NusaMT-7B: Machine Translation for Low-Resource Indonesian Languages with Large Language Models

William Tan **Kevin Zhu**
Algoverse AI Research
william@tan.id, kevin@algoverse.us

Abstract

Large Language Models (LLMs) have demonstrated exceptional promise in translation tasks for high-resource languages. However, their performance in low-resource languages is limited by the scarcity of both parallel and monolingual corpora, as well as the presence of noise. Consequently, such LLMs suffer with alignment and have lagged behind State-of-The-Art (SoTA) neural machine translation (NMT) models in these settings. This paper introduces NusaMT-7B, an LLM-based machine translation model for low-resource Indonesian languages, starting with Balinese and Minangkabau. Leveraging the pre-trained LLaMA2-7B, our approach integrates continued pre-training on monolingual data, Supervised Fine-Tuning (SFT), self-learning, and an LLM-based data cleaner to reduce noise in parallel sentences. In the FLORES-200 multilingual translation benchmark, NusaMT-7B outperforms SoTA models in the spBLEU metric by up to +6.69 spBLEU in translations into Balinese and Minangkabau, but underperforms by up to -3.38 spBLEU in translations into higher-resource languages. Our results show that fine-tuned LLMs can enhance translation quality for low-resource languages, aiding in linguistic preservation and cross-cultural communication.

1 Introduction

Indonesia is home to 726 recorded regional languages, accounting for about 10% of the world's languages [a20, 2024a]. While the official language, Indonesian, is spoken by 80.4% of the population, a significant portion of these Indonesian speakers—about 70.9%—are multilingual, often fluent in various regional languages [a20, 2024c]. However, predictions suggest that in 100 years, 90% of these languages will either be extinct or on the verge of extinction [Miyaoka et al., 2007].

Machine translation systems have the potential to preserve endangered languages, serving as crucial tools for conservation efforts and fostering cross-cultural communication. However, low-resource languages, by definition, lack parallel corpora, which are crucial for traditional Neural Machine Translation (NMT) systems that often need millions of sentences for optimal performance. [Goyle et al., 2024]. While bitext mined datasets like CCMatrix and NLLB have extracted more than 1.4 million pairs, for instance, in the Minangkabau to English direction, a substantial portion of these datasets—98.7% in this case—is plagued by noise and mismatched pairs [Zong et al., 2021].

Recent advancements in generative (decoder-only) Large Language Models (LLMs) have shown promise in machine translation. OpenAI's GPT-3.5 and GPT-4 [Brown et al., 2020], along with advanced fine-tuned LLMs like ALMA-R [Xu et al., 2024b] and Unbabel's TowerInstruct [Alves et al., 2024], have demonstrated remarkable performance in high-resource language translation, outperforming state-of-the-art (SoTA) models like NLLB-200 [Team et al., 2022] and commercial translation products like Google Translate [a20, 2024b] in the FLORES-200 translation benchmark. However, for low-resource languages, even very large models like GPT-4 lag significantly behind SoTA models like NLLB-200. Indeed, recent literature indicates that fine-tuned LLMs in machine

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

translation are extremely sensitive to noisy data, easily picking up on erroneous biases and misalignments when noise is introduced [Zhu et al., 2024a]. Thus, the primary bottleneck is likely not the inherent performance of LLMs, but the lack of high-quality, clean translation data typically found in low-resource language datasets.

Our proposed solution aims to bridge this gap. We introduce NusaMT-7B, a model focused on Indonesian low-resource languages, starting with Balinese and Minangkabau. Built on LLaMa2-7B [Touvron et al., 2023], NusaMT-7B incorporates continued pre-training on non-English monolingual data, supervised fine-tuning, data preprocessing for cleaning parallel sentences, and synthetic data generation. We open-source NusaMT-7B on Huggingface[1] and deploy a free translation web application[2] to showcase our model. We also release the training code[3] and compiled dataset[4].

