Lius: Translation Model Based Instructional Linguistic Using Continual Instruction Tuning In Kupang Malay

Joanito Agili Lopo[1], **Yunita Sari**[2], **Guntur Budi Herwanto**[3]
Department of Computer Science and Electronics
Universitas Gadjah Mada
[1]amalopo99@gmail.com
[2]yunita.sari@ugm.ac.id, [3]gunturbudi@ugm.ac.id

Abstract

Large Language Models (LLMs) offer new potential for translation tasks but often experience performance degradation when handling low-resource languages. To address this limitation, we propose an approach for fine-tuning LLMs on a low-resource language, Kupang Malay. Our approach involves designing a set of instructions by leveraging explicit lexical and semantic features from a bilingual dictionary, and introducing Continual Instruction Tuning (CIT), a training paradigm that enables iterative instruction-based training. Experimental results demonstrate that our model, named Lius[1], yields notable improvements over standard instruction-tuned models by outperforming 4–6 points, and surpassing both Neural Machine Translation (NMT) and Multilingual LLM models by 10–13 points on several evaluation metrics. These findings highlight the potential of our approach to mitigate the reliance on large-scale parallel data in low-resource language translation.

1 Introduction

Neural Machine Translation (NMT) represents a significant advancement in machine translation by leveraging Artificial Neural Networks (ANN) to model the conditional probability of a target sentence y given a source sentence x. Architectures such as Multilayer LSTM (Sutskever et al., 2014), RNN Encoder-Decoder (Cho et al., 2014), Transformer (Vaswani et al., 2017), and Multilayer CNN (Gehring et al., 2017) have demonstrated considerable improvements over earlier statistical approaches. However, NMT systems heavily depend on the availability of large-scale parallel corpora, which are predominantly available for high-resource languages like English, French, or Chinese. Meanwhile, many languages lack adequate parallel data and are often unsupported by commercial translation systems such as Google Translate (Hedderich et al., 2021; Haddow et al., 2022).

The case of Indonesia illustrates this disparity. With over 718 identified local languages, only a small subset is represented in NLP resources. For instance, Javanese, despite having over 84 million speakers, is represented by merely 12 million parallel sentence pairs, contrast to Dutch, which has over 400 million in the OPUS (Tiedemann, 2012) corpus. Initiatives such as NusaX and NusaWrites (Cahyawijaya et al., 2023a; Winata et al., 2023) have begun addressing these gaps, yet the collection of parallel data remains time-consuming and resource-intensive. Several challenges, such as a shortage of qualified annotators, dialectal variation, and inconsistent orthographic standards (Aji et al., 2022; Novitasari et al., 2020), complicate both model development and parallel data collection efforts.

In response, researchers have explored monolingual corpora and bilingual dictionaries as alternatives. Monolingual data, which are more readily available, has been used to support unsupervised or semi-supervised NMT training, demonstrating notable improvements in output quality (Baziotis et al., 2020; Chronopoulou et al., 2021). Bilingual dictionaries have also been utilized to replace rare or low-frequency words during training, leading to enhanced translation performance (Duan et al., 2020; Pourdamghani et al., 2019). Furthermore, multilingual training, transfer learning, and pivot-based strategies (Dabre et al., 2020; Leng et al., 2019; Zoph et al., 2016) provide cross-lingual generalization capabilities that enhance adaptability and robustness, especially for low-resource settings. However, these approaches often lack the fine-grained lexical and semantic mappings available in parallel data and struggle to capture the complex linguistic phenomena in specific languages.

The rise of Large Language Models (LLMs)

[1]https://huggingface.co/collections/joanitolopo/translation-models

such as GPT-3 (Brown et al., 2020) has significantly shifted the paradigm in machine translation. Trained on massive and diverse datasets in an unsupervised approach, LLMs exhibit strong generalization capabilities across various NLP tasks. These models are especially attractive for translation tasks as they are not strictly dependent on parallel corpora. Some studies have shown that, even in zero-shot settings, LLMs can perform competitively with traditional NMT systems trained on large parallel datasets (Lyu et al., 2024). Nonetheless, the performance of LLMs tends to degrade when applied to non-English or low-resource languages due to over-reliance on English during pretraining (Robinson et al., 2023; Jiao et al., 2023; Bang et al., 2023). These models often fail to capture the rich lexical and semantic variation present in low-resource languages.

To address these limitations, this research adopts LLMs while proposing an Instructional Linguistics approach using Continual Instruction Tuning (CIT). This approach aims to enhance the model's understanding of lexical-semantic relations, particularly in low-resource settings where such relations are rarely explicit. As a case study, this work focuses on Kupang Malay, a Malay-based creole spoken in the western part of Timor Island, which remains underrepresented in digital NLP resources.

2 Related Work

Instruction Tuning Recent studies have demonstrated that leveraging monolingual and parallel corpora through approaches such as Low-Rank Adaptation (LoRA) on large language models, including Llama-2 (Touvron et al., 2023), MaLA-500 (Lin et al., 2024), and Mistral (Jiang et al., 2023), enables significant improvement on the low-resource languages task, especially when pretraining is aligned with the linguistic characteristics of the target language (Iyer et al., 2024; Lin et al., 2024; Jiang et al., 2023). Instruction tuning has also emerged as a key technique, where cross-lingual supervision is used to align bilingual or multilingual data in a way that enhances generalization across tasks and languages. This has been implemented through frameworks that extract structured instructions auch as word-level alignments using tools like FastAlign (Dyer et al., 2013), and apply them to models such as BLOOMZ (Muennighoff et al., 2023) and XGLM (Lin et al., 2022), achieving measurable improvements in translation quality, often surpassing previous baselines by up to several BLEU points (Cahyawijaya et al., 2023b; Mao and Yu, 2024; Li et al., 2024; Lin et al., 2022; Muennighoff et al., 2023; Dyer et al., 2013).

