Machine Translation with Large Language Models: Decoder Only vs. Encoder-Decoder

Machine Translators Team

Abstract

This project, titled "Machine Translation with Large Language Models: Decoder-only vs. Encoder-Decoder," aims to develop a multilingual machine translation (MT) model. Focused on Indian regional languages, especially Telugu, Tamil, and Malayalam, the model seeks to enable accurate and contextually appropriate translations across diverse language pairs. By comparing Decoder-only and Encoder-Decoder architectures, the project aims to optimize translation quality and efficiency, advancing cross-linguistic communication tools.The primary objective is to develop a model capable of delivering high-quality translations that are accurate and contextually appropriate. By leveraging large language models, specifically comparing the effectiveness of Decoder-only and Encoder-Decoder architectures, the project seeks to optimize translation performance and efficiency across multilingual contexts. Through rigorous experimentation and analysis, this project aims to advance the field of machine translation, contributing valuable insights into the effectiveness of different model architectures and paving the way for enhanced cross-linguistic communication tools.

Keywords : Machine Translation, Decoder-only, Encoder-Decoder

1   Introduction

Machine Translation (MT) has witnessed significant advancements with the advent of Large Language Models (LLMs), which have revolutionized the field by offering robust capabilities in processing and translating natural language. These models, such as mT5 and LLaMA 2, vary in architecture from decoder-only designs to more complex encoder-decoder frameworks, each tailored to address specific challenges in multilingual translation tasks. This work delves into the comparative analysis of LLMs across different architectural paradigms: decoder-only (1-1 and 1-many) and encoder-decoder (1-1, many-1, 1-many, and many-many). Our primary objectives are twofold: first, to explore how these models perform in bilingual and multilingual language translation scenarios, and second, to evaluate the effectiveness of encoder-decoder transformer models in Neural Machine Translation (NMT) compared to smaller, decoder-only models when trained under similar conditions. Key considerations include assessing the impact of context length—measured in the number of tokens—on translation quality and efficiency for both architectural setups. By conducting comprehensive experiments and performance evaluations on datasets like FLORES-101 and TED Talks, we aim to provide insights into the optimal use cases and trade-offs associated with different LLM configurations in the realm of MT.

The code available on GitHub for In-context-learning (ICL), Baseline Model Development, and sample notebooks for finetuning at the following link [1].

2   Related Work

([Aharoni et al., 2019](#)) investigated the development of a universal NMT system capable of translating between 103 languages using over 25 billion training examples. The study emphasized the effectiveness of transfer learning for low-resource languages while maintaining high-quality translation for high-resource languages. By exploring complexities such as diverse scripting systems, data imbalance, and model capacity, the research compared multilingual NMT with bilingual baselines. The findings highlighted challenges in scaling models, balancing data distribution, and mitigating domain noise, offering insights for future research in universal machine translation. ([Arivazhagan et al., 2019](#)) explored the limits of multilingual NMT by training models to translate between 102 languages and

---

[1] https://github.com/sujaykumarmag/iasnlp

English. Extensive experiments were conducted using the TED Talks multilingual corpus, revealing that massively multilingual models outperform previous state-of-the-art methods in low-resource settings while supporting up to 59 languages. The study analyzed different training setups and highlighted the trade-offs between translation quality and modeling decisions. Results demonstrated that the multilingual models exceeded strong bilingual baselines, indicating promising directions for future research in massively multilingual NMT.

The study by (Zhu et al., 2023a) systematically investigated the performance and factors affecting LLMs in multilingual machine translation, finding that models like GPT-4, despite outperforming the strong supervised baseline NLLB in 40.91% of translation directions, still lag behind commercial systems like Google Translate, particularly for low-resource languages . They used the FLORES-101 dataset to benchmark translation quality and evaluated eight popular LLMs, including ChatGPT and GPT-4 . Their findings also highlighted that cross-lingual exemplars provided better guidance for low-resource translations than same-language exemplars. (Devlin et al., 2018) introduced BERT, a Bidirectional Encoder Representations from Transformers, which pretrains deep bidirectional representations from unlabeled text, conditioning on both left and right context across all layers. This design allows BERT to achieve state-of-the-art results on various natural language processing tasks with minimal task-specific architecture changes. BERT significantly improved performance on tasks such as GLUE, MultiNLI, and SQuAD v1.1, demonstrating absolute improvements in accuracy and F1 scores across different benchmarks (Devlin et al., 2018). The Transformer architecture introduced by (Ashish, 2017) in 2017 revolutionized natural language processing (NLP) by employing self-attention mechanisms. This discovery transformed NLP and established a foundation for subsequently developed language translation models.