Our findings present three key takeaways:

1. Monolingual pre-training and a cleaner, smaller dataset both contribute to improved performance.
2. Backtranslation, a self-learning approach, boosts performance in LLM-based translation.
3. Our combined methods enable our model to outperform most SoTAs in the FLORES-200 benchmark for translation directions into low-resource languages.

This paper includes a case study on the Balinese language to compare our proposed methods. We then extend our model to another low-resource Indonesian language, Minangkabau, and compare its performance against existing SoTA models.

2 Related Work

LLM-based machine translation has seen several innovative developments in low-resource languages. Previous research has focused on improving translation performance by incorporating human preference feedback [Jiao et al., 2023, Zhu et al., 2024b]. Other approaches have leveraged monolingual data for continued pre-training [Xu et al., 2024b, Alves et al., 2024]. Additionally, LLMs like GPT-4 have been used to clean noisy translation data, closely aligning with human cleaning and boosted performance when used to train NMT systems [Bolding et al., 2023].

In the domain of Indonesian low-resource languages, Komodo-7B-Base is a base model of LLaMA2-7B further pre-trained with Masked-Language Modeling (MLM) on 8.79 billion tokens. It's fine-tuned counterpart, Komodo-7B-Instruct, has been trained on multiple tasks including translation. However, it remains closed-source and has not been benchmarked against state-of-the-art models for low-resource languages, leaving its performance unclear.

Our work, however, focuses on enhancing LLM performance through the integration of multiple paradigms, including parallel corpora cleaning, synthetic sentence pair generation, and monolingual pre-training. Together, we provide a comprehensive comparison against various SoTA models on the FLORES-200 benchmark in directions involving Balinese and Minangkabau.

3 Proposed Method

3.1 Continued Pre-Training and SFT

Most pre-trained LLMs are unfamiliar with low-resource languages, making continued pre-training in the target low-resource language essential for teaching the LLM the linguistic principles of these previously unseen languages [Kuulmets et al., 2024]. However, due to limited GPU resources required to pre-train on billions of tokens, we utilize Komodo-7B-base [Owen et al., 2024] These tokens span a diverse set of multilingual corpora, including school textbooks and news articles from 11 regional Indonesian languages, such as Balinese and Minangkabau.

[1] https://huggingface.co/williamhtan/NusaMT-7B

[2] http://indonesiaku.com/translate/

[3] https://github.com/williammtan/nusamt

[4] https://huggingface.co/datasets/williamhtan/NusaMT

During SFT, we use the same translation prompt as ALMA [Xu et al., 2024a], detailed in Appendix A.1.1. The model is tasked with translating a sentence from a source language to a target language, with the loss computed only on the model's generated tokens. Based on Xu et al.'s ablation study on training objectives for LLMs in machine translation, we employ Causal Language Modeling (CLM) loss for fine-tuning, which predicts the next word based only on the preceding context.

3.2 LLM-based Data Preprocessor

We propose an LLM data preprocessor tasked with (1) determining if two sentences share the same underlying meaning and, if so, (2) cleaning the parallel sentences to improve sentence alignment. We selected GPT-4o mini for this task due to its cost-efficiency and the superior performance in data preprocessing tasks demonstrated by its predecessor, GPT-4 [OpenAI et al., 2023]. Initially, we instruct the LLM on the tasks of data cleaning and aligning parallel sentences. We then use few-shot prompting to provide the LLM with representative examples of data cleaning, as shown in Appendix A.1.2. Batch prompting is used to process multiple parallel sentences in a single prompt to reduce overall token size.

3.3 Backtranslation

Backtranslation is a self-training method used to generate additional training data for SFT. It is a data-efficient method to augment new parallel sentence pairs and generate additional training data on more diverse linguistic structures and contexts. To generate synthetic sentence pairs from a source to a target language, we select sentences from monolingual datasets in the target language. After initially training our primary model with SFT, we run inference to translate the target monolingual data back into the source language. Subsequently, we apply our filtering methods and our LLM cleaner a second time to this new synthetic sentence pair to ensure proper alignment. Finally, we fine-tune the model to translate the source sentence back into the target sentence.

