held out test set. For all language modeling, we use the original release of the OSCAR corpus (Ortiz Suárez et al., 2020).

4.3 Training

When performing CLM training, due to our limited compute, we heed the advice of Geiping and Goldstein (2022) and pack samples into 1024-token sequences (the maximum length possible) by separating them with EOS tokens, so to minimize the number of padding tokens and maximize GPU utilisation. With this we are able to fit a batch size of 2 into memory, while actually presenting the model with more than two examples per batch in most cases[3]. We achieve a virtual batch size of 512 by accumulating gradients over 256 steps. We employ single-epoch training (Komatsuzaki, 2019) on a total of 600M tokens, which we estimate to be the number of tokens consumed by our model in a single epoch by running a profiling run on a smaller download. Based on the information in Geiping and Goldstein (2022) and Minixhofer et al. (2022), we decide to use Adam (Kingma and Ba, 2015) with a linear warmup for the first half of training to a peak learning rate of 5e-4, followed by cosine annealing to 0 by the end of training. When performing targeted distillation for MetaICLVA, we reduce the linear warmup to the first 10% of training to help with our voluntary overfitting. As suggested by Izsak et al. (2021), to maximize training time, we evaluate on only 0.5% of the data, logging every 50 steps.

For the ICL training necessary for MetaICLA, we modify Min et al. (2022)'s implementation so to work with adapters. In particular, we use their `HR→LR` training mixture, which consists of 61 tasks sourced from the CrossFit (Ye et al., 2021) and UnifiedQA (Khashabi et al., 2020) benchmarks.

5 Results and Discussion

Fig. 2 shows the performance of GPT2 after around 1k steps of training, evaluated intrinsically in terms of perplexity. For both French and German, we see perplexity decrease to sub-50 values, with the French model reaching a perplexity of ≈ 28. Both models are clearly underfit, still monotonically decreasing by the end of the training. These observations are roughly in-line with Minixhofer et al. (2022)'s findings for smaller variants of GPT2, although we train for much less time and hence are left with higher perplexities. While we believe our preliminary results suggest WECHSEL scales well to larger models in terms of intrinsic evaluation, future work may wish to investigate whether this holds for longer training times. The rest of our work considers, among other questions, the robustness of WECHSEL via extrinsic evaluation on downstream tasks performed by MetaICL.

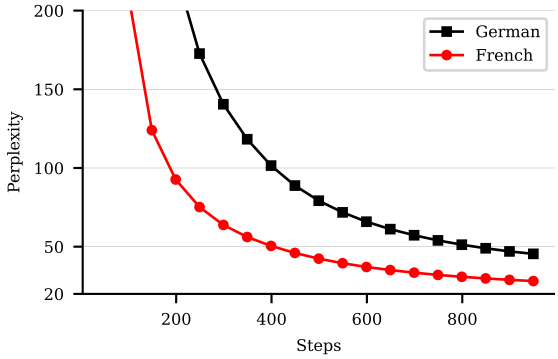


Figure 2: Perplexity on the held out set when performing the recommended CLM training after WECHSEL language-adaptation of GPT2. A step corresponds to an optimizer update. We evaluate every 50 steps.

Fig. 3 shows the performance on each dataset of our benchmark for the two baseline models, MetaICL and Sandwich. As summarized in Table 2, Sandwich performs roughly on par with MetaICL on both target languages, respectively with scores of 0.317 and 0.322 in French and German compared to MetaICL's score of 0.327 in English. We note generally low scores across all tasks. This is particularly perplexing in the case of MetaICL, scoring around 0.1 points less than with the evaluation ensemble used by Min et al. (2022), where the same checkpoint was reported scoring 0.417 in the worst case (a 25 % decrease). While similar values are reached in certain tasks in our benchmark (e.g. most of XGLUE and WINO-X), it is unclear what the origin of this discrepancy is, whether due to differences in evaluation implementation or difficulty of the tasks. Given that Min et al. (2022) simply report macro-averaged scores, it is impossible to verify the latter. Nevertheless, our results suggest that Sandwich-like solutions may be satisfactory for transferring performance from English to other languages given the surprisingly closeness of the scores. The decision between using Sandwich or "properly" adapted models with the same capabilities then becomes an economic one in terms of the

---

[3] This technique is also suggested by HuggingFace in their CLM tutorial: https://huggingface.co/course/chapter7/6.