3 Proposed Methodology and Experimental Results

The proposed methodology aims to evaluate and compare the performance of Encoder-Decoder and Decoder-only models in natural language processing tasks. This methodology is structured into several key phases, including data preparation, model design, training, and evaluation. Each phase is

{
    "1": "After submitting your application, you should receive a registration certificate within several business days. आवेदन जमा करबाक बाद, अहाँकें किछु व्यावसायिक दिनक भीतर एक टा पंजीकरण प्रमाण-पत्र भेटि सकैत अछि।"
    "2": "Early pregnancy bleeding is usually from a maternal source, rather than a fetal one, प्रारंभिक गर्भावस्था में रक्तस्राव आमतौर पर भ्रूण के बजाय मातृ स्रोत से होता है।",
    "3": "The bride's father symbolically offers to the bridegroom a cow as a present. एहन रोगीक लेल, जानिक संक्रमण आ कैंसरक नैदानिक संदेह न्यून रहैत अछि, एक्स-रे एकटा कम महग प्रारम्भिक विकल्पक अछि आ प्रयोगशालाक अध्ययनक सङ्ग एकर व्याख्या होइत अछि।",
}

Figure 1: A sample prompt to the In-Context Learning

described in detail below.

3.1 Incontext Learning

Our first approach in the Machine Translation is In-Context Learning using Few Shot Learning. A brief description about In-Context Learning and its working is given below.

In-Context Learning is a way to use language models to learn tasks given only a few examples (Stanford) and it a prompt engineering tasks for Few-Shot Learning. We give examples of Machine Translation pairs (<X>=<Y>) with a template T, where X is a Source Sentence and Y is a Target Sentence (Zhu et al., 2023b).
The In-Context Exemplars include <X>=<Y> which is a Strong recipe for best outputs from the model as said by Wu et al. (Wu et al., 2023). A prompt P is defined as $T(X_1, Y_1) \oplus T(X_2, Y_2) \oplus ... \oplus T(X_n, Y_n)$ where $\oplus$ refers to concatenation and n being the number of samples (Zhu et al., 2023b).

Our approach to In-Context Learning is a 3-shot learning in which a prompt to the model is structured based on Figure 1.

For the evaluation of the In-context Learning, we used Hindi, Malayalam, Telugu, Tamil and Marathi languages. A 3-short learning using XGLM and mT5 is used with BLEU metric.

3.1.1 Dataset Used

We used BPCC Wiki MT Dataset(AI4Bharat) which had 16k-50k translation samples. It has English to 22 other Indian Language Pairs with a context length of each sentence pair being 40-200 characters long.

3.1.2 ICL Experimental Results

The Architectures that we used are 1. Decoder Only - XGLM. The reason for XGLM is it generates moderate translation with 500 million parameters and builds bi-lingual mapping between non-English and English (Zhu et al., 2023b). 2. Encoder-Decoder - mT5 because of its capability for multilingual translation (mT5-base) and contains 300 million parameters. A sample reference and its prediction in mT5 and XGLM is given in Figure 2. The experimental results of In-Context

Reference	Predicted (mT5)	Predicted (XGLM)
_After _sub mitting _your _application , _you _should _receive _a _registration _certificate _within _several _business _days .	_अपना _आवेदन _जमा _करने _के _बाद, _आपको _कई _व्यावसायिक _दिनों _के _भीतर _पंजीकरण _प्रमाणपत्र _प्राप्त होना _चाहिए। _हिंदी में	_आवेदन _जमा _कर बा क _बाद , _अ हाँ के ं _कि छु _व्यावसायिक _दिन क _भीतर _एक _टा _पंजी करण _प्रमाण- पत्र _भेट ि _सक ं त _अ छि ।

Figure 2: A sample reference results for the Predicted XGLM and mT5

Language-Pair	BLEU - Decoder Only	BLEU - Encoder-Decoder
eng-hin	**1.096078**	**1.04442**
hin-eng	0.443236	0.384031
eng-tel	0.307206	0.292951
tel-eng	0.379685	0.378953
eng-tam	0.307206	0.292951
tam-eng	0.349053	0.364626
eng-mal	0.307206	0.292951
mal-eng	0.330114	0.364626

Figure 3: Experimental Results of In-Context Learning - Language Pairs and their BLEU Scores

Learning for Decoder-only and Encoder-Decoder Architectures are given in Figure 3.

3.2 Finetuning

In this section, we describe the fine-tuning processes of the mT5 and LLaMA 2 models for multilingual and bilingual machine translation tasks. These models were selected because of their superior performance in natural language processing tasks. We meticulously detailed the workflow and evaluation metrics used to gauge their effectiveness. We began with the mT5 model, an encoder-decoder model fine-tuned for translation tasks involving English to Hindi (en-hi) and English to Hindi-Bengali (en-hi-bg) pairs. The process started with data loading, where we prepared datasets for the specified language pairs. The mT5 model was selected for its robust multilingual capability. We configured the training parameters, including hyperparameters, learning rate, batch size, and optimization algorithms, to ensure an optimal setup for fine-tuning. The model was fine-tuned on the datasets, allowing

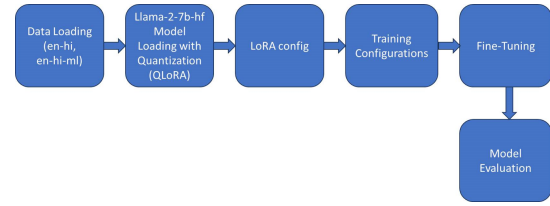Figure 4: Workflow of mT5 Fine-tuning (Decoder only model)

Figure 5: Workflow of Llama2 Fine-tuning (Decoder only model)

it to adapt to the translation tasks. The evaluation metrics used for the fine-tuning task were BLEU, chrF, and TER.

Next, we focused on the LLaMA 2 model, a decoder-only model fine-tuned for English-to-Hindi (en-hi) and English-to-Hindi-Malayalam (en-hi-ml) translation tasks. The process began with data loading for the specified language pairs. The LLaMA-2-7b-hf model was loaded with quantization (QLoRA) to enhance the performance and reduce computational demands. Quantization with QLoRA helps in reducing the model size and computational requirements without significantly compromising the model's performance. We set up LoRA configurations to facilitate efficient fine-tuning, followed by establishing training configurations, including hyperparameters and optimization settings. The model was then fine-tuned on the datasets with evaluation metrics, such as BLEU scores, used to assess performance. The LLaMA 2 model exhibited superior performance in one-to-one (1-1) translation tasks compared with one-to-many (1-many) tasks. Visualization of the loss over time indicated a steady decrease, confirming effective fine-tuning for both the bilingual and multilingual mT5 models. To feed data into the models, we used specific prompt templates: source text #hi#> target text for English-to-Hindi translations and source text #ml#> target text for English-to-Malayalam translations, where #hi#> and #ml#> signify the target language.

3.2.1 Finetuning Experimental Results

For the fine-tuning task, the mT5 model demonstrated significant performance with a BLEU score of 14.1444 and a chrF score of 33.8278 for the bilingual translation task between English and Hindi. This indicates the model's strong capability in handling the en-hi translation pair. Additionally, we extended our experiments to include six different combinations involving English, Hindi, and Bengali for the multilingual mT5 model, exploring

3

(a) mT5 Multilingual eng, hi, bg (all pairs)
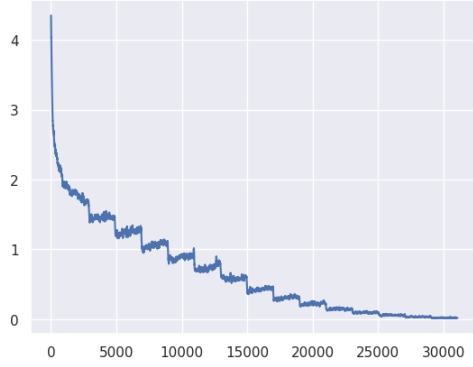
(b) mT5 Bi-lingual eng-hi one-to-one

Figure 6: Loss Convergence plots for Fine-tuning

various language pairings to assess the model's robustness across different translation tasks.

In the case of the LLaMA 2 model, our evaluations revealed that the 1-1 configuration outperformed the 1-many model. This suggests that the one-to-one translation setup is more effective for maintaining high translation quality, outperforming the many-to-one setup in our experiments.

3.3 Baseline Model Development

The baseline models for this study are built upon pre-trained models that have been trained on ex-

Model	BLEU	chrF	TER
Llama2-finetuned-one-many(en-hi)	0.0265	7.1217	94.0950
Llama2-finetuned-one-many(en-ml)	0.0409	6.8530	96.4312
Llama2-finetuned-one-one(En-Hi)	0.0955	9.2282	90.4864
mT5-bi-lingual(en-hi)	11.7107	31.0639	74.1626
mT5-bi-lingual(hi-en)	**14.1444**	**33.8278**	74.7157
mT5 many-many(en-hi)	3.4802	19.6184	84.7821
mT5 many-many(en-bg)	1.0885	16.2382	91.9398
mT5 many-many(hi-bg)	0.7545	15.6990	92.9326
mT5 many-many(hi-en)	5.2237	23.2258	84.6685
mT5 many-many(bg-en)	3.9469	21.5855	86.7228
mT5 many-many(bg-hi)	2.1458	16.9235	88.2083

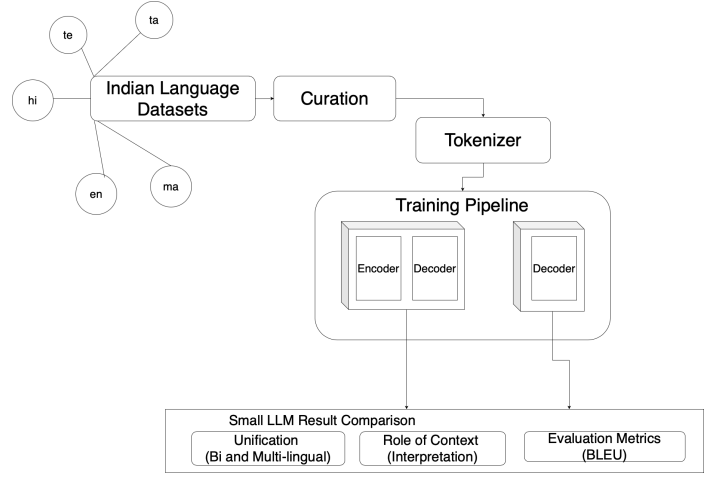Figure 7: Results for Finetuning the models

Figure 8: Proposed Methodology for Baseline Model Comparison

tensive datasets. Specifically, we utilize the mT5 model, which is pre-trained to understand multiple languages. This inherent multi-lingual capability of mT5 is leveraged in our fine-tuning process to adapt the model to the specific tasks at hand.

However, the problem statement for our study remains incomplete. Our primary objective is to compare the performance of Encoder-Decoder and Decoder-only models under similar training conditions. This comparison aims to evaluate the effectiveness of these models in multi-task learning scenarios, particularly in multi-lingual machine translation (MT). To ensure a comprehensive comparison, we will examine not only the overall performance metrics of the models but also their ability to handle different context lengths. Quantitative metrics will be employed to measure and interpret the models' performance with varying context lengths, providing deeper insights into their strengths and limitations. This thorough evaluation will help us understand which model architecture is better suited for multi-lingual and multi-task learning applications.

The proposed methodology of our multi-lingual Encoder-Decoder based architecture can be viewed in Figure 8. We used the same datasets from the IndicTrans2 (Gala et al., 2023) which is in-tune with our Indian baselines study. We go to the next step which is data-curation where we curate the dataset for better translation.