4 Experiments

4.1 Data

Table 1: Parallel sentence counts before and after LLM cleaning across datasets and language pairs including English (en), Indonesian (id), Balinese (ban), and Minangkabau (min)

Dataset	**Before Cleaning**				**After Cleaning**			
	ban ↔ en	**ban ↔ id**	**min ↔ en**	**min ↔ id**	**ban ↔ en**	**ban ↔ id**	**min ↔ en**	**min ↔ id**
NLLB Mined	7.4k	2.2k	5.7k	16.5k	4.4k	1.5k	3.4k	9.9k
NLLB SEED	6.0k	6.0k	6.0k	6.0k	5.8k	5.8k	5.7k	5.8k
BASAbaliWiki	23.4k	36.6k	0	0	18.7k	29.3k	0	0
Bible verses	7.1k	9.3k	8.2k	7.6k	5.9k	7.5k	6.6k	6.0k
NusaX	0.9k	1k	1k	0.9k	0.8k	0.8k	0.9k	0.7k
TOTAL	**44.9k**	**55.2k**	**20.9k**	**31.0k**	**35.6k**	**44.9k**	**16.6k**	**22.4k**

For our parallel data, we aggregated both human-annotated and automatically matched bitext datasets. We initially selected Balinese for our ablation study and subsequently expanded to Minangkabau using the best-performing method for additional benchmarking. Each low-resource language has four translation directions: to and from English and Indonesian. The number of parallel sentences for each dataset is shown in Table 1. It is important to note that all parallel sentences undergo a filtering pipeline as described in Appendix A.2.

We used AllenAI's NLLB bitext dataset [AllenAI, 2024] (licensed under ODC-BY), sourced from metadata released by Meta AI as part of the NLLB project, as it is the largest dataset available for low-resource languages. Additionally, we used the human-annotated NLLB SEED dataset [Maillard et al., 2023] (licensed under CC-BY-SA), which contains 6,062 sentences across English and multiple low-resource languages, including Balinese and Minangkabau.[5] We also extracted Bible verse bitexts

[5]Since SEED does not include Indonesian, and given the SoTA performance for en→id, English SEED sentences were translated into Indonesian with the NLLB-3.3B model for additional bitext [Team et al., 2022].

from Alkitab.mobi [Yayasan Lembaga SABDA (YLSA), 2018] (released under copyright for non-profit scholarly and personal use only), a collection of Bibles translated into regional Indonesian languages, where parallel sentences were generated automatically based on identical Bible line numbers. Finally, we used NusaX (licensed under CC-BY-SA), a parallel corpus annotated by Indonesian language experts across English and 10 Indonesian languages, including Balinese and Minangkabau [Indra Winata et al., 2023]. For Balinese directions, we also sourced BASAbaliWiki (licensed under CC-BY-SA), a Balinese wiki containing articles with translations in Indonesian and English [a20]. In each article, we generated bitext by using LASER3 to find the nearest neighbors of each sentence and create possible sentence pairs, setting a similarity threshold of 0.7.

For monolingual data used in backtranslation, we aggregated sentences from Wikipedia dumps (CC BY-SA) and the Glot500 dataset, which was collected from other existing multilingual datasets (all of which we used were licensed under CC BY-NC or CC BY) [Rogers et al., 2023].

We also report the parallel sentence counts before and after LLM cleaning in Table 1. Across all language pairs, there is a significant decrease in total parallel sentences—most notably from 31k to 22.4k sentences in the min↔id pair. However, in human-annotated datasets like NLLB SEED and NusaX, minimal parallel sentences were filtered out, indicating that the LLM cleaner was proficient in retaining truly aligned sentence pairs.