Prompt Engineering Instead of relying on a heavy training process, some studies have designed instruction-specific approaches for low-resource language translation, aiming to improve model interpretability and output quality. By combining monolingual corpora, bilingual dictionaries, and syntactic patterns, models like BLOOMZ (Muennighoff et al., 2023) and ChatGPT benefit from reduced ambiguity and improved lexical alignment (Guo et al., 2024; Muennighoff et al., 2023). Similar gains have been observed through the use of semantic embeddings such as LASER (Heffernan et al., 2022) to refine word correspondences in specific languages (Merx et al., 2024; Heffernan et al., 2022). Advances in prompt design, including domain-specific, morphologically-informed, and context-aware instructions, have shown to further support model reasoning and translation accuracy across tasks and models like GPT-4 and Mistral (Peng et al., 2023; Zhang et al., 2024a; Huang et al., 2023; He et al., 2024; Jiang et al., 2023).

3 Kupang Malay Language

Kupang Malay is a Malay-based creole spoken in the western part of Timor Island and is widely used as a lingua franca across various ethnic communities. Although it originated through creolization rather than as a native ethnic language, it has evolved into a stable and distinct linguistic system (Rafael, 2019). The language exhibits minimal morphological complexity such as limited to four prefixes, with no suffixes or infixes, and a vowel system similar to Indonesian, excluding the schwa vowel (Jacob and Grimes, 2003). Several dialects such as Air Mata, Alor Malay, and Basa Kupang reflect the diverse vernacular influences of its speakers.

Kupang Malay is classified as a low-resource language in the field of natural language processing (NLP). It lacks substantial parallel corpora in major multilingual datasets such as OPUS (Tiedemann, 2012), WikiMatrix, and CCMatrix (Schwenk et al., 2021a,b). Existing resources for Kupang Malay are sparse and include only a few datasets such as Taxi1500 (Ma et al., 2024), PanLex (Kamholz et al., 2014), and the Bhinneka Korpus (Lopo and Tanone, 2024), each providing limited data coverage for the

language. Despite being spoken by an estimated 5.3 million people in East Nusa Tenggara (NTT), Kupang Malay has yet to receive meaningful technological support. Bridging this gap through NLP development could not only promote digital inclusion for its speakers but also serve as a gateway to revitalizing over 70 other underrepresented languages in the region.

4 Instructional Linguistic

Instructional Linguistics is an approach that applies linguistic theories and principles to design instruction to guide the model during training. There are four prompts that have been derived, i.e., Context-based §4.1, Semantic-based §4.2, Phonetic-based §4.3, and List-Group-Label-based §4.4. The process of instruction generation requires a sentence representation that serves as a proxy for extracting lexical and semantic features relevant to the instruction. We employ the KeyBERT method (Grootendorst, 2021; Sharma and Li, 2019)[2] using Indonesian BERT model [3] (Wilie et al., 2020) to identify the most semantically relevant word in a sentence.

4.1 Context-Based Prompt

The context-based prompt, inspired by the work of Guo et al. (2024) and Merx et al. (2024), incorporates sentence examples in the target language that contain the semantic representation of the input sentence. Specifically, given an input sentence in Indonesian S_{id} and its corresponding semantic representation W^*, the equivalent term in Kupang Malay, denoted as W^*_{mkn}, is retrieved using a bilingual dictionary. However, if no equivalent W^*_{mkn} is found, the original W^* is retained without translation. Subsequently, the context-based instruction retrieves n example sentences from the bilingual dictionary that contain the word W^*_{mkn}. These example sentences, S^1_{mkn}, $S^2_{mkn}, \ldots, S^n_{mkn}$, serve to reinforce the model's understanding of the relationship between the source sentence representation and its contextual realization in the target language.

4.2 Semantic Mapping-Based Prompt

A Semantic-based prompt inspired by the work of Ghazvininejad et al. (2023), which provides bilingual word pairs in the prompt. Firstly, the input S_{id} is represented by a salient word W^*. This word is then searched for the parallel in Kupang Malay as

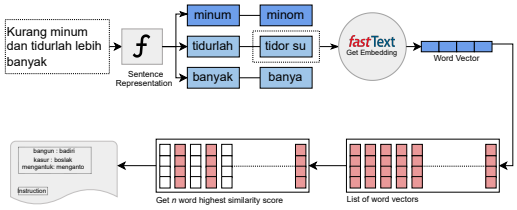


Figure 1: Illustration of semantic mapping-based instruction.

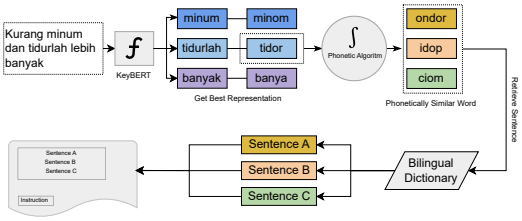


Figure 2: Illustration of phonetic-based instruction construction.

W^*_{mkn} using a bilingual dictionary. Next, W^*_{mkn} is mapped to its vector representation $E(W^*_{mkn})$ using FastText embeddings (Joulin et al., 2017) trained on monolingual Kupang Malay text. The model then searches for the n nearest neighboring words in Kupang Malay based on cosine similarity between $E(W^*_{mkn})$ and all other word vectors $E(W_{mkn})$ in the target language. The retrieved nearest neighbors, $W^*_{mkn_1}$, $W^*_{mkn_2}, \ldots, W^*_{mkn_i}$, are then represented in Kupang Malay. Each of these words is subsequently matched with their Indonesian counterparts using the bilingual dictionary. An illustration of the construction is shown in Figure 1.

4.3 Phonetic-Based Prompt

Phonetic-based prompt enables the model to map sentence representations from the source language to the target language by identifying phonetically similar words, which is expected to improve model performance. This idea is motivated by the work of Atkinson and Raugh (1975) in the pedagogy field. Given Indonesian character c^i_{id} at position i and the Kupang Malay character c^i_{mkn} at the same position, the instruction begins by finding the phonetic representation of the word W^*_{mkn} and applying the phonetic rule R (see Equation 1) extracted from the bilingual dictionary.

$$R\left(c^i_{id}\right) \rightarrow c^i_{mkn} \; \forall i, \; if \; c^i_{id} \neq c^i_{mkn} \qquad (1)$$

[2] https://github.com/MaartenGr/keyBERT

[3] https://huggingface.co/indobenchmark/indobert-large-p2

After the phonetic rule R is obtained, the phonetic representation of the sentence W^*_{mkn} is derived as

$$R(W^*_{mkn}) = \{R(c_1),\ R(c_2), \ldots,\ R(c_n)\}$$

where $\{c_1,\ c_2, \ldots,\ c_n\}$ are the characters in the word. To compute the phonetic similarity between W^*_{mkn} and a candidate word W_{mkn} in the target language, a sequence matching algorithm[4] is used. The similarity score is calculated using Equation 3, where L is the length of the longest matching subsequence between $R(W^*_{mkn})$ and $R(W_{mkn})$. If the similarity score $sim(W^*_{mkn},\ W_{mkn}) \geq \tau$ (with threshold $\tau = 0.7$), the two words are considered phonetically similar. An illustration of construction is shown in Figure 2.

$$d = \max\left(\text{len}(R(W^*_{mkn})),\ \text{len}(R(W_{mkn}))\right) \quad (2)$$

$$sim(W^*_{mkn},\ W_{mkn}) = \frac{L}{d} \quad (3)$$

4.4 List-Group-Label-Based Prompt

This prompt is inspired by Rekrut (1996) in the pedagogical field, which introduces technical vocabulary through the categorization of sentences or objects within instructional content. The prompt generation begins by identifying a list of words related to a specific term, then grouping these words and assigning a label to each group. Specifically, given an input sentence representation $W^* = W^*_1, W^*_2, \ldots, W^*_n$, each representation W^*_n is used to retrieve n closest words $W_{i1}, W_{i2}, \ldots, W_{in}$ in the Kupang Malay language using FastText embeddings. This approach enriches the model with a broader understanding of vocabulary and context, especially when learning cross-lingual semantics. The illustration of this technique is illustrated in Figure 3.

5 Continual Instruction Tuning (CIT)

Continual Instruction Tuning (CIT) is a training paradigm proposed in this study to optimize the ability of Large Language Models (LLMs) in processing and leveraging our Instructional Linguistic §4 approach. CIT draws inspiration from the work of Lee et al. (2024) and Feng et al. (2023), who explored LLM training through curriculum learning, rubrics, and self-correction mechanisms. Unlike

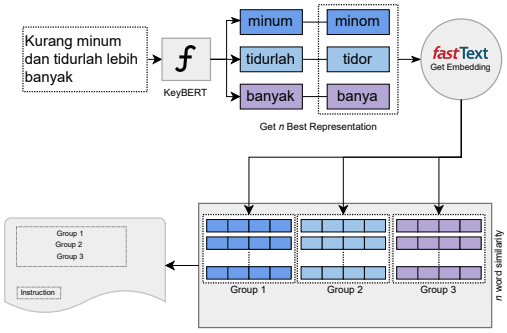


Figure 3: Illustration of List-Group-Label based instruction.

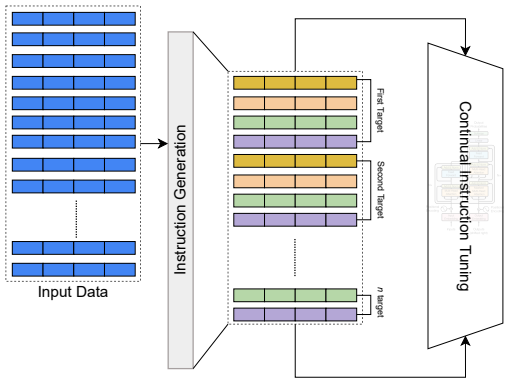


Figure 4: Continual Instruction Tuning adopts a continual training paradigm in which the model learns iteratively, processing one data sample for every four instruction prompts.

conventional fine-tuning, which typically involves a single input–target pair, CIT is designed to continuously incorporate multiple instruction forms for the same input. Specifically, each input x is associated with a set of instructions I_1, I_2, I_3, I_4, all corresponding to the same translation target y. The training process is carried out sequentially, where the model is fine-tuned with the pairs $(x, I_1) \rightarrow y$, followed by $(x, I_2) \rightarrow y$, and so on, until $(x, I_4) \rightarrow y$. Overview of the CIT-based fine-tuning process is illustrated in Figure 4.

6 Experimental Settings

6.1 Data Acquisition

Parallel Data The parallel sentences from Indonesian[5] to Kupang Malay were collected and processed from several sources, namely the Bhin-

[4] https://github.com/python/cpython/blob/3.13/Lib/difflib.py

[5] https://huggingface.co/collections/joanitolopo/kupang-malay-dataset

neka Korpus (Lopo and Tanone, 2024), BibleNLP[6], and The Language Archive[7]. A total of 66,521 sentence pairs were compiled, with 53,217 used for training and 13,304 for testing. Since the translation direction in this study is from Indonesian to Kupang Malay, English sentences in BibleNLP and The Language Archive were first translated into Indonesian using the NLLB model[8]. To ensure translation quality, back-translation was performed and evaluated using SacreBLEU and BERTScore[9]. The statistics of the parallel data and the back-translation results are presented in Table 1. In addition, a bilingual dictionary was compiled from Kamus Pengantar Bahasa Melayu Kupang by Jacob and Grimes (2003), consisting of approximately 3,200 lexical entries.