We wanted to write a model from scratch which we able to provide some construction, as using the pretrained model is more black boxed and less interpretable. As creating and experiment-

4

ing with the model comes out with their own challenges, we took some stable baseline models and we equated the parameters. We use XLNet as a base model for implementing the MT task based learning as a Decoder-only model (Wu et al, 2021) and for Encoder-Decoder only model we use the IndicBART as a base model (Dabre et al, 2021) with the shared tokenizer.

Model Name	Trainable Parameters
XLNet Baseline	147,490,318
Indic-BART Baseline	145,339,392

Table 1: Trainable parameters for XLNet and Indic-BART baseline models

We utilized the datasets from IndicTrans2 (Gala et al., 2023), aligning with our focus on Indian language baselines. Following dataset selection, we curated the data to enhance translation quality. Our initial intention was to develop a model from scratch to ensure transparency and interpretability, as pre-trained models often function as black boxes. However, building and experimenting with custom models introduces significant challenges. To address these, we employed stable baseline models and ensured parameter equivalence for fair comparison. For the Decoder-only model, we chose XLNet as the base, implementing it for multi-task learning in machine translation, as demonstrated by (Wu et al., 2021) For the Encoder-Decoder model, we used IndicBART as the foundation, based on the work of (Dabre et al., 2021), Both models shared a common tokenizer to maintain consistency in data processing. This approach allowed us to systematically compare the performance and interpretability of Decoder-only and Encoder-Decoder models under similar conditions, providing insights into their respective strengths and weaknesses in multi-lingual and multi-task learning scenarios. As shown in Table 1, the XLNet Baseline model has 147,490,318 trainable parameters, while the Indic-BART Baseline model has 145,339,392 parameters.

3.3.1 Baseline Experimental Results

In this section, we present the experimental results of our study, focusing on comparisons between one-to-one, one-to-many, many-to-one Encoder-Decoder, and Decoder-only models. Figure 10 illustrates the performance comparison between one-to-one and one-to-many Encoder-Decoder models. The BLEU scores for one-to-many model (Figure
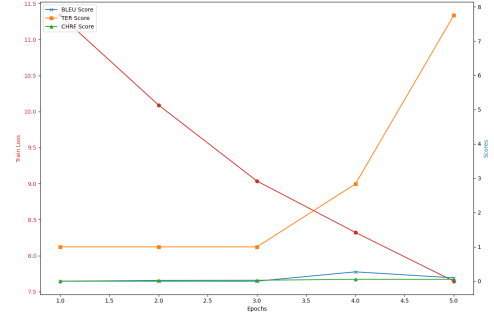
(a) Encoder-Decoder One-to-One Eng to Hindi Results
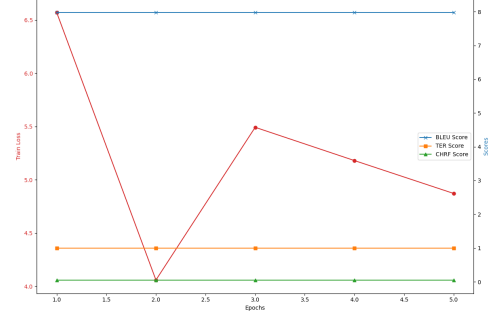

(b) Decoder only One-to-One Eng to Hindi Results

Figure 9: Loss Convergence, BLEU, chrF, TER

10(a)) indicate a slight improvement in translation quality over the one-to-one model (Figure 10(b)), particularly in handling multiple outputs from a single input. Figure 9 presents the results for many-to-one and Decoder-only models. The many-to-one model (Figure 9(a)) shows robust performance in aggregating multiple inputs into a single output, while the Decoder-only model (Figure 9(b)) excels in generating fluent translations with reduced computational complexity.