4.2 Training Setup

In the initial SFT stage, we trained the model across all language directions simultaneously. To reduce GPU memory usage, we utilized Low-Rank Adaptation (LoRA) with a rank of 16, which reduces the number of trainable parameters to only 0.1% (7.7 million from 7 billion parameters) [Hu et al., 2021]. We also used Deepspeed with ZeRO stage 2 offloading using bfloat16 to further optimize memory usage and enable multi-GPU training [SC , 2020]. The dataset was randomly split into training, testing, and validation sets with 90%, 5% and 5% splits respectively. Training was conducted over 3 epochs with a learning rate of 0.002 and a per-device batch size of 10. The best model weights were selected based on the lowest CLM loss on the validation set. For training, two Nvidia RTX 4090 GPUs were rented through the vast.ai cloud GPU platform. Training took 18 hours on our combined dataset, which includes Balinese and Minangkabau.

4.3 A Balinese Case Study

Table 2: spBLEU score comparison of the LLaMA2-7B SFT model with various enhancements, including monolingual pre-training (+ Mono), backtranslation (+ BT), and LLM cleaning (+ Cleaner)

Models	**ban → en**	**en → ban**	**ban → id**	**id → ban**
Llama2-7B SFT	27.63	13.94	27.90	13.68
+ Mono	31.28	18.92	28.75	20.11
+ Mono + BT	33.97	20.27	29.62	20.67
+ Mono + Cleaner	33.23	19.75	29.02	21.16
+ Mono + Cleaner + BT	**35.42**	**22.15**	**31.56**	**22.95**

To study the effects of our proposed method, we compare fine-tuning the base LLaMA2-7B model with the addition of monolingual pre-training, LLM cleaning, and backtranslation methods. We present our findings in Table 2, using the spBLEU metric, which is the traditional BLEU metric applied over text tokenized by the FLORES-200 SentencePiece [Goyal et al., 2022].

The results indicate that the Komodo-7B-base model, with additional monolingual pre-training, achieves substantial gains over the base LLaMA2-7B model across all translation directions. Notably, we observe up to a 45% increase in spBLEU in the Indonesian to Balinese direction. Additionally, we find that the LLM cleaning method alone raises spBLEU scores by an average of 5%. This suggests that a reduced training size with reduced noise can indeed boost model performance, supporting the "less is more" (LIMA) philosophy [Zhou et al., 2023]. We also report a 4.7% increase in spBLEU through backtranslation, demonstrating the LLM's capacity to continue learning through synthetically

generated data. Furthermore, when applying both methods in conjunction, we observe an average performance increase of 13% spBLEU over the Mono + SFT baseline.

4.4 Benchmarking

Baselines. We now evaluate and benchmark NusaMT-7B, our model with the LLM cleaner and backtranslation, additionally trained on Minangkabau language directions. We benchmark the SoTA NLLB-200 models, including the 3.3B, 1.3B, and the distilled 600M variant as well as the very large GPTs from OpenAI—GPT-3.5-turbo, GPT-4, and the latest GPT-4o using zero-shot prompts.

Table 3: spBLEU scores of NusaMT-7B compared against SoTA models (NLLB-600M, NLLB-1.3B, NLLB-3.3B) and large GPT models (GPT-3.5-turbo, GPT-4o, GPT-4)

Models	**ban → en**	**en → ban**	**ban → id**	**id → ban**	**min → en**	**en → min**	**min → id**	**id → min**
GPT-3.5-turbo, zero-shot	27.17	11.63	28.17	13.14	28.75	11.07	31.06	11.05
GPT-4o, zero-shot	27.11	11.45	27.89	13.08	28.63	11.00	31.27	11.00
GPT-4, zero-shot	27.20	11.59	28.41	13.24	28.51	10.99	31.00	10.93
NLLB-600M	33.96	16.86	30.12	15.15	35.05	19.72	31.92	17.72
NLLB-1.3B	37.24	17.73	32.42	16.21	38.59	22.79	34.68	20.89
NLLB-3.3B	**38.57**	17.09	**33.35**	14.85	**40.61**	**24.71**	**35.20**	22.44
NusaMT-7B (Ours)	35.42	**22.15**	31.56	**22.95**	37.23	24.32	34.29	**23.27**

We report our benchmarking in Table 3. For all translations into higher-resource languages, NusaMT-7B scores higher than the GPT models and NLLB-600M, but is either outperformed by NLLB-1.3B or NLLB-3.3B. This could be due to the additional learning that NLLB models transferred from the directions involving similar languages—besides Minangkabau and Balinese—into high-resource languages. However, in translations into Balinese, NusaMT-7B achieves spBLEU scores of 22.1 and 22.9 spBLEU score from English and Indonesian directions, respectively, outperforming all the SoTA models, including the larger NLLB-3.3B by up to +6.69 spBLEU. Similarly, in the Indonesian to Minangkabau direction, NusaMT-7B outperforms NLLB-3.3B by +0.83 spBLEU. Overall, while NusaMT-7B still lags behind SoTAs in translations toward high-resource languages, we observe significant performance gains in translations toward low-resource languages.

5 Conclusion

In this paper, we introduced NusaMT-7B, a large language model fine-tuned for low-resource Indonesian languages, with a focus on Balinese and Minangkabau. Our method combines continued pre-training on monolingual data, SFT, and data manipulation techniques using our LLM cleaner and backtranslation. The results from our experiments demonstrate significant performance improvements in translation quality, particularly in directions toward Balinese. Our findings also support the LIMA hypothesis, showing that a smaller, higher-quality dataset can indeed increase model performance. This study presents a promising direction for enhancing machine translation in low-resource settings, contributing to the preservation of the many endangered languages in Indonesia and beyond.

6 Limitations

There are several limitations to our study. Due to limited GPU resources, we used the Komodo-7B-base model, which restricts us from fully assessing the required content and size of monolingual data for optimal performance. We also did not benchmark against the NLLB-54B Mixture of Experts (MOE) model, NLLB's largest model [Team et al., 2022]. In addition, comparisons with models like NLLB are limited by differences in training data, as our model incorporates additional external sources beyond NLLB's SEED and mined bitext datasets. Our findings are also based solely on the spBLEU metric, which may not fully align with translation performance. Furthermore, the ablation study discussed in 4.3 was conducted only on language directions involving Balinese; therefore, the performance gains from our chosen techniques may not generalize to other low-resource languages. It is also important to note that, compared to NMT models, our model utilizes significantly more parameters (7 billion), and thus is less computationally efficient during training and inference.

References

Basabaliwiki. URL https://dictionary.basabali.org/Main_Page.

Data bahasa - peta bahasa, 2024a. URL https://petabahasa.kemdikbud.go.id/databahasa.php.

Google translate, 2024b. URL https://translate.google.com/?sl=id&tl=en&op=translate.

Indonesia, 2024c. URL https://www.ethnologue.com/country/ID/.

AllenAI. Nllb, 04 2024. URL https://huggingface.co/datasets/allenai/nllb.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, and Martins de. Tower: An open multilingual large language model for translation-related tasks, 2024. URL https://arxiv.org/abs/2402.17733.

Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings, 2019. URL https://aclanthology.org/P19-1309.pdf.

Quinten Bolding, Baohao Liao, Brandon Denis, Jun Luo, and Christof Monz. Ask language model to clean your noisy translation data, 2023. URL https://aclanthology.org/2023.findings-emnlp.212.pdf.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and ... Dario Neelakantan. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 01 2022. doi: 10.1162/tacl_a_00474. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00474/110993/The-Flores-101-Evaluation-Benchmark-for-Low.

Vakul Goyle, Parvathy Krishnaswamy, Kannan Girija Ravikumar, Utsa Chattopadhyay, and Kartikay Goyle. Neural machine translation for low resource languages, 08 2024. URL https://arxiv.org/abs/2304.07869.

Amir Hossein, François Yvon, and Hinrich Schütze. Glotlid: Language identification for low-resource languages, 06 2024. URL https://arxiv.org/pdf/2310.16248v3.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Prasojo, Pascale Fung, Timothy Baldwin, Jey Lau, Rico Sennrich, and Kata Ai. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages, 2023. URL https://aclanthology.org/2023.eacl-main.57.pdf.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Parrot: Translating during chat using large language models tuned with human translation and feedback. *arXiv (Cornell University)*, 12 2023. doi: 10.18653/v1/2023.findings-emnlp.1001. URL https://aclanthology.org/2023.findings-emnlp.1001/.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. Teaching llama a new language through cross-lingual knowledge transfer, 2024. URL https://arxiv.org/abs/2404.04042.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.154.

Osahito Miyaoka, Osamu Sakiyama, and Michael E Krauss. *The Vanishing Languages of the Pacific Rim*. Oxford University Press, 04 2007.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and ... Barret Altenschmidt. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. Komodo: A linguistic expedition into indonesia's regional languages, 03 2024. URL https://arxiv.org/pdf/2403.09362.

OpusFilter: A configurable parallel corpus filtering toolbox, 07 2020. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. doi: 10.18653/v1/2020.acl-demos.20. URL https://aclanthology.org/2020.acl-demos.20.

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors. *Glot500: Scaling multilingual corpora and language models to 500 languages*, volume 1, 07 2023. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.61. URL https://aclanthology.org/2023.acl-long.61.

ZeRO: memory optimizations toward training trillion parameter models, 11 2020. SC '20: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. URL https://dl.acm.org/doi/10.5555/3433701.3433727.

Weiting Tan, Kevin Heffernan holger, and Schwenk philipp Koehn. Multilingual representation distillation with contrastive learning, 2023. URL https://aclanthology.org/2023.eacl-main.108.pdf.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, and ... Jeff Licht. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and ... Thomas Bhosale. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Haoran Xu, Young Kim, Amr Sharaf, Hassan Hany, and Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models, 02 2024a. URL https://arxiv.org/pdf/2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation, 2024b. URL https://arxiv.org/abs/2401.08417.

Yayasan Lembaga SABDA (YLSA). Alkitab yang terbuka (ayt), 2018. URL http://alkitab.mobi/. Copyright © 2018 Yayasan Lembaga SABDA (YLSA). For non-profit scholarly and personal use only.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, and ... Omer Efrat. Lima: Less is more for alignment, 05 2023. URL https://arxiv.org/pdf/2305.11206.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice?, 2024a. URL https://arxiv.org/abs/2404.14122.

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. A preference-driven paradigm for enhanced translation with large language models, 2024b. URL https://arxiv.org/abs/2404.11288.

Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors. *CCMatrix: Mining billions of high-quality parallel sentences on the web*, 08 2021. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). doi: 10.18653/v1/2021.acl-long.507. URL https://aclanthology.org/2021.acl-long.507.

A Appendix

A.1 Prompts

A.1.1 Translation Prompt

```
Translate this from [source language] to [target language]:
[source language]: [source]
[target language]:
```

[Few-shot prompt]:

```
Translate this from English to Balinese:
English: Astaire continued to act in the 1970s.
Balinese: Astaire sasai maakting ring warsa 1970-an.
```

A.1.2 Cleaner Prompt

```
You are an expert in aligning and cleaning parallel sentences in different
languages. You will receive two sentences: one in a source language and
one in a target language.

Your task is:
1. On the first line, respond with "True" if the sentences have the same
meaning, otherwise respond with "False".
2. If the first line is "True", provide the cleaned and aligned sentences
on the second and third lines respectively by fixing syntax errors, removing
noise (such as unnecessary phrases, punctuation or ambiguous
numbers), and normalizing text (e.g., capitalization).

Here are some examples to guide you:
[Few-shot prompt]

Now, clean the following sentence pairs:
[Batch-prompt]
```

[Few-shot prompt]:

```
Indonesian: Dengan harga yang bisa dibilang menengah, apa saja yang ditwarkannya?
Balinese: Suratan puniki nénten indik Kabupatén miwah kota ring Kepulauan Riau.

Indonesian: Bahasa daerah memiliki karakteristik yang unik.
Balinese: (32:2) Basa daerah madue "karakteristik" sane soleh.

False

True
Indonesian: Bahasa daerah memiliki karakteristik yang unik.
Balinese: Basa daerah madue karakteristik sane soleh.
```

A.2 Filtering

We used OpusFilter, a parallel corpus processing toolkit, [Pro, 2020] to implement several simple filtering methods to remove noisy, low-quality or erroneous parallel sentences.