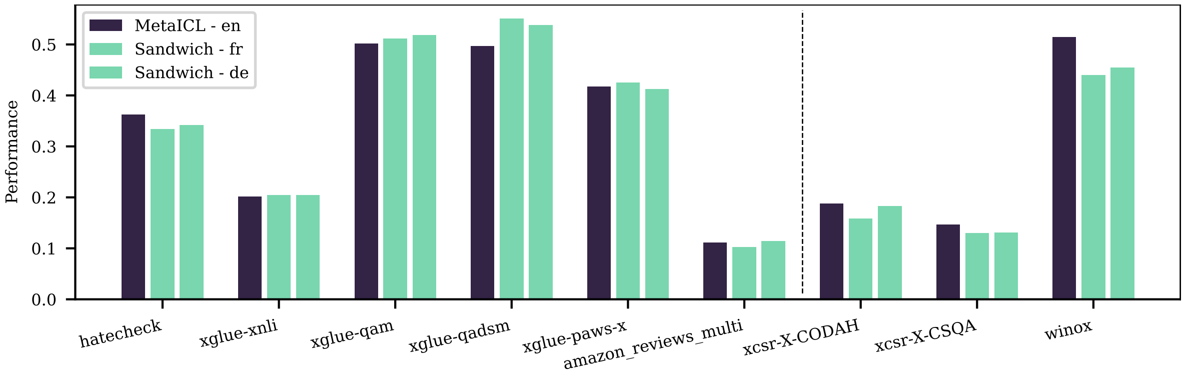


Figure 3: Performance (max is 1) on our multi-task benchmark of our two baseline models, MetaICL and Sandwich. The dashed line separates whether a given task uses accuracy (left) or F1-score (right) as the performance metric.

Table 2: Average performance (max is 1) on our multi-task benchmark for the models considered in this work. We use "W" as a shorthand for "geWECHSELt". We report average difference in performance for each proposed alternative to Sandwich. Negative values indicate underperformance compared to Sandwich.

	en	fr	de
MetaICL	0.327	-	-
Sandwich	-	0.317	0.322
Difference in Performance w.r.t. Sandwich			
MetaICL-W	-	-0.020	-0.026
GPT2-W+MetaICLA	-	-0.041	-0.042
GPT2-W+MetaICLVA	-	-0.036	-0.045

cost of API calls (for the former) versus the cost of inference plus training (for the latter).

Fig. 4 shows the difference in performance on each dataset of our benchmark between the proposed models and Sandwich. In general, we observe that the proposed models underperform across almost all tasks in both French and German, with the trends aligning at a task-level (e.g. all models underperform on QAM, by roughly the same amount). As reported in Table 2, the best of our proposed models is MetaICL-geWECHSELt, which underperformed Sandwich by roughly 0.02-0.03 points depending on the language. This undermines the motivation for the other two models, which were designed to avoid catastrophic forgetting by separating language and ICL capabilities via adapters. The results suggest that the trade-off between catastrophic forgetting and needing to train ICL-adapters leans in favour of the former in this compute regime. In this sense, we can conclude that WECHSEL does not suffer tremendously due to catastrophic forgetting when adapting fine-tuned DOTs such as the MetaICL variant of GPT2.

Our work is mainly limited by its preliminary nature. Apart for considering more appropriate (*sc.* larger) compute scales, future work could investigate training ICL-adapters more thoroughly, for example by performing hyperparameter optimization or incorporating more recent adapter research such as AdapterDrop (Rücklé et al., 2021) and AdapterFusion (Pfeiffer et al., 2021).

Similarly, we note that despite our best efforts our ICL evaluation benchmark faces some limitations. Due to the design restrictions mentioned in section 4.2, and the lack of non-English datasets, we only evaluate on 9 tasks, despite focusing on high-resource languages like German and French. This is a small number of tasks compared to what is done in English, where 10s and 100s of tasks can comprise a multi-task benchmark. Furthermore, many of our datasets are either human or machine translated versions of English, which can lead to noise due to "translationese" (Koppel and Ordan, 2011). We hope to raise awareness of the complete English dominance of the NLP data landscape.

We are also interested in a more complete treatment of PHODIVA. For instance, future work could explore different forms of student-teacher distillation, consider other forms of loss criterions, use beam search rather than greedy sampling and/or take inspiration from similar solutions such as Khrulkov et al. (2021)'s work on generative models. We believe work in this direction could benefit from simplifying the problem setting first, by considering a smaller, encoder-only transformer fine-tuned on a single downstream task on a single-language.

Other future work may consider different adaptation approaches that have recently emerged. For