Source	# Sentences	SacreBLEU	BERTScore
Bible NLP	1,229	35.96	0.94
Bhinneka Corpus	4,000	–	–
The Language Archive	56,740	32.69	0.93
Tapaleuk	4,552	24.34	0.90
Total Parallel Data	**66,521**	–	–
Train	53,217	–	–
Test	13,304	–	–
Total Instruction Data	**227,172**	–	–
Train	212,868	–	–
Test	13,304	–	–
Experience Replay	1,000	–	–

Table 1: Data statistics and back-translation evaluation. For **test instruction data** is randomly chosen over four instructional linguistic during testing.

Monolingual Data This study also employs monolingual data in Kupang Malay to support the development of a FastText model, which plays a crucial role in the instruction construction process within the instructional linguistic approach. The monolingual dataset is collected from several primary sources in Kupang Malay, including Tapaleuk News, the Jakarta Field Station (Station, 1980–2014), Taxi1500-RawData (Ma et al., 2025), as well as collections of poetry (Manhitu, 2005b) and pantun[10] (Manhitu, 2005a). The complete statistics of the monolingual dataset used are presented in Table 16.

[6] https://github.com/BibleNLP/ebible

[7] https://archive.mpi.nl/tla/

[8] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[9] https://huggingface.co/FacebookAI/roberta-large

[10] A form of traditional Indonesian poetry

6.2 Backbone Models

We utilize the Cendol mT5 model (Cahyawijaya et al., 2024), an open-source multilingual LLM supporting Indonesian and 18 local languages. Specifically, three variants of the Cendol mT5 model are employed; ranging from 350 and 580 million parameters to 1.2 billion parameters, respectively. To maintain model's generalization, the Experience Replay technique (Rolnick et al., 2019) is applied. We used 1,000 additional training samples from the original Cendol model's training collection[11] and combined them with the constructed instruction data for training. In addition, due to computational and storage constraints, we used Brain Floating Point 16-bit (BF16) for reduced precision computation, gradient checkpointing to store partial activations before the backward pass, and gradient accumulation over two steps. All training parameters are detailed in Table 18.

6.3 Evaluation Suite

Model Baseline The model baseline consists of comparisons between models trained with standard instructions and those with linguistic instructional prompts. The models using standard instructions were trained using the Cendol-mT5 model in small, base, and large parameter sizes. To ensure comparability, all model training parameters were kept consistent (Table §18) and following the instruction template for the baseline model, adopted from Cahyawijaya et al. (2024) for translation tasks.

Comparison with Other Models We conduct evaluation using both zero-shot and few-shot prompting on multilingual LLMs. Specifically, the models used include BLOOMZ-7B1-MT and mT0-XXL-MT (Muennighoff et al., 2023), Sailor-7B-Chat (Dou et al., 2024), Aya-Expanse-8B (Dang et al., 2024), SeaLLMs-v3-7B-Chat (Zhang et al., 2024b), Cendol-LLaMA2-7b-Inst, and Cendol-mT5-XL-Inst (Cahyawijaya et al., 2024). Furthermore, to the best of our knowledge, the only existing Kupang Malay translation models are the multilingual Madlad400-3B-MT and Madlad400-7B-MT models by Google (Kudugunta et al., 2023). These models were trained on 100 billion sentences across 419 languages, including 25.4 thousand Kupang Malay sentences, all sourced from Bible texts.

[11] https://huggingface.co/datasets/indonlp/cendol_collection_v2

Model	SacreBLEU	chrF++	TER	ROUGE-L
Standard Prompt				
Cendol-mT5-small-tuned	6.62	23.79	102.5	19.97
Cendol-mT5-base-tuned	8.77	26.91	98.08	25.23
Cendol-mT5-large-tuned	**8.97**	**27.30**	**97.13**	**25.94**
Independent Prompt				
Lius-Large-Contextual	**9.47**	**27.78**	96.90	26.55
Lius-Large-Phonetic	9.42	27.63	97.52	26.48
Lius-Large-List-Group-Label	9.37	27.85	**96.41**	**26.86**
Lius-Large-Semantic	9.21	27.70	96.83	26.55
Instructional Linguistic Prompt				
Lius-Small-MT	8.72	26.70	97.96	24.99
Lius-Base-MT	11.63	30.26	92.90	29.75
Lius-Large-MT	**12.99**	**32.01**	**90.36**	**31.55**

Table 2: Performance comparison between Lius and Baseline Models. Ours is denoted as **Instruction Linguistic Prompt**, while **Independent Prompt** is independent instructional linguistic variants.

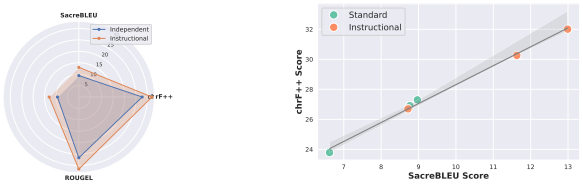
Human Evaluation We conduct further evaluation by human evaluators to maintain the quality of translations. More specifically, evaluators were asked to rate the translation outputs using two evaluation metrics: adequacy and fluency. Each metric follows a four-level scale and conducted by recruiting several native speakers to assess and rate the translation outputs. The rankings were analyzed using inter-annotator agreement, measured with the Kappa coefficient.

7 Result & Analysis

7.1 Comparison with Baseline

Overall, the model performance across all experiments is shown in Table 2. Lius models achieved superior performance across all evaluation metrics, with the 1.2 billion parameters model demonstrating the highest average performance by 2–6 points on each metric and among the Lius variants. Notably, even the smaller variant with 300 million parameters, the model outperformed the standard instruction model of the same size by two points. This indicates that our approach remains effective even for smaller models in low-resource scenarios. In addition, all independent instructional outperformed the standard instruction models. For instance, SacreBLEU scores for these variants ranged from 9.21 to 9.47, surpassing the best-performing standard model. This highlights the benefit of providing task-specific instructions over generic templates.

Semantic and Contextual Analysis Lius improving model performance by approximately 4–6 points across all evaluation metrics. This indicates that our approach enables the model to capture deeper semantic features and comprehend sentence context. For instance, the lower score on the TER metric suggests that fewer edits are required to match the model's translation output with the reference. This can be proved as denoted in Figure 5b, which shows the relationship between SacreBLEU and chrF++ scores, where models with higher BLEU scores also tend to have higher chrF++ scores.



(a) (b)

Figure 5: (a) Lius performance and its independent instructional models using Large model. (b) Correlation between BLEU and chrF++ metrics.

Independent Prompt Contribution We found that each independent instructional is slightly better than the standard baseline, which suggests that it contributes significantly to the performance of the Lius model. To better understand this finding, four example translations were selected and presented in Table 15. There are different variations across the outputs from each instruction type. For example, in the target sentence *"Botong sangka dia pung badan ba'isi, tagal su ampa bulan sonde datang bulan"*, the Semantic and Phonetic instructions signal the word *"kotong"* (we), which semantically aligns with *"botong"* (we). Furthermore, Lius is also capable of capturing the semantic meaning of the sentence. For instance, the phrase *"balakang"* (restroom) is consistently translated as *"kamar kici"* (small room/restroom) in Kupang Malay.

Training Process The training results of each model are illustrated in Figure 11. In general, models exhibit a steady downward trend in loss, indicating that the models are effectively learning and improving their translations over time. In terms of convergence, the Small and Base models tended to converge faster than the Large model, primarily due to having fewer parameters to adjust. In addition, the computational cost and training efficiency of the Large model, denotes in Table 10, requires significantly more storage capacity as well as increased infrastructure during training. For instance, training duration increases exponentially, from 1

hour and 9 minutes to over 6 hours, reflecting a rise in computational complexity.

Copying Behaviour Furthermore, we investigated a model's tendency to directly copy tokens from the input sentence or Copying behaviour, which can degrades translation quality (Liu et al., 2021). We compute both copy accuracy and copy rate for each model, a higher copy value indicates a stronger tendency to replicate input tokens. We inspected that, both metrics exhibit a downward trend as model size increases. For instance, `Lius` with small parameter size shows the highest copy accuracy at 23.62%, while the large one decreases to 20%. Similarly, the Copy Rate decreases from 23.54% in the Small model to 19.38% in the Large model.

7.2 Existing NMT & LLMs Comparison

We conduct comparison using Madlad400-3B-MT and Madlad400-7B-MT to measure the further capabilities of `Lius`. As shown in Table 3, the multilingual NMT models achieve substantially lower scores across all evaluation metrics, whereas the `Lius` model demonstrates up to a sixfold improvement. This implies that although Madlad400 has a larger number of parameters its performance in translating Kupang Malay text remains highly limited.

Similarly, `Lius` model significantly outperforms regional LLMs across all evaluation metrics with significant margin on zero-shot and few-shot prompting. For instance, in few-shot scenario, multilingual models such as mT0-XXL-MT and BLOOMZ-7B1-MT relatively underperform, with SacreBLEU scores of only 2.15 and 1.13 respectively, and extremely high TER values (173.29 and 372.48). This indicates that internally, `Lius` has better semantic understanding, task awareness, and better alignment to translation tasks rather than the other multilingual and regional LLMs.

7.3 Generalization Towards Unseen Data

Unseen Language To evaluate multilingual capability of the model, we conduct an experiment involving Javanese, Sundanese, and Indonesian in a bidirectional translation setup using 100 randomly sampled sentences. The model's performance evaluation is shown in Figure 6. Surprisingly, the evaluation results indicate that the model retains strong multilingual signals. For example, translation from Javanese to Kupang Malay achieves SacreBLEU scores of 10.17, with chrF++ of 26.76 and ROUGE-L of 23.10. This result highlight the model's generalization on cross-lingual transfer.



Figure 6: Bidirectional translation performance of the `Lius` Model.

Unseen Task In addition to its multilingual capability, we investigated the `Lius` multitasking ability. Based on the analysis presented in Table 17, the model demonstrates its multitasking capability despite being fine-tuned specifically for translation tasks. For instance, in the sentiment analysis task, the model consistently classifies text with "negative" sentiment, while in the topic modeling task, the model accurately categorizes the input as "sad". These indicate the model's ability to capture cross-lingual emotional nuance in other languages, such as complex regional language and non-standard lexical variations.

7.4 Robustness Testing

We conducted a series of evaluations on the `Lius` model to assess its robustness under various perturbations applied to the input text. Overall, as shown in Figure 7, the model performs reasonably well on modified sentences, with the performance gap relatively small compared to the unmodified input. These results demonstrate the model's ability to reconstruct structured and contextually appropriate sentences despite input variations.

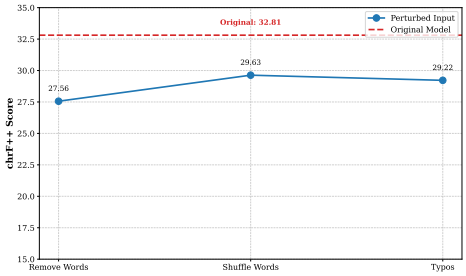


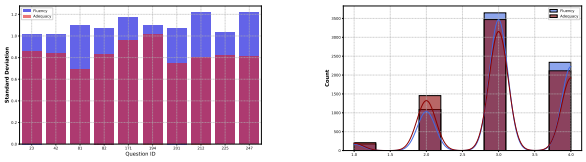
Figure 7: Comparison of model performance on perturbed inputs: Word Shuffling, Word deletion, and Typographical errors.

Model	Zero-shot				Few-shot			
	SacreBLEU	chrF++	TER	ROUGEL	SacreBLEU	chrF++	TER	ROUGEL
Multilingual & Regional LLMs								
BLOOMZ-7B1-MT	3.14	15.03	148.08	6.37	1.13	13.17	372.48	4.69
mT0-XL	4.25	21.00	130.35	12.26	2.25	19.19	163.03	12.07
mT0-XXL-MT	4.86	22.93	114.60	14.40	2.15	18.31	173.29	11.11
Sailor-7B-Chat	3.47	22.36	154.16	13.81	3.47	22.36	154.16	13.81
Aya-Expanse-8B	3.35	23.81	178.34	16.10	–	–	–	–
SeaLLMs-v3-7B-Chat	4.44	21.80	146.95	15.95	1.12	10.88	294.83	4.46
Cendol-LLaMA2-7b-inst	1.87	11.98	122.50	4.31	1.45	5.96	134.57	2.23
Cendol-mT5-XL-Inst	3.72	20.68	143.90	11.00	–	–	–	–
Multilingual NMT								
Madlad400-3B-MT	2.22	20.96	136.81	13.94	–	–	–	–
Madlad400-7B-MT	2.22	20.36	128.79	13.32	–	–	–	–
Lius Models								
Lius-Small-MT	9.20	27.75	98.76	25.55	8.72	26.70	97.96	24.99
Lius-Base-MT	11.88	31.23	92.83	30.52	11.63	30.26	92.90	29.75
Lius-Large-MT	**13.27**	**32.80**	**89.64**	**32.62**	**12.99**	**32.01**	**90.36**	**31.55**

Table 3: Performance comparison of multilingual, regional LLM and NMT, and `Lius` models on zero-shot and few-shot prompting settings. There are two models excluded in Few-shot prompting due to limited computational time.

7.5 Human Evaluation

Quality Rating Distribution Human Evaluation is conducted by involving native speakers to assess the translation results of the model. In general, the average translation scores fall within the range of 3 to 4, presented in Figure 8b, indicating that the majority of the translation outputs are considered reasonably good in terms of both fluency and adequacy. However, the adequacy score distribution appears slightly more dispersed, suggesting that conveying meaning accurately in some translations remains a challenge. Furthermore, some evaluators tend to consistently assign higher or lower score, may reflect subjectivity, and individual preferences. On the other hand, there is greater disagreement among raters in assessing sentence naturalness, as shown in Figures 8a, which the standard deviation of fluency ratings is generally higher than that of adequacy.



(a) (b)

Figure 8: (a) Variability score for fluency and adequacy. (b) Distribution score for translation quality based on fluency and adequacy.

Inter-annotator Agreement In addition to analyzing score distribution and inter-rater variability, the ordinal *Cohen's Kappa* coefficient was also calculated to evaluate the level of agreement among evaluators in scoring each translation output. To mitigate potential bias from evaluators who consistently assign overly high or low scores, score normalization was applied prior to the calculation.

The results indicate that the weighted Kappa value for the adequacy is 0.2099, while for fluency it is 0.1339. According to the interpretation scale by Landis and Koch (1977), both values fall within the slight agreement category, suggesting a low level of agreement among evaluators. We suspected that, internally, Kupang Malay does not yet have a widely recognized standardized guideline, even though it has evolved into a distinct language (Jacob and Grimes, 2003). This lack of standardization allows for divergent perceptions among speakers when evaluating translation quality.

8 Conclusion

This study proposes a method called the linguistic instructional approach to train LLMs based on the Continual Instruction Tuning paradigm, using the Cendol-mT5 architecture. The model outperforms standard instructions, exceeding both NMT models and multilingual LLMs in zero-shot and few-shot prompting scenarios, with performance gaps rang-

ing from 10 to 13 points across various metrics. It retains multilingual abilities and performs well on tasks like sentiment analysis and question answering. Human evaluation, however, shows it still falls short of native-speaker quality, with the absence of standardized Kupang Malay guidelines posing a major challenge.

References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Richard C. Atkinson and Michael R. Raugh. 1975. An application of the mnemonic keyword method to the acquisition of a russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory*, 1(2):126–133.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023a. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023b. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2021. Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Xin Mao, Ziqi Jin, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-East Asia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 424–435, Miami, Florida, USA. Association for Computational Linguistics.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Tao Feng, Zifeng Wang, and Jimeng Sun. 2023. Citing: Large language models create curriculum for instruction tuning. *Preprint*, arXiv:2310.02527.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Maarten Grootendorst. 2021. MaartenGr/KeyBERT: BibTeX.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Exploring very low-resource translation with LLMs: The University of Edinburgh's submission to AmericasNLP 2024 translation task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 209–220, Mexico City, Mexico. Association for Computational Linguistics.

J.A. Jacob and C.E. Grimes. 2003. *Kamus pengantar bahasa Kupang*. Paradigma series B / Centre for Regional Studies, UKAW. Artha Wacana Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual

lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

J. Richard Landis and Gary G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33(2):363–374.

Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. 2024. Instruction tuning with human curriculum. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1281–1309, Mexico City, Mexico. Association for Computational Linguistics.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183, Florence, Italy. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592. Place: Cambridge, MA Publisher: MIT Press.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the copying behaviors of pre-training for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4265–4275, Online. Association for Computational Linguistics.

Joanito Agili Lopo and Radius Tanone. 2024. Constructing and expanding low-resource and underrepresented parallel datasets for indonesian local languages. *Preprint*, arXiv:2404.01009.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Chunlan Ma, Ayyoob Imani, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schuetze. 2025. Taxi1500: A dataset for multilingual text classification in 1500 languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 414–439, Albuquerque, New Mexico. Association for Computational Linguistics.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schütze. 2024. Taxi1500: A multilingual dataset for text classification in 1500 languages. *Preprint*, arXiv:2305.08487.