4 Conclusion and Future Work

The Encoder-Decoder model has demonstrated reliable performance in our experiments, providing trustworthy results. However, the training paradigms for Decoder-only models differ significantly, as they are typically trained on next-word or next-character prediction tasks. The inherent differences in learning paradigms raise the question of how to achieve convergence in multilingual machine translation. Decoder-only models handle the starting positions of the source and target texts separately, posing unique challenges.
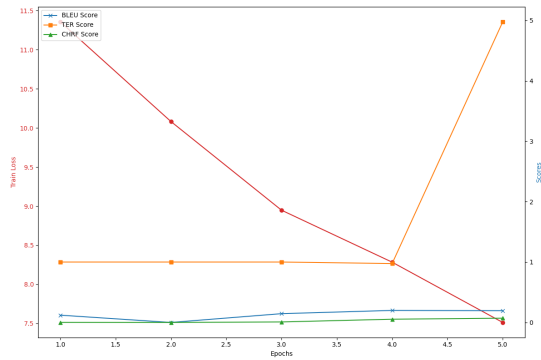
Additionally, the development of novel methods such as Streaming Self-Attention (SSA) marks a significant advancement. SSA enables the model to determine when it has sufficient context from
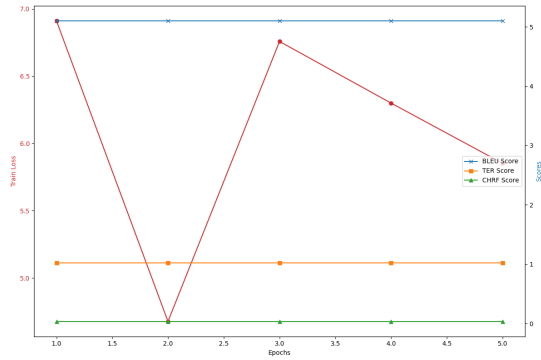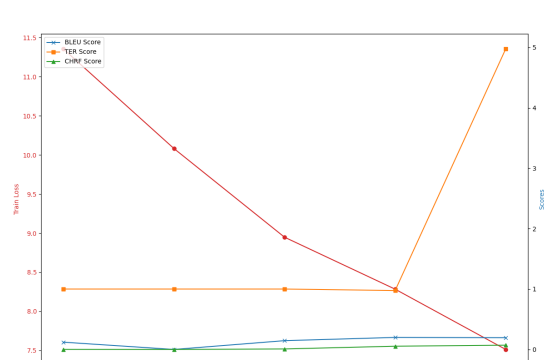
(a) Decoder only One-to-Many English to (Hindi, Marathi) Translation
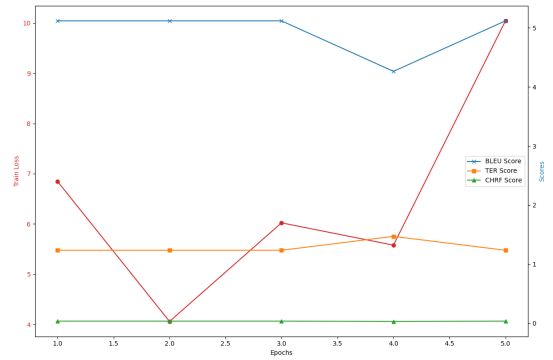
(b) Decoder only Many-to-One (Hindi, Marathi) to English Translation

(c) Encoder-Decoder One-to-Many English to (Hindi, Marathi) Translation

(d) Many-to-One (Hindi, Marathi) to English Translation

Figure 10: Loss Convergence, BLEU, chrF, TER

the original text to begin translating accurately. This technique addresses some of the inherent challenges in translating long texts and could be crucial for improving the performance of Decoder-only models in multilingual settings. Future work should focus on refining these models and exploring ways to harmonize the learning paradigms of Encoder-Decoder and Decoder-only architectures. Further research is needed to fully exploit the potential of SSA and other innovative techniques to enhance translation accuracy and efficiency. This research will contribute to the broader goal of advancing machine translation technologies for multilingual applications.

Acknowledgements

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

AI4Bharat. [link].

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Vaswani Ashish. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:I.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush

6

Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Stanford. [link].

Nier Wu, Hongxu Hou, Ziyue Guo, and Wei Zheng. 2021. Low-resource neural machine translation using xlnet pre-training model. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, pages 503–514. Springer.

Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. 2023. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023a. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023b. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.