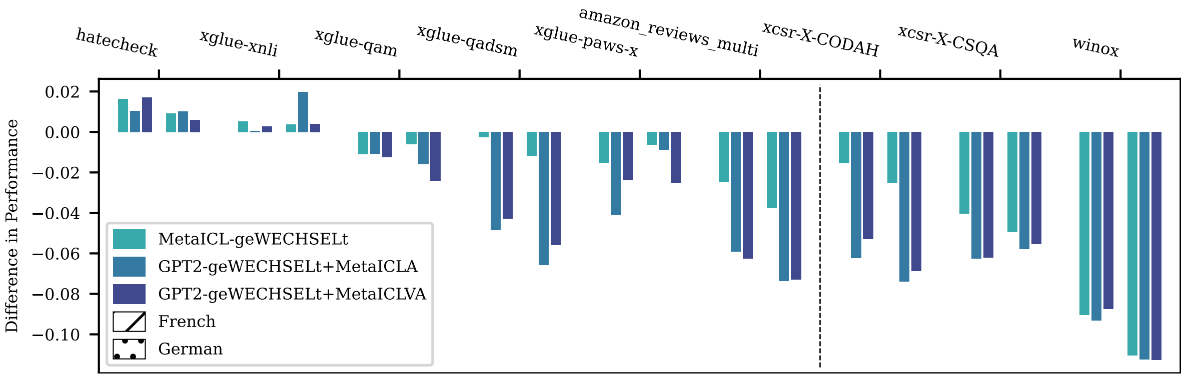


Figure 4: Performance gap on our multi-task benchmark between each of the language-adapted models and the "Sandwich" baseline. Positive values indicate that the adapted models are outperforming the baseline, while negative values indicate the reverse. The dashed line separates whether a given task uses accuracy (left) or F1-score (right) as the performance metric.

example, Marchisio et al. (2022)'s Mini-Model adaptation has yet to be tested on decoder-only transformers, and it would be interesting to see how it compares to WECHSEL in this regard. Other, slightly more distant approaches such as meta-learning a-la X-MAML (Nooralahzadeh et al., 2020) may provide different results.

Perhaps a clear limitation of this direction of research is that the setting remains monolingual. Future work could explore whether it is possible to adapt a monolingual model to multiple languages simultaneously, and how such adaptations would compare to monolingual-to-monolingual adaptation in terms of resources and performance. Finally, undermining all of this work is our restriction to results on a single random seed. Future work with more seeds and more compute would be necessary to draw more definitive conclusions.

6 Conclusion

We explore the problem of language-adapting a monolingual DOT previously fine-tuned to perform in-context learning. To this end, we stress test the current SOTA adaptation method, WECHSEL, scaling to previously untested model sizes, applying it to a fine-tuned variant of GPT2 (MetaICL) and evaluating extrinsically on a multi-task benchmark. While we find that WECHSEL successfully scales to larger model sizes, we note that at our compute regime, WECHSEL-adapted MetaICL underperforms compared to simply sandwiching the English model between translation API calls. We experiment with separating ICL fine-tuning and language adaptation to address potential catastrophic forgetting through the use of adapters, but find these approaches unsuccessful. In doing so, we propose PHoDiVA, a novel method for post-hoc disentanglement through vessel adapters. We share PHoDiVA in this rudimentary form as a starting point for future work in this direction.

7 Acknowledgements

We would like to thank Sami Jullien and Mozhdeh Ariannezhad for their helpful feedback and general supervision throughout this project. We would also like to thank Prof. Katia Shutova for acting as the examiner. Finally we thank the University of Amsterdam for providing the opportunity and resources necessary for the produced output.

References

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. 2022. Git Re-Basin: Merging Models modulo Permutation Symmetries.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M. Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan

Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. *arXiv:2202.01279 [cs]*.

Ankur Bapna and Orhan Firat. 2019. Simple, Scalable Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

BigScience Workshop. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

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. *arXiv:2005.14165 [cs]*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2022. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems.

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. *arXiv:1911.02116 [cs]*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv:1809.05053 [cs]*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *International Conference on Learning Representations*.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. GePpeTto Carves Italian into a Language Model.

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting Monolingual Models: Data can be Scarce when Language Similarity is High. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

Wietse de Vries and Malvina Nissim. 2021. As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model.

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.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-dickstein, Kevin Murphy, and Charles Sutton. 2022. Language Model Cascades.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Jonas Geiping and Tom Goldstein. 2022. Cramming: Training a Language Model on a Single GPU in One Day.

Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2022. Cross-lingual Transfer of Monolingual Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 948–955, Marseille, France. European Language Resources Association.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. *arXiv:1902.00751 [cs, stat]*.

Peter Izsak, Moshe Berchansky, and Omer Levy. 2021. How to Train BERT with an Academic Budget.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The Multilingual Amazon Reviews Corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries With a Single QA System.

Valentin Khrulkov, Leyla Mirvakhabova, Ivan Oseledets, and Artem Babenko. 2021. Disentangled Representations from Non-Disentangled Models.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Aran Komatsuzaki. 2019. One Epoch Is All You Need.

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A New Benchmark Datasetfor Cross-lingual Pre-training, Understanding and Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021b. Few-shot Learning with Multilingual Language Models. *arXiv:2112.10668 [cs]*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2022. Mini-Model Adaptation: Efficiently Extending Pretrained Models to New Languages via Aligned Shallow Training.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.

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.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. *arXiv:2104.08773 [cs]*.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-Shot Cross-Lingual Transfer with Meta Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making Sense of Language-Specific BERT Models.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the Efficiency of Adapters in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal 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 Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv:2110.08207 [cs]*.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5986–5995. PMLR.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-Box Tuning for Language-Model-as-a-Service. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20841–20855. PMLR.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-X: A Unified Hypernetwork for Multi-Task Multilingual Transfer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks. *arXiv:2204.07705 [cs]*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

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

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2022. Self-adaptive In-context Learning.

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.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A Few-shot Learning Challenge for Cross-task Generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR.