A Multilingual Perspective on Prompt Variability

Afra Baas
12417505
afra.baas@student.uva.nl

Amittai-Shlomo Aharoni
13652311
amittai-shlomo.aharoni@student.uva.nl

Emile Ghedamsi Dhifallah
14044994
emile.dhifallah@student.uva.nl

Fabian van Stijn
11906448
fabianvanstijn@gmail.com

Karsten Langerak
12567418
karsten.langerak@student.uva.nl

1 Introduction

In recent years, NLP has witnessed a significant shift from traditional fine-tuning approaches to the emerging paradigm of few-shot learning though prompting [Tu et al., 2022]. This transition has been driven by several factors, including the need to address data scarcity [Xia et al., 2020], improve model generalization [Raffel et al., 2019], and achieve more flexible and efficient learning capabilities. Fine-tuning, while effective for adapting pre-trained models to specific tasks, often requires a substantial amount of task-specific labeled data [Liu et al., 2022b]. However, this approach is not always feasible or practical, as collecting extensive labeled data for every task is challenging and time-consuming, especially in low-resource languages. Prompt tuning has emerged as a powerful alternative, enabling models to be fine-tuned using task-specific prompts that provide explicit instructions or constraints.

Nevertheless, despite the recent success of prompt tuning, there are still many open questions regarding the optimal design of prompts and the best way to incorporate them into pre-trained models. The behavior of prompt-tuned models can be influenced by subtle changes in the prompt phrasing or structure, which may lead to unexpected variations in their outputs and up to even 40% accuracy differences [Zhao et al., 2021]. The causes of these variations are not always clear, and it is often challenging to predict how performance will change for a given prompt. This lack of predictability can make it challenging to precisely control the model's behavior and ensure consistent results across different prompts. Researchers and practitioners are actively working on mitigating this issue by exploring methods to enhance the interpretability and control of prompt-tuned models ([Zhao et al., 2021], [Zhong et al., 2022], [Liu et al., 2022a]).

Since language and grammar are language-specific, prompts designed for English might not work well for other languages. Therefore, there is a need to evaluate the effectiveness of prompts across different languages. Some multilingual few shot learning models have been developed for such a task [Lin et al., 2022].

In this study, our objective is to investigate the variability in the effectiveness of prompts across a range of languages. Additionally, recent research has focused on evaluating the performance of single models across multiple languages. We aim to address the research question - what is the impact of varying specific linguistic properties, such as active vs. passive voice, non-modal vs. modal verbs, and substituting synonyms in prompts, on the performance of pre-trained language models (LMs) across different languages in the tasks of Sentiment Analysis (SA) and Natural Language Inference (NLI)? By investigating the effects of these linguistic variations on model performance, we can gain insights into the sensitivity and adaptability of pre-trained LMs to different grammatical constructions in various languages. For this research, four models were selected, namely BLOOM [wor], LLAMA [Touvron et al., 2023], BLOOMZ [Muennighoff et al., 2022] and Flan-t5 [Chung et al., 2022]. We will evaluate the performance of these pre-trained LMs on two tasks, Sentiment Analysis (SA) and Natural Language Inference (NLI) using prompts written in languages English, German and French. We selected SA because it is generally regarded as a straightforward task for models to excel in [Pang et al., 2002] and so it represents a good base-

line. On the other hand, NLI was chosen because a model's strong performance in this area signifies a robust semantic understanding. We will vary one linguistic property at a time, such as e.g. active vs. passive voice and non-modal vs. modal verb to evaluate the impact on performance. The code for our study can be found on our github repository [1].

2 Related Work

2.1 Robustness in Prompting

Recent research has demonstrated the effectiveness of prompt engineering in improving the performance of LMs. In particular, the study by [Sanh et al., 2021] showed that multitask prompted training can enable zero-shot task generalization. [Brown et al., 2020] showcased that, by pretraining on large-scale datasets, Language Models (LMs) significantly enhances their task-agnostic, few-shot performance, occasionally even matching or surpassing the effectiveness of previous state-of-the-art fine-tuning approaches. [Qi et al., 2022] demonstrated the usefulness of prompt-learning based frameworks for multilingual settings. BLOOM, a 176B-parameter open-access multilingual language model developed by [wor], has shown impressive performance on several NLP tasks, including SA (on the SST-2 task) and NLI (on the MultiNLI task), as well as other wide range of tasks including NER, POS tagging, dependency parsing, SRL, coreference resolution, question answering, paraphrase identification, STS, commonsense reasoning, reading comprehension, language modeling, text classification, and machine translation. However, studies constantly show that slight variation in wording can result in drastic changes in prompt performance and it is still difficult to uncover what properties affect prompt performance across different models ([Gonen et al., 2022], [Mishra et al., 2022]).

2.2 Multilingual Prompting

To the best of our knowledge, no study has systematically evaluated prompt effectiveness across multiple languages. The main paper that surveys multilingual prompting for task generalization was [Muennighoff et al., 2022]. The paper compared between different models (BLOOMZ, BLOOMZ-MT, mT0-13B, mT0-13B-MT) on tasks like XNLI, XCOPA, XStoryCloze, and XWinograd using prompts in English (EN), machine-translated (MT), and human-translated (HT) prompts. It highlights that BLOOMZ performs better on English prompts compared to non-English prompts. However, when BLOOMZ is finetuned on machine-translated multilingual prompts (BLOOMZ-MT), it significantly improves performance on multilingual prompts, albeit with a slight decrease in performance on English prompts. The paeer also mentions the performance gains achieved by mT0 models when using MT prompts, with mixed results across different tasks. The paper reports on the impact of multitask finetuning on generative tasks. It mentions that generative performance initially jumps but then decreases, particularly on tasks requiring longer answers. It also discusses the performance of models on the HumanEval task, where multitask finetuning does not lead to significant performance improvements, except for smaller models like BLOOM-560M vs. BLOOMZ-560M. Furthermore, it mentions that BLOOM models perform better on languages that are more prevalent in the pretraining data. For example, XCOPA and XNLI show significantly better performance on high-resource languages like English, Spanish, or French. Relavant to us are German (de: 0.21%), English (en: 46.23%) and French (fr: 15.73%). We could not find follow up papers to BLOOMZ that examined the multilingual performance of prompting.

Previous works have also compared existing multilingual models, namely DV003, BLOOMZ and TULRv6, on XNLI tasks using prompting [Ahuja et al., 2023]. However, [Ahuja et al., 2023] did not make an attempt to formulate semantically meaningful prompts in different languages, nor did they evaluate the models' performance based on the syntactical structure of the prompts. Instead, the paper translated prompts from the English dataset to different languages. Therefore, this project aims to fill this gap in the literature.

3 Method

3.1 Prompt Design

To facilitate our study, we devised prompts in three different languages: English, German, and French. These languages were selected due to their high resource availability, e.g. datasets like MARC and XNLI contain all three off these languages. The fact that the samples in different languages are from the same dataset ensures that the examples is similar enough. e.g. the english and german amazon

[1] https://github.com/afra-baas/ATCS_group3

reviews are similar enough because their from the same dataset, MARC. Additionally, our familiarity with these languages will be helpfull in the analysis of the results.

The grammatical characteristics we will be comparing are active vs. passive voice, non-modal (auxiliary) vs. modal verb and common vs rare synonyms. Also compared to effect of only using different model words. The rare synonyms are chosen with the help of chatgpt. We will further refer to each of these 7 categories as prompt type.

The prompt variations were designed to ensure a significant level of similarity in sentence structure and content while modifying specific characteristics to differentiate between prompt type, e.g. active and passive constructions.

The prompts were inspired by Promptsource templates, which provided a systematic and standardized approach to generate diverse prompts.

Along with the instructions with specified the possible answers we want to model to give. For the Sentiment Analysis set, the possible answers are "A yes" or "B no" in each respective language. We mapped the dataset labels positive and negative to target words yes and no respectively. This works because the prompt instruction always directly or indirectly asks if the review is positive.
An example of an active English prompt for SA is:

Review: {content}[2]
Did the user write a positive review? Possible answers:
A yes
B no
Answer:

For the Natural Language Inference set, the answer options are "A yes," "B no" or "C maybe" representing entailment, neutral, and contradiction respectively.
An example of an active English prompt for NLI is:

Given the premise: premise \n Do we assume that this is true: hypothesis? Possible answers: \n A yes \n B no \n C maybe \n Answer:

The English prompts were translated into German and French using ChatGPT 3.5 Turbo, and subsequently assessed by native speakers of German and French. The complete set of prompts used to obtain the results in Section **??** can be found in Appendix A **??**.[3]

3.2 Significance tests

The Friedman test is employed to assess significant differences within a single prompt type, specifically determining if there is a notable difference in performance among the 10 active prompts. On the other hand, the Wilcoxon test is utilized to determine if there is a significant difference in performance when comparing two different prompt types, such as an active prompt and a passive prompt.

4 Experimental Setup

The main idea of our method/ Experimental setup is to decide on instruction prompts for each grammatical type, mentioned in Section **??**. Then take examples from the task the prompt instructs on and evaluated the performance of a model, by looking that the probability of each label being the next token. Our study takes a sentence from the MARC or XNLI dataset and inserts it into the instruction prompts we choose, see example in Appendix A.

4.1 Datasets

For our research, we utilized two multilingual datasets: the Multilingual Amazon Review Corpus (MARC) [Keung et al., 2020] and Cross-lingual Natural Language Inference (XNLI)[]. Both datasets were obtained from the Hugging Face library [Wolf et al., 2020].

The MARC dataset [Keung et al., 2020] was employed for sentiment analysis. It consists of reviews from various product categories on Amazon, available in six different languages. The training dataset contains 200,000 reviews, while the test and development datasets consist of 5,000 reviews each. All reviews are truncated to 200 tokens. The reviews in MARC are rated between 1 and 5 stars, with each star representing 20 percent of the data. For our study, we extracted the review content and used the star rating as the label. We focused on 1-star reviews as negative and 5-star reviews as positive, aiming to incorporate a greater polarity in the evaluation. Moreover, we only considered reviews containing a maximum of 100 tokens. Our experiments were conducted on a balanced sample size of 200 reviews, with 100 1-star reviews and 100 5-star reviews randomly sampled.

[2] {content} is replaced with each review to be predicted.

[3] Prompts used for the ablation studies can be found on our github https://github.com/afra-baas/ATCS_group3

The XNLI dataset [Conneau et al., 2018] was employed for natural language inference. It contains inferences in 15 different languages, with 122,000 training examples, 2,490 validation examples, and 5,010 test examples. The XNLI dataset was created by translating the Multi-NLI dataset [Conneau et al., 2018]. Each example in XNLI comprises a premise sentence, a hypothesis sentence and a label indicating entailment, neutral, or contradiction. For natural language inference, we randomly sampled 200 datapoints, with an equal distribution of 1/3 for each label category.

4.2 Models

The study involved the assessment of four different models, namely two pre-trained models, namely BLOOM (bigscience/bloom-560m) and LLAMA (huggyllama/llama-7b), and two instruction-tuned models, namely BLOOMZ (bigscience/bloomz-560m) and Flan-t5 (google/flan-t5-base).

We decided on 2 intruction tuned, and 2 models which are not since related work shows that instruction tuned models perform way better on prompt instructions. We wanted to observe if the models ability to better use prompt instructions will gave an effect on if certain prompt types have a better or worse effect on performance.

4.3 Evaluation

The pretrained models outputs logits, which contain the probability of each word in the vocabulary of the model to be the next word. We look at the softmax of the logits of the last token, as this best represents the probability of the models (complete) answer. add reference/cite
We than compare probability the model assigned to each possible answer[4]. The answer with the highest probability was considered the model's prediction.

Due to biased results observed in some models (refer to figure in Appendix B (norm vs no norm ablation)), we implemented logit scaling to mitigate the bias. Logit scaling adjusts the answer probabilities based on the bias introduced by the prompt instructions (add ref). This scaling was performed before selecting the answer with the highest probability.

The primary metric used to measure model performance was accuracy. This metric allowed us to assess the overall effect of prompt characteristics on model performance across different languages.(Additionally, we used F-score)

In this study, we employed the seeds 42, 33, and 50 as part of the experimental setup to ensure robustness in seed variability and reproducibility.

5 Results

5.1 variability within grammatical category across

Is there a significant difference in performance between the various prompts with one prompt type? It is anticipated that there will be variability within one prompt type, and more so for models that have not been instruction tuned, bloom and llama.

5.2 Grammatical performance across languages

Is there a significant difference in performance between the various prompt types?

It is anticipated that the use of diverse grammatical characteristics will introduce variability in the model's performance on the NLP tasks.

Figure 2 demonstrates a congruence with this finding, indicating a statistically significant distinction in accuracy when comparing the utilization of the English active prompt versus the English passive prompt. (tempory)

LLAMA: Friedman chi-square score: 5.0 p-value: 0.4158801869955079 Test Statistic: 0.0 p-value: 1.0

5.3 Robustness over languages

If a prompt characteristic changed the performance of the model, did this hold across languages?

The results indicate that in English, employing an active prompt yields better performance compared to using a passive prompt. Similarly, Figure 3 illustrates that the same holds true for German, with an active prompt outperforming a passive prompt. Figure 4 further confirms this trend, demonstrating that the advantage of using an active prompt persists in French as well. (tempory)

5.4 Ablation studies

We did an ablation study on the instruction prompts, to evaluate the effect of considering different possible answers or order of answers.

overview ablation study: - prompt instructions with no possible answers specified
- prompt instructions with Yes or no? / yes, no or

[4] A, B, C is the mean case of this paper. However, for the ablation with possible answer yes, no the probability of these words is taken.

Model	Prompt Type	English		German		French	
Bloom	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
Bloomz	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
t5	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
flan	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
Llama	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance

Table 1: Mean Accuracy and Variance for Each Prompt Type and Model on the Sentiment Analysis task

Model	Prompt Type	English		German		French	
Bloom	null	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
Bloomz	null	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
t5	null	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
flan	null	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
Llama	null	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	active	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	passive	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	auxiliary	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	common	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	rare_synonyms	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance
	identical_modal	Mean Acc	Variance	Mean Acc	Variance	Mean Acc	Variance

Table 2: Mean Accuracy and Variance for Each Prompt Type and Model on the Natural Language Inference task

maybe? specified
- prompt instructions with Possible answers: A yes B no / Possible answers: A yes B no C maybe specified
- prompt instructions with No or yes? / No, maybe or yes? specified

To observe the baseline performance on the models on the task, by giving no instruction prompts. The model in simply given the classification content, so review for the SA task and premise and hypothesis for the NLI task, and the probability of yes and no is extracted.

Furthermore, we considered a different approach to extracting the answer prediction of the model. Instead of looking at the models output logits to obtain which answer has the highest next token probability. We calculate to probability of the full prompt with each answer, and take the prediction comes from the full prompt and answer with highest probability.

6 Conclusion

Based on our results in figure 1 and figure 2 we can conclude that, for the grammatical characteristics we choose to evaluate, there is no significant difference in performance.

Meaning active and passive instruction prompts will give similar performance across languages. This shows a consistence in grammatical importance across European languages.

7 Discussion

7.1 Limitations and Future work

- Our research is done on a sample size 200, due to long runtimes. Future work can go deeper into the robustness of our conclusions, when evaluating on bigger sample sizes.

- We only evaluated on the tasks SA and NLI, future work can consider evaluating on more tasks to see if there is also a correlation to be found between tasks characteristics and important grammatical characteristics. To then evaluated is the same grammatical characteristics hold across languages.

- There a more ablation studies that can be done, such as different answer optional like A entailment B neutral C contradiction.

- Look for grammatical characteristics that are known to have a significant effect on the performance and only than compare across languages.

- include low resource languages aswell as high resource languages into the evaluation across languages.

- consider few-shot instruction prompting (instead of only one-shot prompting).

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. Mega: Multilingual evaluation of generative ai, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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. Scaling instruction-finetuned language models, 2022.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations, 2018.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation, 2022.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus, 2020.

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. 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, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.616.

Chi-Liang Liu, Hung yi Lee, and Wen tau Yih. Structured prompt tuning, 2022a.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=rBCvMG-JsPd.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk's language, 2022.

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. Crosslingual generalization through multitask finetuning, 2022.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, page 79–86, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118693.1118704. URL https://doi.org/10.3115/1118693.1118704.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.134. URL https://aclanthology.org/2022.acl-long.134.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.401.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Congying Xia, Chenwei Zhang, Jiawei Zhang, Tingting Liang, Hao Peng, and Philip S. Yu. Low-shot learning in natural language processing. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, pages 185–189, 2020. doi: 10.1109/CogMI50398.2020.00031.

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

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Panda: Prompt transfer meets knowledge distillation for efficient model adaptation, 2022.

Appendix A

			0
de		active	Prämisse: premise Nehmen Sie diese Prämisse als Wahrheit an. Impliziert diese Wahrhei
		auxiliary	Nehmen Sie das Folgende als wahr an: premise Impliziert diese Wahrheit die folgende A
		common	Nehmen wir an, es ist wahr, dass: premise Daher impliziert es auch: hypothesis? Möglich
	NLI	identical$_{m}odal$	Angesichts der: premise Dürfen wir annehmen, dass: hypothesis? Mögliche Antworten:
		modal	Nehmen Sie das Folgende als wahr an: premise Könnte diese Wahrheit die folgende Aus
		passive	Prämisse: premise, Diese Prämisse wird als Wahrheit angenommen. Wird die folgende A
		rare$_{s}ynonyms$	Stellen Sie die folgende Behauptung auf: premise Folglich impliziert es auch: hypothesis
		active	Bewertung: content Ist diese Bewertung positiv? Mögliche Antworten: A ja B nein Antw
		auxiliary	Bewertung: content Drückt diese Bewertung positive Stimmung aus? Mögliche Antworte
		common	Bewertung: content Betrachten Sie diese Bewertung als positiv? Mögliche Antworten: A
	SA	identical$_{m}odal$	Bewertung: content Dürfen diese Bewertung als positiv betrachtet werden? Mögliche An
		modal	Bewertung: content Könnte diese Bewertung positiv sein? Mögliche Antworten: A ja B
		passive	Bewertung: content Wird diese Bewertung als positiv angesehen? Mögliche Antworten:
		rare$_{s}ynonyms$	Bewertung: content Begutachtest du diese Bewertung als bejahend? Mögliche Antworten
en		active	Premise: premise Take this premise as truth. Does this truth entail the following statemen
		auxiliary	Take the following as truth: premise Does this truth entail the following statement: hypot
		common	Assume it is true that: premise Therefore, does it also entail: hypothesis? Possible answe
	NLI	identical$_{m}odal$	Given: premise Can we assume that: hypothesis Possible answers: A yes B no C maybe
		modal	Take the following as truth: premise Could this truth entail the following statement: hypo
		passive	Premise: premise, This premise is taken as truth. Is the following statement entailed by th
		rare$_{s}ynonyms$	Hypothesize the following is truthful: premise Consequently, does it conjointly entail: hy
		active	Review: content Is this review positive? Possible answers: A yes B no Answer:
		auxiliary	Review: content Is this review positive? Possible answers: A yes B no Answer:
		common	Review: content Do you consider this review to be positive? Possible answers: A yes B n
	SA	identical$_{m}odal$	Review: content Can this review be considered positive? Possible answers: A yes B no A
		modal	Review: content Could this review be positive? Possible answers: A yes B no Answer:
		passive	Review: content Is this review considered positive? Possible answers: A yes B no Answe
		rare$_{s}ynonyms$	Review: content Do thou envision this review to be positive? Possible answers: A yes B
fr		active	Prémisse: premise Prends cette prémisse comme vérité. Cette vérité implique-t-elle l'affi
		auxiliary	Prenez ce qui suit comme vérité: premise Cette vérité implique-t-elle l'affirmation suivan
		common	Supposons que c'est vrai: premise Par conséquent, est-ce que cela implique aussi: hypoth
	NLI	identical$_{m}odal$	Étant donné: premise Pouvons-nous supposer que cela est vrai: hypothesis? Réponses po
		modal	Prenez ce qui suit comme vérité: premise Est-ce que cette vérité pourrait impliquer l'affi
		passive	Prémisse: premise, Cette prémisse est considérée comme étant vraie. Le fait suivant est-i
		rare$_{s}ynonyms$	Présumez que ce qui suit est véridique: premise En conséquence, est-ce que cela entraîne
		active	Critique: content Cette critique est-elle positive? Réponses possibles: A oui B non Repon
		auxiliary	Critique: content Cette critique est-elle positive? Réponses possibles: A oui B non Repon
		common	Critique: content Considérez-vous cette critique comme positive? Réponses possibles: A
	SA	identical$_{m}odal$	Critique: content Pourrait-on considérer cette critique comme positive? Réponses possibl
		modal	Critique: content Est-ce que cette critique pourrait être positive? Réponses possibles: A
		passive	Critique: content Est-ce que cette critique est considérée comme positive? Réponses poss
		rare$_{s}ynonyms$	Critique: content Estimes-tu cette critique comme étant favorable? Réponses possibles: A