

# TACO: ENHANCING CROSS-LINGUAL TRANSFER FOR LOW-RESOURCE LANGUAGES IN LLMs THROUGH TRANSLATION-ASSISTED CHAIN-OF-THOUGHT PROCESSES

**Bibek Upadhayay, Vahid Behzadan, Ph.D.**

SAIL Lab

University of New Haven

West Haven, CT 06516, USA

{bupadhayay, vbehzadan}@newhaven.edu

## ABSTRACT

Creating multilingual LLMs poses a significant challenge. Pretraining or fine-tuning LLMs to adopt new languages is evidently very costly. Furthermore, there exist limitations concerning benchmark datasets and the metrics used to measure model performance in multilingual settings. This paper proposes cost-effective solutions to both aforementioned challenges. Firstly, we introduce the Multilingual Instruction-Tuning Dataset (MITS), comprised of Alpaca-52K, Dolly-15K, and Vicuna Benchmark translations into 132 languages. Secondly, we propose a new method called *TaCo: Translation-Assisted Cross-Linguality*, which utilizes translations in a chain-of-thought process to instruction-tune LLMs on new languages through a curriculum-learning process. As a proof of concept, we experimented with the instruction-tuned Guanaco-33B model, performing further instruction tuning using our proposed TaCo method in three low-resource languages and one high-resource language. Our results indicate that the TaCo method impresses GPT-4 with an 82% score for a low-resource language in the Vicuna Benchmark dataset, doubling the performance in contrast to instruction tuning alone. Furthermore, TaCo shows promise in creating multilingual LLMs, even for low-resource languages. We have released our datasets and model adapters<sup>1</sup>, encouraging the research community to utilize these resources to advance work on multilingual LLMs.

## 1 INTRODUCTION

Languages are not mere tools for communication; they are rich repositories of cultural heritage, historical archives, and mediums for preserving traditions, reflecting the unique cognitive abilities of human beings. Languages encapsulate ancestral knowledge, traditions, and unique worldviews, offering insights into day-to-day lives and socio-cultural diversities. The alarming rate at which languages are disappearing, with rare languages fading into oblivion approximately every week, signals a global crisis (UNESCO (IESALC), 2022). In this context, Large Language Models (LLMs) have emerged as a promising strategy for the preservation of these linguistic treasures. As the size of LLMs increases, they have shown to improve in downstream tasks not only in English but also in non-English languages (Goyal et al., 2021; OpenAI, 2023; Conneau & Lample, 2019). OpenAI’s GPT-4, for instance, performs across 26 languages, achieving state-of-the-art scores in the Multitask Multilingual Language Understanding (MMLU) dataset (Hendrycks et al., 2020).

LLMs can play a pivotal role in analyzing and revitalizing low-resource and endangered languages by teaching them vocabulary, grammar, and making use of available texts and resources. These models can facilitate the generation of additional language resources, enhance linguistic research, and create interactive applications to attract more users, thereby providing a technological buffer

<sup>1</sup><https://github.com/UNHSAILLab/TaCo>

against language extinction. However, the availability of sufficient resources to train LLMs in non-English languages, especially those that are low-resource, remains a significant challenge.

The scarcity of rich datasets for training LLMs in non-English languages is a critical barrier. Proprietary models such as PaLM (Anil et al., 2023) and GPT-4 (OpenAI, 2023), despite their capabilities in understanding and performing downstream tasks in non-English languages, are hindered by their closed model architecture and the inaccessibility of their post-processed datasets to the research community. This limitation restricts academic and independent researchers from building upon or tailoring these models for specific uses. Moreover, these proprietary models cover only a limited number of non-English languages. For instance, GPT-4’s performance on languages beyond the 26 evaluated in the MMLU dataset remains unspecified, and our preliminary experiments have shown that models like GPT 3.5 often produce responses that mix similar languages, such as Sanskrit and Hindi. Similarly, models based on the newest technologies, such as Google’s Bard, reveal their training limitations when tested with low-resource languages. To counteract these challenges, the open-source community has made significant strides by releasing LLMs with accessible model weights. Models such as BLOOM (Scao et al., 2022), POLYLM (Wei et al., 2023), and Glot-500 (ImaniGooghari et al., 2023) are pushing the boundaries towards truly multilingual LLMs by including non-English languages in their training datasets and employing techniques such as instruction tuning to enhance their versatility (Peng et al., 2023).

However, achieving multilingualism in LLMs is fraught with challenges, not the least of which is the substantial resource requirement for pretraining. Moreover, as the number of languages in a model’s repertoire increases, a phenomenon known as the curse of multilinguality becomes apparent, indicating a decrease in performance for low-resource languages (Conneau et al., 2019). Aligning languages with diverse syntactic structures further complicates the matter (Dufter & Schütze, 2020).

In addressing these challenges, we introduce a novel methodology named TaCo, leveraging the translation in the chain-of-thought process to foster a more inclusive multilingual model. Using established LLMs’ properties—namely, emergent capabilities with increased size (Wei et al., 2022a) and the elicitation of reasoning behaviors through chain-of-thought prompting (Wei et al., 2022b)—TaCo employs a curriculum learning strategy (Bengio et al., 2009). It utilizes a fine-tuned Guanaco-33B model (Dettmers et al., 2023) on the OASST1 dataset (Köpf et al., 2023) in conjunction with instruction tuning via Low-Rank Adaptation (LoRA)(Hu et al., 2021) for efficient fine-tuning. Distinct from other methodologies that apply LoRA adapters for multilingualism or perform instruction tuning on base models (Li et al., 2023; tloen, 2023), we leverage the curriculum learning with advanced capabilities of the fine-tuned Guanaco-33B model. This approach streamlines the process of teaching the model to translate and generate responses in respective languages, minimizing the need for intensive model training from scratch and thereby saving on overall training costs. Through these efforts, we aim to contribute to preserving the linguistic diversity that enriches the fabric of human culture and cognition.

We present our overall contribution as follows:

1. We present the Multilingual Instruction-Tuning Dataset (MITDS), which consists of the translated Alpaca-52K (Peng et al., 2023) and Dolly-15K (Conover et al., 2023). These datasets, translated using Google Cloud Translation, cover 132 languages.
2. We propose a new method called TaCo for instruction-tuning LLMs to learn new languages.
3. We also present the Multilingual Vicuna Benchmark, which is a dataset comprised of translation of the Vicuna Benchmark (Chiang et al., 2023) in 132 languages and made them publicly available.
4. We evaluate four TaCo models on the Vicuna Benchmark, for 3 low-resource languages (Nepali, Sanskrit, and Maithili), as well as a high-resource language: Persian.
5. We release the adapters for the aforementioned four language models for public use.

## 2 RELATED WORK

Training LLMs for multilingual purposes involves two major approaches: pretraining on multilingual data and fine-tuning in a new language. Kenton & Toutanova (2019) introduced the concept of a multilingual-BERT pretrained model, which laid the foundation for subsequent advancements.

Conneau et al. (2019) discussed the trade-offs involved in adding languages to improve performance on low-resource languages, called *curse of multilinguality*. They noted that past a certain point, performance begins to decline. Their work with the XLM-R model involved training on extensive data to bolster cross-lingual representation. Improving upon these efforts, Chi et al. (2022) introduced discriminative pretraining tasks in the creation of XLM-E, which enhances cross-lingual transferability while reducing computational cost. Dufter & Schütze (2020)’s approach utilized smaller BERT models, incorporating synthetic data alongside natural data, and adjusted the masking strategy to create efficient multilingual BERT models.

Considering the impact of model size, Goyal et al. trained XLM-RXL and XLM-RXXL models, demonstrating improved cross-lingual understanding in larger models Goyal et al. (2021). Their findings highlight the capabilities of large-scale models in zero-shot learning and underscore the enhanced performance achievable through fine-tuning. Scao et al. (2022) introduced BLOOM, a large 176B-parameter open-source model, trained on a diverse corpus to support multiple languages and tasks, showing remarkable task generalization. Following this, Muennighoff et al. (2022) furthered this initiative by fine-tuning BLOOM and mT5 models on multitask prompts, resulting in the BLOOMZ and mT0 variants that demonstrated improved performance across languages and tasks. Li et al. (2023) embarked on training Bactrian-X models using the LoRA approach, aiming at bridging language representation in a vast array of languages. Similarly, Wei et al. (2023) released POLYLM, a model trained on a large multilingual dataset, employing a curriculum strategy to foster both commonsense generalization and multilingualism. ImaniGooghari et al. (2023) targeted predominantly low-resource languages with the Glot500-m model, creating a tailored pretraining environment to evaluate multilingual models’ performance effectively. Lastly, Kudugunta et al. (2023) explored the creation of a vast multilingual machine translation dataset, underpinning training models with up to 10.7B parameters, focusing on minimizing errors through a mix of original and back-translated data.

### 3 EXPERIMENT

**Dataset:** We initially translated two datasets, Alpaca-52K-GPT4 and Dolly-15K, into 132 languages using the Google Cloud Translation service. Due to computational resource constraints, we focused our experiment on three low-resource languages—Sanskrit, Nepali, and Maithili—and one high-resource language, Persian. Aware of the risk of *translationese*, a common issue in machine translations where the output deviates from native linguistic norms through literal translations, foreign syntax, and odd phrases (Borah; Koppel & Ordan, 2011), we conducted a manual evaluation for translation quality. We selected a sample of 1000 sentences from each language and assessed them using the BLEU score, calculated with ScarceBLEU (Post, 2018), ChrF (Popović, 2017), and the Translation Error Rate (TER) (Snover et al., 2006), by translating from English to the target language and back (en-xx-xx-en).

In our multilingual fine-tuning process, we employed a chain-of-thought approach to transform the translated instruction dataset. This method involves breaking down the complex problem of cross-linguality into manageable reasoning steps, akin to a chain of thought. By combining the English dataset with another language, we crafted question-answer pairs called as TaCo datasets. Here, instructions are given in a non-English language, while outputs are provided in both English and the original language. The process unfolds by first translating the instruction into English, formulating the response in English, and then translating this back into the non-English language. We avoid translating the *input* from the target language (xx) into English in our *output*. This method fosters the model’s comprehension and utilization of the original language nuances. By combining the Alpaca-52K and Dolly-15K datasets, we have curated a comprehensive collection of 67K question-answer prompts. The example of a TaCo data point is given in Fig. 1 (Left).

**Training:** Our training follows a curriculum learning strategy where we train the *pre-finetuned* model which has already produced state-of-the-art results in generating responses in English. Accordingly, we selected the Gaunaco-33B fine-tuned model and further instruction-tuned it with our TaCo method. We transformed the dataset to align with our TaCo method as mentioned earlier. The hyperparameters for the model’s training are as follows: a learning rate of  $3e-4$  and a cutoff length set to 2000. Our training targeted four LoRA modules: q\_proj, k\_proj, v\_proj, and o\_proj, with the LoRA parameters  $r$ ,  $\alpha$ , and dropout set at 32, 64, and 0.05, respectively.

**Evaluation:** In our model evaluation we used Vicuna Benchmark Chiang et al. (2023), however, the benchmark is available only in English. We translated the benchmark dataset into 132 languages using Google Cloud Translation and have made them publicly available. To evaluate our TaCo models, we followed the single answer evaluation method in which we asked the GPT-4 model to judge the answers generated by our models. We generated answers for the questions in the following four languages: Nepali, Persian, Sanskrit, and Maithili. An example of the output generated from the model is depicted in the Fig. 1 (Right).

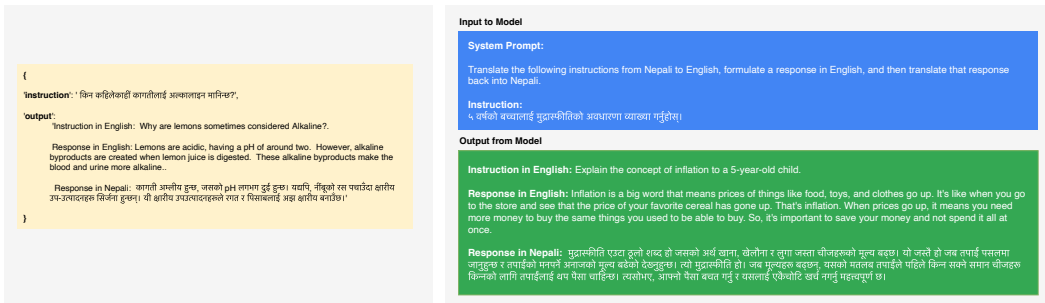


Figure 1: Left: An example of a data point in the TaCo Dataset. Right: Instruction and the response generated using TaCo method.

As we can see, the input to the model is in the native language, with the system prompt provided in English. This is because the underlying model being used was fine-tuned in the English language and is therefore capable of understanding the system prompt and acting accordingly. In the model’s response, we can observe that the instruction in Nepali is initially translated to English, and then a response is generated in English followed by a response in Nepali. One of the main challenges faced in this approach is the token limit while generating responses. Open-ended question responses are generally longer, and when implemented using the TaCo method, need to be first generated in English and then translated into the respective language, often exceeding the maximum token limit in the LLaMA model.

This typically results in longer English response generation and incomplete response generation in the native language due to model token limitations. We resolved this by appending a simple instruction to the prompt, specifying that the model response should not exceed six sentences. After exhaustive trials, we found that six sentences is the sweet spot. Moreover, we also modified the system prompt in the evaluation benchmark by adding an extra line: ‘Do not allow the length of the responses to influence your evaluation’, which will make the GPT-4 not judge based on the length of the answer.

## 4 RESULTS

Our Vicuna Benchmark results are summarized in Table 1, detailing performance by category and language. The average scores were 88% for Nepali, 80% for Sanskrit, 82% for Maithili, and 84% for Persian across nine categories. Except for math, where the Sanskrit model scored the lowest, all TaCo models achieved scores above 70%. We excluded coding from the evaluation since programming languages are usually in English, allowing for a more accurate assessment of model performance.

To evaluate the effectiveness of the TaCo method, we conducted a comparative analysis by instruction-tuning the Guanaco-33B model on two languages of varying resource availability—Nepali, a low-resource language, and Persian, a high-resource language. The instruction-tuning was performed utilizing translations of the Alpaca-52K and Dolly-15K datasets into the respective languages, maintaining consistency in hyperparameters across both models. The results, as illustrated in Table 1, demonstrate that our TaCo method markedly surpasses traditional approaches, nearly doubling the average performance score.

Table 1: Comparison of scores (in %) of TaCo models vs. Instruction-Tuned (IT) models.

Category	Nepali	Sanskrit	Persian	Maithili	Nepali-IT	Persian-IT
Common Sense	85.5	83.0	86.0	78.0	40.0	63.0
Counterfactual	89.0	83.0	84.6	83.0	18.0	46.0
Fermi	77.0	77.0	82.0	75.0	31.0	30.0
Generic	90.0	91.3	90.0	93.0	63.0	44.0
Knowledge	95.5	94.2	92.5	88.0	47.0	52.5
Math	96.7	33.3	63.3	70.0	43.3	56.7
Roleplay	83.5	89.0	86.0	78.0	36.0	38.5
Writing	88.0	89.5	86.5	93.0	36.0	41.0
<b>Overall Average</b>	<b>88.1</b>	<b>80.0</b>	<b>83.9</b>	<b>82.2</b>	39.3	46.5

## 5 DISCUSSION

The LLMs performance increases with their sizes and is capable of solving the multilingual grade school math problems using chain-of-thought process, which can also be applied to other common-sense problems (Shi et al., 2022; Wei et al., 2022b). Similarly, we initially experimented with few-shot chain-of-thought examples in our base Guanaco-33B model, but the model was unable to generate non-English response. One of the probable reason could be the lack of exposure to the language during pretraining and finetuning. Nonetheless, further LoRA instruction tuning revealed the model’s capability to generate responses in that particular language. But instruction-tuned model, capable of generating responses in non-English languages, encountered several issues such as broken grammar, lack of information (like facts, history, and significant events), hallucinations, and repetitive sentences during generation. However, most of these issues were resolved when we implemented our proposed TaCo method. This improvement is evident in our average results, which display an accuracy of 80% or above for each language and an average accuracy of 80% in the common sense category across all four languages.

The curriculum learning strategy emphasizes starting with basic principles and advancing to more complex topics. We adopt this approach in teaching models to respond in English before tackling low-resource languages, using the state-of-the-art Guanaco-33B model. This mirrors human second-language learning—initially analyzing and responding in one’s native language before switching to English. Our TaCo method reflects this by first translating to English, generating the response, and then translating back to the non-English language, adhering to a chain-of-thought process that simplifies complex problems through sequential tasks, with translation at both the beginning and end. This approach capitalizes on the pre-training knowledge of the English language to generate responses in other languages, utilizing its understanding to provide more accurate and factual information while minimizing errors. Consequently, it achieves an average accuracy of 92% in the knowledge category across four languages.

Despite its proficiency in handling non-English questions with an impressive 83% success rate by GPT-4 standards, the model faces significant limitations. One major issue is the token limit, as responses in both languages often exceed the allowed number of tokens, restricting the length of the responses. Furthermore, the model’s creativity in the native language diminishes, limiting its ability to craft poetry or rhymes in the target language, as it operates primarily from an English perspective. Another challenge is the increased token count and time required for generating responses in non-English languages, which we observed to be particularly true for the four languages we tested. This not only increases costs but also the time needed for response generation.

## 6 CONCLUSION AND FUTURE WORK

In this study, we introduced TaCo, a novel approach for crafting multilingual models through translation in the chain-of-thought process, impressively impacting GPT-4 with an 82% accuracy on the Vicuna Benchmark dataset. TaCo effectively generates multilingual models at a reasonable cost, with model weights and datasets now publicly available. Despite its advancements, the study recog-

nizes limitations, including a lack of robustness, absence of toxicity tests, and challenges posed by token limits. The large size of our released models indicates that future efforts will aim for efficiency in smaller models. In conclusion, the emergent behavior of LLMs, amplified by translation in the chain-of-thought process, can enable multilingualism within these models.

## REFERENCES

- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.
- Angana Borah. Translationese: A brief introduction. <https://anganaborah.medium.com/translationese-a-brief-introduction-db4575c0ee69>. Accessed: 16-Sep-2022.
- Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6170–6182, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.427. URL <https://aclanthology.org/2022.acl-long.427>.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.
- Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Un-supervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm, 2023. URL <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.
- Philipp Dufter and Hinrich Schütze. Identifying elements essential for bert’s multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437, 2020.
- Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*, 2021.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint arXiv:2305.12182*, 2023.

- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2, 2019.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations—democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.
- Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 1318–1326, 2011.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*, 2023.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*, 2023.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.
- OpenAI. Gpt-4 technical report, 2023.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pp. 612–618, 2017.
- Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, 2006.
- tloen. alpaca-lora, 2023.
- UNESCO (IESALC). A decade to prevent the disappearance of 3,000 languages. <https://www.iesalc.unesco.org/en/2022/02/21/a-decade-to-prevent-the-disappearance-of-3000-languages/>, 2022. Accessed: 2023-04-01.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.
- Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*, 2023.

Table 2: Translation Evaluation Metrics for Low Resources Languages

Language	BLEU	CHRF++	TER
Sanskrit	65.23	84.62	19.43
Nepali	69.68	87.37	15.02
Persian	62.42	80.72	20.61
Maithili	63.65	84.88	19.58

## A APPENDIX

### A.1 TRANSLATION EVALUATION METRICS

We present the translation evaluation metric scores for three low-resource languages—Sanskrit, Nepali, and Maithili—alongside Persian, a high-resource language, in Table 2.

### A.2 EVALUATION LOSS FOR TACO MODELS

In Fig. 2, we plot the evaluation loss for four models. As can be seen, there is a steady decline in the loss throughout the steps in all four models.

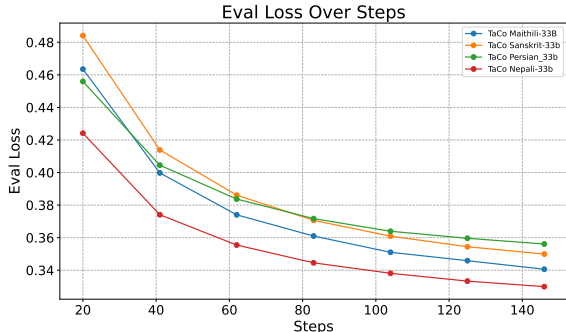


Figure 2: Evaluation Loss over steps for four different TaCo models.

### A.3 COMPARISON OF TACO MODELS VS INSTRUCTION-TUNED MODEL

In Fig.3, we plot the scores of categories from each model. We observe that the instruction-tuned-only models perform inferiorly compared to the TaCo Models. In Fig. 4, Fig. 5, and Fig. 6, we illustrate responses from the instruction-tuned model in Nepali alongside those from the TaCo Nepali model. We also provide a Google translation in English for each response. We appended the instruction *'Your response must be less than 6 sentences'* to ensure that the model’s response does not exceed the token limit. In the responses from the instruction-tuned-only model, it can be observed that the model begins to provide an answer but starts repeating the same answer after a few sentences. In contrast, the responses from the TaCo models are concise and to the point.



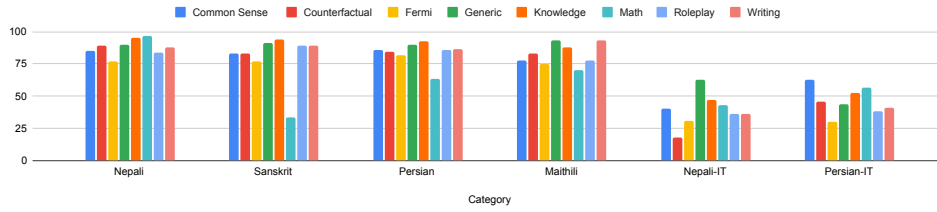


Figure 3: Comparison of Vicuna Benchmark scores by categories for different TaCo models and the Instruction Tuned (IT) model.

**Instruction: म कसरी मेरो समय व्यवस्थापन कौशल सुधार गर्न सक्छु?**

Instruction-tuned in Nepali Language	Finetuned using TaCo Method in Nepali Language
<p><b>Response:</b></p> <p>तपाईंले आफ्नो समय व्यवस्थापन कौशल सुधार गर्न सक्ने थुप्रै तरिकाहरू छन्: 1. स्पष्ट लक्ष्यहरू सेट गर्नुहोस्: तपाईंले आफ्नो समय कसरी व्यवस्थापन गर्ने भने बारे स्पष्ट र संक्षिप्त लक्ष्यहरू सेट गरेर सुरु गर्नुहोस्। यसले तपाईंलाई आफ्नो समय कसरी व्यवस्थापन गर्ने भने कुरामा ध्यान केन्द्रित गर्न मद्दत गर्दछ। 2. प्राथमिकता सेट गर्नुहोस्: तपाईंले आफ्नो समय व्यवस्थापन गर्ने सबैभन्दा महत्त्वपूर्ण कुराहरू पहिचान गर्नुहोस् र तिनीहरूलाई प्राथमिकता दिनुहोस्। यसले तपाईंलाई आफ्नो समय कसरी व्यवस्थापन गर्ने भने कुरामा ध्यान केन्द्रित गर्न मद्दत गर्दछ। 3. प्राकृतिक ब्रेकहरू लिनुहोस्: प्राकृतिक ब्रेकहरू लिनु भनेको तपाईंलाई आफ्नो समय व्यवस्थापन गर्ने प्रक्रियालाई अझ प्रभावकारी बनाउन मद्दत गर्न सक्छ। यसले तपाईंलाई आफ्नो समय व्यवस्थापन गर्ने प्रक्रियालाई अझ रमाइलो बनाउन मद्दत गर्दछ, जसले तपाईंलाई आफ्नो लक्ष्यहरू प्राप्त गर्न उत्प्रेरित गर्न सक्छ। 4. समय ट्याक गर्नुहोस्: समय ट्याक गर्दा, तपाईंले आफ्नो समय कसरी व्यवस्थापन गर्ने भने कुरामा ध्यान केन्द्रित गर्न मद्दत गर्न सक्छ। यसले तपाईंलाई आफ्नो समय व्यवस्थापन गर्ने प्रक्रियालाई अझ प्रभावकारी बनाउन मद्दत गर्दछ, जसले तपाईंलाई आफ्नो लक्ष्यहरू प्राप्त गर्न उत्प्रेरित गर्न सक्छ। 5. तपाईंको समय व्यवस्थापन प्रविधि अडिमाइज गर्नुहोस्: तपाईंको समय व्यवस्थापन प्रविधि अडिमाइज गर्न सक्ने थुप्रै तरिकाहरू छन्। यसले तपाईंलाई आफ्नो समय व्यवस्थापन गर्ने प्रक्रियालाई अझ प्रभावकारी बनाउन मद्दत गर्दछ, जसले तपाईंलाई आफ्नो लक्ष्यहरू प्राप्त गर्न उत्प्रेरित गर्न सक्छ। तपाईंको समय व्यवस्थापन कौशल सुधार गर्न सक्ने थुप्रै तरिकाहरू मध्ये एक भनेको तपाईंलाई आफ्नो समय व्यवस्थापन गर्ने प्रक्रियालाई अझ प्रभावकारी बनाउनको लागि तपाईंको समय व्यवस्थापन प्रविधि अडिमाइज गर्नु हो। यसले तपाईंलाई आफ्नो समय व्यवस्थापन गर्ने प्रक्रियालाई अझ प्रभावकारी बनाउन मद्दत गर्दछ, जसले तपाईंलाई आफ्नो लक्ष्यहरू प्राप्त गर्न उत्प्रेरित गर्न सक्छ। तपाईंको समय व्यवस्थापन प्रविधि अडिमा</p>	<p><b>Response:</b></p> <p>Instruction in English: How can I improve my time management skills? Your response must be less than 6 sentences..</p> <p>Response in English: Here are some tips to help you improve your time management skills:</p> <ol style="list-style-type: none"> <li>1. Prioritize your tasks: Make a list of everything you need to do and rank them in order of importance.</li> <li>2. Set realistic goals: Break down large tasks into smaller, more manageable chunks.</li> <li>3. Eliminate distractions: Turn off your phone, close unnecessary tabs, and find a quiet place to work.</li> <li>4. Use a planner or calendar: Keep track of your appointments, deadlines, and to-dos.</li> <li>5. Take regular breaks: Take short breaks throughout the day to recharge and refresh your mind.</li> </ol> <p>Response in Nepali: यहाँ तपाईंको समय व्यवस्थापन कौशल सुधार गर्न मद्दत गर्ने केही सुझावहरू छन्: 1. आफ्नो कार्यहरूलाई प्राथमिकता दिनुहोस्: तपाईंले गर्न आवश्यक पर्ने सबै कुराको सूची बनाउनुहोस् र तिनीहरूलाई महत्त्वको क्रममा क्रमबद्ध गर्नुहोस्। 2. यथार्थपरक लक्ष्यहरू सेट गर्नुहोस्: ठूला कार्यहरूलाई साना, थप व्यवस्थित टुक्राहरूमा विभाजन गर्नुहोस्। 3. विचलितहरू हटाउनुहोस्: आफ्नो फोन बन्द गर्नुहोस्, अनावश्यक ट्याबहरू बन्द गर्नुहोस्, र काम गर्न शान्त ठाउँ खोज्नुहोस्। 4. योजनाकार वा पात्रो प्रयोग गर्नुहोस्: आफ्नो मेटाघट, समयसीमा, र कार्यहरूको ट्याक राख्नुहोस्। 5. नियमित ब्रेक लिनुहोस्: आफ्नो दिमागलाई रिचार्ज गर्न र ताजा गर्न दिनभरि छोटो ब्रेक लिनुहोस्।</p>