Yohanes Manhitu. 2005a. Pantun Melayu Kupang oleh Yohanes Manhitu — geocities.ws. https://www.geocities.ws/johnmanhitu2001/PANTUN_MELAYU_KUPANG_YOHANES_MANHITU.html. [Accessed 27-01-2025].

Yohanes Manhitu. 2005b. Yohanes Manhitu Pung Puisi Bahasa Melayu Kupang — geocities.ws. https://www.geocities.ws/johnmanhitu2001/PUISI_PAKE_BAHASA_MELAYU_KUPANG.html. [Accessed 27-01-2025].

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 131–138, Marseille, France. European Language Resources association.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. Translating translationese: A two-step approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3057–3062, Florence, Italy. Association for Computational Linguistics.

Agnes Maria Diana Rafael. 2019. Interferensi fonologis penutur bahasa melayu kupang ke dalam bahasa indonesia di kota kupang. *Jurnal Penelitian Humaniora*, 20(1):47–58.

Martha D. Rekrut. 1996. Effective vocabulary instruction. *The High School Journal*, 80(1):66–74.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *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)*, pages 6490–6500, Online. Association for Computational Linguistics.

Prafull Sharma and Yingbo Li. 2019. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. *Preprints*.

Jakarta Field Station. 1980–2014. Jakarta field station. Item in collection "MPI EVA corpora". The Language Archive. https://hdl.handle.net/1839/3a60a94c-0b69-4513-bf28-de0dbea23471.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024a. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024b. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *Preprint*, arXiv:2407.19672.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Computational Resources

All experiments were conducted using a single 40GB A100 GPU for all model training variations, and a single 8GB GeForce RTX 4060 Ti for evaluation and additional inference resources. The training and evaluation processes took a total of approximately ∼13 days, including two days each for training the small, base, and large model sizes. Additionally, the Weights & Biases platform was used to monitor model training and GPU utilization, providing visualization and model tracking capabilities.

B Robustness Testing Examples

Typographical errors As shown in Table 4, words such as "dan" are altered to "daan", "bajunya" to "bjunya", and "kemudian" to "emudian". Nevertheless, the model is capable of capturing the intended meaning from such noisy input. In some cases, the model successfully maps distorted words to semantically relevant expressions, such as transforming "emudian" into *"abis itu"* in Kupang Malay. This demonstrates the model's capacity to understand and reconstruct meaning despite distortions introduced by typographical noise.

Word deletion As shown in Table 5, the phrase "kau akan mendapatkan tinjuku" is reduced to "kau tinjuku", which drastically alters the intended meaning. The output generated after word deletion reveals that the model struggles with such distortions, occasionally inserting unrelated elements such as *"bini"* and *"matua"*. This illustrates the model's limitations in grasping holistic meaning when faced with incomplete input, leading to outputs that are not only inaccurate but also enriched with hallucinated content.

Original Text	Typos	Output
Setelah itu, baptua melipat koran dan kemudian menyimpan-nya di dalam bajunya kemudian pulang.	Setelah itu, baptua melipat koran daan kemudian menyimpan-nya di dalam bjunya emudian pulang.	*Abis itu, baptua lipat surat ko simpan di dia pung buk* abis itu pulang.

Table 4: Examples of Translations with Typographical Errors

Original Text	Deletion	Output
Biarkan dia pergi atau kau akan mendapatkan tinjuku, Ama Jola sudah ingin melompat pada orang tua.	Biarkan dia pergi atau kau tinjuku, Ama Jola ingin melompat pada tua.	*Biar itu su jalan ko lu puku be pung bini, Ama Jola mo malompa sang matua*

Table 5: Examples of Translations with Word Deletion

Word Shuffling As shown in Table 6, where the word "pada" is randomly misplaced. Nevertheless, the model demonstrates the ability to reconstruct structured and contextually appropriate sentences. Although the translation of the word *"bagitu"* is inaccurate, the model still manages to produce diverse outputs that reflect resilience to structural perturbations in the input.

Original Text	Shuffled	Output
tapi aku tidak berpikir untuk menikah pada saat itu.	tapi aku tidak berpikir pada menikah untuk saat itu.	*tapi itu waktu be son berpikir untuk menika bagitu*.

Table 6: Examples of Translations with Shuffled Input

C Human Evaluation Process

Human evaluation was conducted by involving native speakers to assess the translations produced

by the Lius model. We evaluated based on two main questions: **In your opinion, is the grammar, spelling, and sentence structure of the translation correct?**—which corresponds to the *fluency* metric; and **In your opinion, has all the information in the source text been accurately translated?**—which corresponds to the *adequacy* metric. The evaluation took place over approximately two months, from January 21 to March 21, 2025. The process was carried out through a web-based annotation platform built using the Python programming language[12] and the Flask framework[13]. The user interface of the annotation platform is shown in Figure 9.

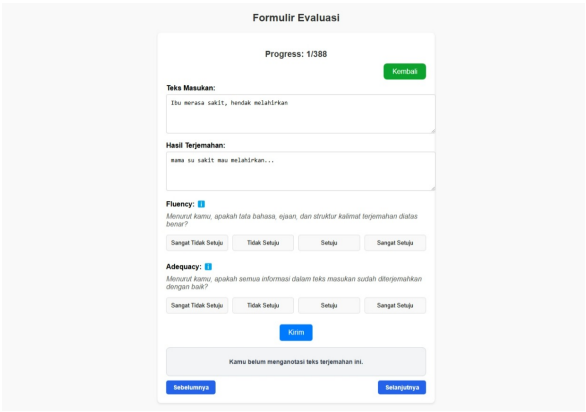


Figure 9: User Interface of Annotation Application

Table 7: Evaluation Scale for Adequacy and Fluency

Score	Adequacy	Fluency
4	All meaning preserved	Very fluent
3	Most meaning preserved	Fluent
2	Some meaning preserved	Moderately fluent
1	Little meaning preserved	Not fluent

The evaluators came from diverse backgrounds, including university students, lecturers, high school teachers, civil servants, and government language development officials. They were mostly residents of Kupang Malay-speaking areas, such as the cities of Kupang, SoE, and parts of Sabu Island. Of the 40 respondents registered in the system, 18 successfully completed the full evaluation. Four evaluators were excluded due to providing identical scores across all translation samples. Examples of the evaluation results for input, translation, and their corresponding fluency and adequacy scores are presented in Table 8.