Heuristics. We apply a few simple heuristics to remove likely noisy sentences. Specifically, we set a length filter between 15 and 500 characters to remove sentences with less than approximately three words and those above the maximum 256 tokens (given an approximate 2.5 characters per token). We also specify a word length ratio of 2 and remove sentences containing words longer than 20 characters, as they often indicate errors. Finally, we deduplicate sentence pairs and remove sentences with excessive punctuation or numerical content beyond a 20% threshold.

Language Identification (LID). We intend to preserve only the sentences clearly in the desired source or target language. Thus, we apply the GlotLid V3 LID [Hossein et al., 2024] on both monolingual and parallel corpora, using only sentences with a language score above a 0.9 threshold, therefore removing sentence pairs that may contain ambiguity and noise.

Laser Score. The LASER3 encoder [Team et al., 2022] was chosen due to its availability in all the FLORES-200 languages, including Balinese and Minangkabau, as well as its low error rate compared to other multilingual sentence encoders [Tan et al., 2023]. LASER3 encodes sentences in multiple languages and evaluates the quality of a sentence using xsim [Artetxe and Schwenk, 2019], as shown in 1,

$$\text{score}(x, y) = \text{margin}\left(\cos(x, y), \left(\sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2k}\right)\right) \tag{1}$$

where x denotes the source, y denotes the target sentence embeddings, and NN_k represents the k nearest neighbors of x in other languages. We chose to use the ratio margin function, which is defined as ($\text{margin}(a, b) = \frac{a}{b}$) and set $k = 3$. Then, a threshold of 1.09 was chosen, a slightly higher threshold than NLLB's 1.06, since we only want to select high-quality sentence pairs. Bitext with scores lower than this threshold value is filtered out.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions, detailing the development of NusaMT-7B for low-resource languages, including specific methods like monolingual pre-training, supervised fine-tuning, and data cleaning, which are all validated by the results presented. The abstract also highlights how the methods can be applied to other languages and datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations such as the lack of resources for monolingual pre-training, not evaluating more advanced models, and constraints in assessing the optimal size of monolingual data with Komodo-7B-base. Additionally, it reflects on the scope of the claims, referring to the language directions and metrics used, and notes on the model's computational efficiency in comparison to NMT models.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is empirical in nature and does not propose new theoretical results or proofs, focusing instead on practical implementation and evaluation of machine translation methods.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the datasets, training setup, and we release our model and datasets used, making it feasible for others to reproduce the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the code for model training on github with specific instructions for the environment and training setup. The data sources are outlined clearly and the compiled parallel sentence dataset is released through huggingface.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All critical experimental details, such as the data splits, specific hyperparameters (learning rates, batch sizes, lora rank), and loss functions are thoroughly documented.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper does not provide statistical significance measures like error bars or confidence intervals.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the use of two Nvidia RTX 4090 GPUs and provides estimates of training duration, giving a clear picture of the computational resources required for the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics by ensuring responsible use of data, addressing potential biases, and emphasizing the model's intended use for language preservation without harmful applications.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper briefly discusses the positive impact on language preservation and cross-cultural communication. However, since this paper focuses on foundational research, a large emphasis is not tied to particular applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release assets that pose high risks for misuse, and therefore, the need for specific safeguards is not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper appropriately credits existing datasets and models, mentioning the sources and licenses where applicable, ensuring proper use and acknowledgment of all assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release our model and compiled dataset through huggingface and included details such as training and dataset sources in the README. The code also provides simple documentation regarding the methods to run the training and validation scripts.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or human subjects, so this consideration does not apply.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research does not involve human subjects or require IRB approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.