The SETU-ADAPT Submissions to the WMT24 Low-Resource Indic Language Translation Task

Neha Gajakos, Prashanth Nayak[a]
Rejwanul Haque[b], **Andy Way**
ADAPT Centre, Dublin City University, Dublin, Ireland
[a]KantanAI, Dublin, Ireland
[b]South East Technological University, Carlow, Ireland
neha.gajakos@adaptcentre.ie,pnayak@kantanai.io
rejwanul.haque@setu.ie,andy.way@adaptcentre.ie

Abstract

This paper presents the SETU-ADAPT's submissions to the WMT 2024 Low-Resource Indic Language Translation task. We participated in the unconstrained segment of the task, focusing on the Assamese-to-English and English-to-Assamese language pairs. Our approach involves leveraging Large Language Models (LLMs) as the baseline systems for all our MT tasks. Furthermore, we applied various strategies to improve the baseline systems. In our first approach, we fine-tuned LLMs using all the data provided by the task organisers. Our second approach explores in-context learning with few-shot prompting. In our final approach we explore an efficient data extraction technique based on a fuzzy match-based similarity measure for fine-tuning. We evaluated our systems using BLEU, chrF, WER, and COMET. The experimental results showed that our strategies can effectively improve the quality of translations in low-resource scenarios.

1 Introduction

Advances in deep learning have led to major improvements in present-day MT systems. However, developing reasonable-quality MT systems for low-resource languages, especially those from the Indic language family, remains a challenge (Pal et al., 2023). India, home to numerous ancient and morphologically rich languages, presents unique obstacles for MT development due to the intricate morphology, syntax, and scarcity of parallel data for many regional languages (Suman et al., 2023; Ahmed et al., 2023). This motivated us to participate in the WMT 2024 Low-Resource Indic Language Translation task and contribute to the advancements in indic MT systems.

Large-pre-trained models are becoming the norm in MT due to their accuracy, scalability, and usage flexibility. Hence, for our experiments we chose LLMs as our baseline MT systems. More specifically, we used IndicTrans2[1] as the baseline for building all our MT systems. We carried out our experiments for Assamese-to-English and English-to-Assamese language pairs.

We conducted experiments applying different methodologies for improving the performance of our MT systems. Our primary approach involves fine-tuning LLMs using all the available data. However, Assamese is a very low-resource language, and obtaining good quality data is challenging. Since there is limited availability of domain-specific parallel data, in our second approach we generated synthetic data by retrieving a large corpus of monolingual data from OPUS[2]. We then performed similarity search in order to identify domain-specific sentences of target language from the generic data and back-translated them into the source language. Our third approach involves investigating in-context learning using few-shot prompting. We augmented the prompt with samples whose source-side is similar to the source sentence to be translated.

The rest of the paper is organised as follows: we discuss related works in Section 2. We detail the data sets used in Section 3. Our models and experimental setups are described in Sections 4 and 5. The results are reported and findings are discussed in Section 6. Section 7 concludes this work and discusses avenues for future work.

[1] https://github.com/AI4Bharat/IndicTrans2?tab=readme-ov-file#indictrans2

[2] https://opus.nlpl.eu/NLLB/as&en/v1/NLLB

Proceedings of the Ninth Conference on Machine Translation, pages 762–769
November 15-16, 2024 ©2024 Association for Computational Linguistics

2 Related Work

In this section, we discuss the papers that are related to our work. Burchell et al. (2022) introduced a framework that differentiates between lexical and syntactic diversity in back translation. Their research highlights that while both types of diversity improve Neural MT (NMT) performance, lexical diversity is more critical. They also demonstrated that nucleus sampling, a method that balances diversity with adequacy, provides superior results for low-resource and mid-resource language pairs.

Ahmed and Buys (2024) introduced the concept of "Synthetic Pivoting" to address the limitations of traditional pivot-based methods, which often face challenges due to structural mismatches between the pivot and low-resource languages. Synthetic Pivoting generates synthetic pivot sentences that better align with the structure of both the source and target languages, resulting in more accurate translations. This method has substantially improved translation quality, particularly for Southern African languages, by simplifying the translation process and effectively utilising high-quality synthetic data.

Suman et al. (2023) focused on improving the translation quality for low-resource Indic languages: Manipuri and Assamese. They leveraged linguistic and scriptural similarities between these languages and Bengali to improve translation outcomes. By utilising pre-trained models on Bengali and incorporating transliteration techniques, they were able to overcome the challenges posed by the limited resources available for Manipuri and Assamese. Their experiments showed that their approaches were effective in improving translation.