**Google Translation in English**

**Instruction: How can I improve my time management skills?**

Instruction-tuned in Nepali Language	Finetuned using TaCo Method in Nepali Language
<p><b>Response :</b></p> <p>There are several ways you can improve your time management skills: 1. Set clear goals: Start by setting clear and concise goals about how you will manage your time. It helps you focus on how to manage your time. 2. Set priorities: Identify the most important things you manage your time with and prioritize them. It helps you focus on how to manage your time. 3. Take Natural Breaks: Taking natural breaks can help you manage your time more effectively. It helps you make the process of managing your time more enjoyable, which can motivate you to achieve your goals. 4. Track Time: Tracking time can help you focus on how to manage your time. It helps you manage your time more effectively, which can motivate you to achieve your goals. 5. Optimize your time management technique: There are many ways you can optimize your time management technique. It helps you manage your time more effectively, which can motivate you to achieve your goals. One of the most important things to improve your time management skills is to optimize your time management techniques to make your time management process more efficient. It helps you manage your time more effectively, which can motivate you to achieve your goals. Optimizing Your Time Management Techniques</p>	<p><b>Response:</b></p> <p>Here are some tips to help improve your time management skills:</p> <ol style="list-style-type: none"> <li>1. Prioritize your tasks: Make a list of everything you need to do and rank them in order of importance.</li> <li>2. Set realistic goals: Break large tasks into smaller, more manageable chunks.</li> <li>3. Eliminate distractions: Turn off your phone, close unnecessary tabs, and find a quiet place to work.</li> <li>4. Use a planner or calendar: Keep track of your appointments, deadlines, and tasks.</li> <li>5. Take regular breaks: Take short breaks throughout the day to recharge and refresh your mind.</li> </ol>

Figure 4: Example response from models on 'How can I improve my time management skills?'

9