Moslem et al. (2023) explored using LLMs for adaptive translation. Their research demonstrated that in-context learning with LLMs enables real-time adaptation to specific terminology and stylistic preferences during inference. They showed that this eliminated the need for extensive fine-tuning. They found that few-shot in-context learning, especially when combined with fuzzy matches from translation memories, can outperform traditional encoder-decoder models regarding translation quality, particularly for high-resource language pairs.

Zhang et al. (2023) investigated the potential of fine-tuning LLMs for MT, focusing on decoder-based models that had not been extensively studied before. They evaluated 15 publicly available LLMs using methodologies such as zero-shot prompting, few-shot learning, and fine-tuning, with a particular emphasis on the QLoRA (Dettmers et al. (2023)) fine-tuning method. QLoRA proved a highly effective technique, reducing memory usage by quantising the model to 4-bit precision and limiting the number of trainable parameters. Their findings showed that fine-tuning LLMs, especially using QLoRA, significantly outperformed zero-shot and few-shot approaches, particularly in document-level translation tasks.

3 Data

We utilised the data provided by WMT organisers for our experiments. The data statistics are detailed in Table 1.

Assamese ↔ English	
Files	Sentences
Train	50,000
Valid	2,000
Test (2023)	2,000
Test (2024) - Blind Test	500

Table 1: *Statistics of the datasets used.*

4 Models used

4.1 IndicTrans2

We used IndicTrans2, a Transformer-based (Vaswani et al., 2023) Multilingual NMT model trained on the BPCC dataset,[3] as our baseline MT system. We used the `ai4bharat/indictrans2-indic-en-1B` and `ai4bharat/indictrans2-en-indic-1B` checkpoints for our systems. For building our MT systems we set the following hyperparameters:

- the data was tokenised to a fixed length of 128 tokens, where sequences longer than 128 tokens were truncated and shorter ones were padded to ensure uniform length across batches,
- the learning rate: 2×10^{-5},
- the batch size: 16,
- the training ran for 3 epochs satisfying our stopping criterion,
- a weight decay of 0.01 for improving the model's generalisation capabilities.

[3] https://ai4bharat.iitm.ac.in/bpcc/

We fine-tuned the model in order to adapt it to the domain and styles of data of Assamese-to-English translation task.

4.2 GPT-4o

GPT-4o (OpenAI et al., 2024) is a language model from OpenAI based on Transformer, which serves as the foundation for many language models today. It comprises multiple layers of self-attention mechanisms and feed-forward neural networks, enabling the model to efficiently process and generate text sequences. The model has been trained on a diverse and extensive dataset, allowing it to capture various linguistic patterns and contextual knowledge. We used GPT-4o for our in-context learning strategy. We used the following set of hyper-parameters for our experiment: (i) the temperature was set to 0.2, which controls the randomness of the output, ensuring more deterministic responses, and (ii) all other hyperparameters were not explicitly set and were set to the default values.

5 Experiments

In this Section we discuss our experiments. As discussed in Section 4, we used IndicTrans2 as our baseline model. We selected this model as the baseline due to its superiority as far as translation performance on low-resource Indian languages like Assamese is concerned (cf. Figure 1). We evaluated our MT models using the test data described in Section 3. We used BLEU (Papineni et al. (2002)), chrF (Popović (2015)), WER, and COMET[4] (Rei et al. (2020)) metrics for evaluation. The following subsections describes our MT systems.

5.1 Assamese-to-English

5.1.1 Primary

Our primary MT system for the Assamese-to-English translation task is the fine-tuned IndicTrans2 model (cf. Section 4). In other words, we fine-tuned the baseline model on the domain data provided by the organisers. Our data sets were detailed in Section 3. We used the same set of hyperparameters that we described in Section 4.

5.1.2 Contrastive System One

as for our second system, we implemented an in-context few-shot learning approach, using which we generated English translations of Assamese sentences using `OpenAI's GPT-4o`. More specifically, for few-shot learning we create prompts for the model with a few samples of translation pairs (source and target) whose source-side is similar to the source sentence we want to translate. We will now explain how we obtained training instances, whose source-side is similar to the sentence to be translated. We first convert all the Assamese training set sentences into dense vector embeddings using `sentence-transformers/all-MiniLM-L6-v2` [5]. The resulting embeddings were then indexed using FAISS,[6] enabling efficient similarity searches to retrieve the most relevant examples.

Furthermore, for each Assamese sentence of the test set, we used FAISS to retrieve the top five closest sentences from the training data based on the cosine similarity of their embeddings. Then, we constructed a detailed prompt for the GPT-4o model using the sentence-pairs that were retrieved from the training set. In Figure 2, we show an example of prompt used for in-context learning.

5.1.3 Contrastive system two

For building our second Contrastive system we used our primary MT model (see Section 5.1.1) as our baseline. We adapted this MT system by fine-tuning it with a synthetic data. In order to create the synthetic data, we used a large English corpus comprising 5,000K sentences from the OPUS repository's NLLB project. We took 500k sentences for that large corpus for our experiment. We further filtered the sentences to include only those whose lengths are of 100 to 500 characters. With this, we omitted very short and excessively long sentences.

To extract domain-similar sentences from the now filtered corpus, we performed a semantic search on it using the validation set. All the corpus sentences were first converted to 768-dimensional dense vector embeddings using the `sentence-transformers-qa-mpnet-base-dot-v1` [7] model. We chose the `qa-mpnet-base-dot-v1` model over the `all-MiniLM-L6-v2` model used in 5.1.2 because it can store more detailed information about a sentence and capture semantic relationships across a wide range of contexts,

[4] COMET version 3.19.1 supports Assamese language.

[5] sentence-transformers/all-MiniLM-L6-v2: `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

[6] FAISS: `https://github.com/facebookresearch/faiss`

[7] sentence-transformers-qa-mpnet-base-dot-v1: `https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1`

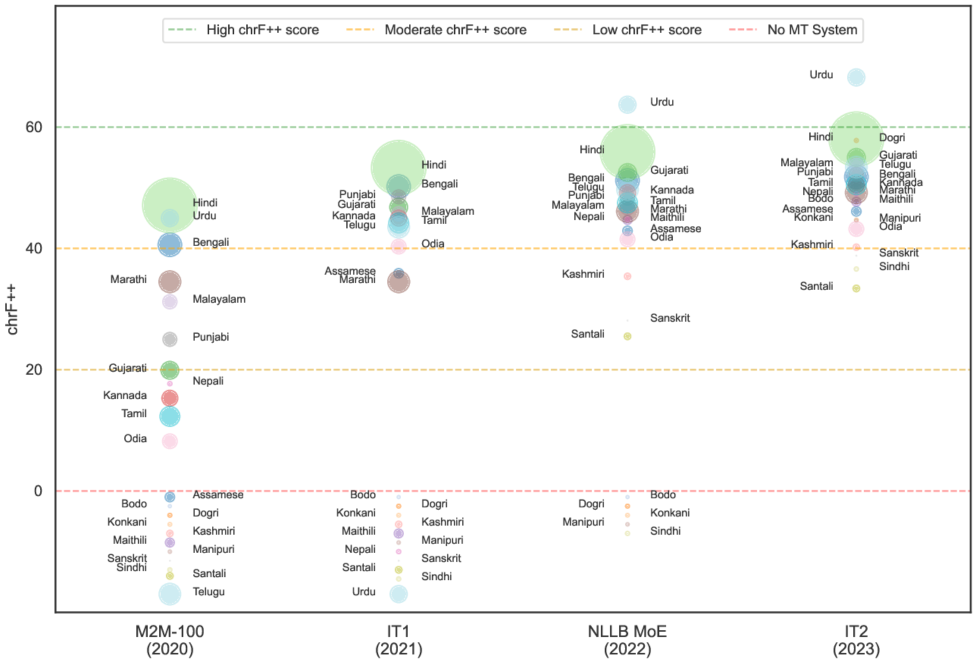

Figure 1: *A visual representation of the advancements in machine translation systems for Indic languages using the IN22-Gen Evaluation set in the En-Indic direction. IT1, IT2 refers to IndicTrans1 and IndicTrans2 respectively. Negative chrF++ values indicate poor translation quality or situations where the translation system fails to generate meaningful or accurate translations. Adapted from (Gala et al., 2023)*

essential for extracting accurate and richer sentence representations from OPUS. These sentence embeddings were then indexed using FAISS. Later, we performed similarity searches by querying the FAISS index with embeddings of the validation sentences. We retrieved the top five nearest neighbours from the corpus for each validation sentence based on cosine similarity. We then removed sentences with similarity scores below 0.2. This process ensured that only the most relevant and contextually similar sentences were selected.

The final fuzzy matching English sentences were then back-translated into Assamese using our primary checkpoint. These new English-Assamese sentence pairs were used to create a new checkpoint by fine-tuning the primary system checkpoint and translation capabilities.

5.2 English to Assamese

5.2.1 Primary

We build out primary systems using an MT approach similar to the one we used in Constrative system one (5.1.2) of the Assamese-to-English translation section, where we utilised `OpenAI's` `GPT-4o` model. The primary difference lies in the prompt structure. In Figure 3, we show the sample prompt that was modified to treat English sentences as inputs and Assamese sentences as outputs.

6 Results

This section presents the evaluation results of the MT systems for both the Assamese-to-English and English-to-Assamese tasks. We performed the initial evaluation using the test pairs from the WMT 2023 dataset. Additionally, we present the results of our evaluation of the blind test set provided by the organisers. The results are reported in terms of BLEU, chrF, WER, and COMET metrics.

To ensure the reliability of our findings using the 2023 dataset, we conducted a statistical evaluation across pairs of models. This involved using bootstrap resampling (Koehn, 2004), calculating BLEU scores, and performing paired t-tests. For each test, we generated 100 bootstrap samples, each containing 100 randomly selected sentences from the dataset without repetition. This method maintains the original dataset's integrity while ensuring diversity in each sample. The results of these statistical analyses are also presented in this section. In all comparisons, we tested the null hypothesis that there is no difference in performance between the systems by calculating p-values. A low p-value (less than 0.05) indicates that we can reject the

```
Give only the final English
    sentence in a single line.
Context:
Assamese 1: <Assamese sentence 1>
Translation in English 1: <
    English translation 1>
Assamese 2: <Assamese sentence 2>
Translation in English 2: <
    English translation 2>
...
Assamese 5: <Assamese sentence 5>
Translation in English 5: <
    English translation 5>

What is the English translation
    for Assamese: <input sentence
    >?
```

Figure 2: *Prompt structure for GPT-4o model: Assamese-to-English*

null hypothesis, suggesting that the observed differences are statistically significant and not due to random variation.

Four models were evaluated for the Assamese-to-English translation task: the baseline, primary, contrastive model one and contrastive model two. The evaluation results are summarised in Table 2.

Model	BLEU ↑	chrF ↑	WER ↓	COMET ↑
B	0.2946	0.5646	0.7000	0.8064
P	0.3418	0.5748	0.6455	0.8086
C1	0.3110	0.5690	0.7035	0.8157
C2	0.3221	0.5724	0.6556	0.8075

Table 2: *Evaluation Results for Assamese-to-English Translation using WMT2023 test pair.*
***B** = Baseline, **P** = Primary (5.1.1), **C1** = Contrastive 1 (5.1.2), **C2** = Contrastive 2 (5.1.3). ↑ indicates higher is better, and ↓ indicates lower is better.*

As shown in Table 2, the primary model (P) outperforms the baseline (B) in all metrics except COMET, where Contrastive system one (C1) achieves slightly higher scores than Contrastive system two (C2) . The BLEU and WER improvements suggest that the primary MT model provides more accurate and fluent translations compared to those by the baseline and contrastive models.

The statistical analysis further supports these findings. When comparing model **B** and **P**, the BLEU score of **P** (0.3418) was higher than that of **B** (0.2946), with a t-statistic of -10.71 and a p-value of 1.72e-09. Similarly, when comparing **P** to **C1** (0.3110), the t-statistic was -10.17 with a p-value of 7.70e-20. In the comparison with **C2** (0.3221), the t-statistic was -8.64, and the p-value was 5.25e-08. Across all comparisons, the null hypothesis was rejected, indicating that **p** consistently performed better than the other models.

```
Give only the final Assamese
    sentence in a single line.
Context:
English 1: <English sentence 1>
Translation in Assamese 1: <
    Assamese translation 1>
English 2: <English sentence 2>
Translation in Assamese 2: <
    Assamese translation 2>
...
English 5: <English sentence 5>
Translation in Assamese 5: <
    Assamese translation 5>

What is the Assamese translation
    for English: <input sentence>?
```

Figure 3: *Prompt structure for GPT-4o model: English to Assamese*

For the English-to-Assamese translation task, two models were evaluated: the baseline and the primary model. The results are summarised in Table 3.

Model	BLEU ↑	chrF ↑	WER ↓	COMET ↑
B	0.1432	0.4948	0.8105	0.8263
P	0.1768	0.4815	0.7457	0.8220

Table 3: *Evaluation Results for English-to-Assamese Translation using WMT2023 test pair.*
***B** = Baseline, **P** = Primary (5.2.1). ↑ indicates higher is better, and ↓ indicates lower is better.*

In Table 3, the primary model shows a noticeable improvement over the baseline in BLEU and WER, indicating better translation accuracy and reduced word errors. However, the chrF and COMET scores are slightly lower than those of the baseline.

The statistical significance tests compares Baseline and Primary (BLEU scores of 0.1432 for the Baseline and BLEU scores of 0.1768 for the Primary) with a t-statistic of -53.11 and a p-value of 1.45e-74. These results clearly indicate that Primary produces better translations than those by the Baseline. The null hypothesis, which assumes no difference in performance between the two systems, was rejected, confirming that the Primary system outperforms the Baseline.

We now present our results on the blind test set provided by the WMT organisers. The results for Assamese-to-English translation in WMT24 Low-Resource Indic Language Translation Task are summarised in Table 4. We observe that contrastive system two generally achieves the best results, leading to 0.3227 BLEU, 0.7563 METEOR, and 0.6573 chrF points, indicating better overall translation quality and semantic accuracy. The primary system closely follows the best-performing system (contrastive system two), performing slightly better in TER (33.56 points) and RIBES (0.3778 points), suggesting that translations require fewer edits, though it falls slightly behind contrastive system two on other metrics (0.3180 BLEU, 0.7537 METEOR, and 0.6551 chrF points). Contrastive system one consistently underperforms the other two systems, with lower scores across all metrics, particularly 39.03 TER and 0.7219 METEOR points.

Model	BLEU ↑	TER ↓	RIBES ↑	METEOR ↑	chrF ↑
P	0.3180	33.56	0.3778	0.7537	0.6551
C1	0.2981	39.03	0.3713	0.7219	0.6437
C2	0.3227	33.63	0.3720	0.7563	0.6573

Table 4: *Evaluation Results for Assamese-to-English Translation (2024).*
***P** = Primary (5.1.1), **C1** = Contrastive 1 (5.1.2), **C2** = Contrastive 2 (5.1.3). ↑ indicates higher is better, and ↓ indicates lower is better.*

The results for English-to-Assamese translation in WMT24 Low-Resource Indic Language Translation Task are summarised in Table 5. We only had one system for this direction, where we obtained 0.1612 BLEU, 65.96 TER, 0.2641 RIBES, 0.3927 METEOR, and 0.5673 chrF points on the test set.

Model	BLEU ↑	TER ↓	RIBES ↑	METEOR ↑	chrF ↑
P	0.1612	65.96	0.2641	0.3927	0.5673

Table 5: *Evaluation Results for English-to-Assamese. **P** = Primary (5.2.1). ↑ indicates higher is better, and ↓ indicates lower is better.*

7 Conclusion

In this work, we presented our MT models developed for the WMT 2024 Low Resource Indic Translation Task, focusing on the Assamese-to-English and English-to-Assamese language pairs. We conducted a comparative analysis using experimental setups to explore strategies such as fine-tuning, back translation, and in-context learning with few-shot prompting. All of these methods demonstrated significant performance improvements in translation.

For our future work, we intend to investigate synthetic pivoting methods for Indic languages and implement QLoRA technique to improve our current in-context learning approach, both discussed in Section 2. We believe that these techniques hold the potential to address the challenges associated with low-resource language translation and further improve the performance of our models.

References

Khalid Ahmed and Jan Buys. 2024. Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158, Torino, Italia. ELRA and ICCL.

Mazida Ahmed, Kuwali Talukdar, Parvez Boruah, Prof. Shikhar Kumar Sarma, and Kishore Kashyap. 2023. GUIT-NLP's submission to shared task: Low resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 935–940, Singapore. Association for Computational Linguistics.

Laurie Burchell, Birch, and Alexandra Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki 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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Dhairya Suman, Atanu Mandal, Santanu Pal, and Sudip Naskar. 2023. IACS-LRILT: Machine translation for low-resource Indic languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 972–977, Singapore. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.