[12] https://www.python.org/
[13] https://flask.palletsprojects.com/en/stable/

D Instruction Template

The instruction template refers to direct commands used for translation tasks and constitutes a core component of the prompt template, marked with the label **INSTRUCTION**. A total of nine types of instruction template variations were developed, including styles such as formal-direct commands, direct questions, and narrative-style prompts. Each variation consists of 3 to 5 instructions, resulting in a total of 50 distinct instructions. Examples of these instruction template variations are shown in Table 9. The **INPUT** represents the sentence to be translated, **SOURCE** denotes the source language (Indonesian), and **TARGET** refers to the target language (English). For each data instance, the selection of an instruction template is performed randomly.

Examples of single input data entries for each type of instruction can be seen in Table 11, Table 12, Table 13, and Table 14. For instance, in the case of the List-Group-Label-based instruction, the sentence `Aku makan kue tiga potong.` serves as the input text, while sentences such as `Label 1: kala, siri, cop` and `Label 3: fiik, binatang, babika` represent the List-Group-Label-type instruction. The sentence `Berikut adalah kategori kata dalam bahasa Melayu Kupang:` is the prompt template, and `Teks berikut membutuhkan terjemahan dari bahasa Indonesia ke bahasa Melayu Kupang: Terjemahan:` serves as the instruction template.

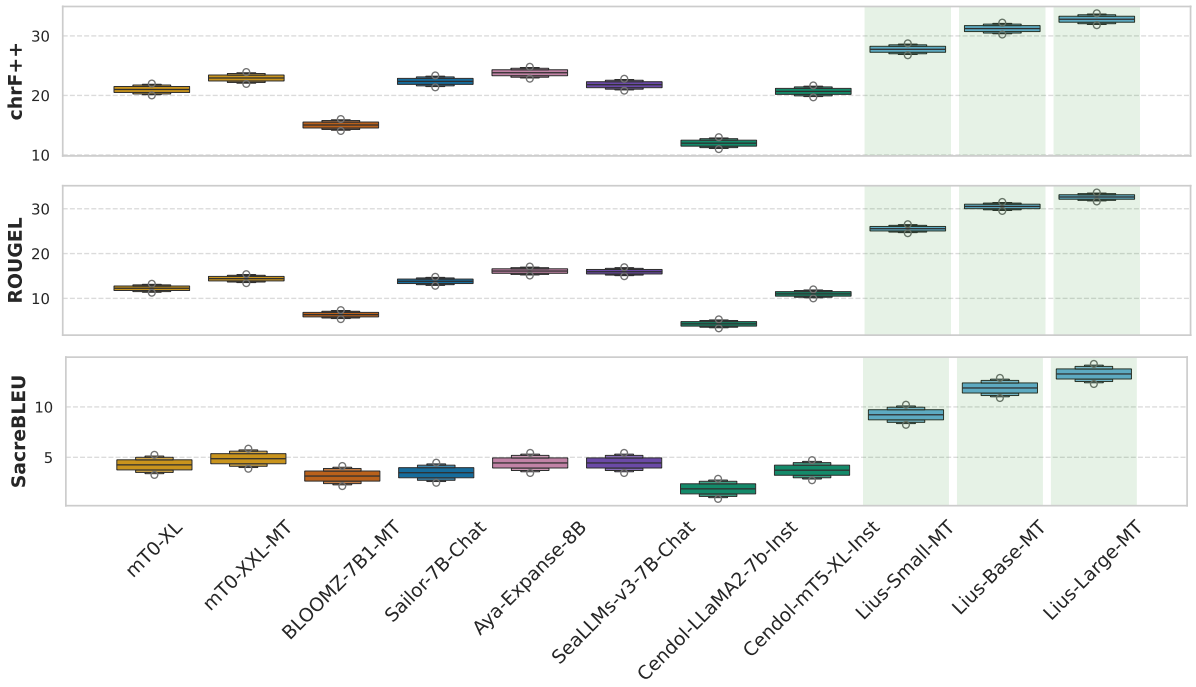


Figure 10: Model Performance on Zero-shot Prompting

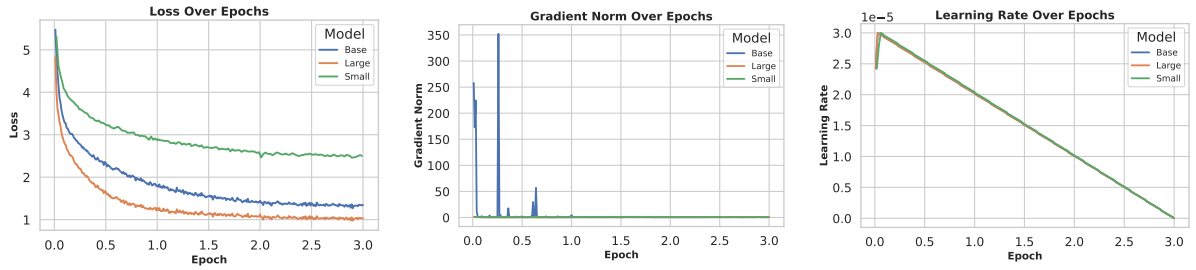


Figure 11: Training Evaluation across Lius Model Variants

ID	Input Text	Translation Output	Fluency	Adequacy
4986	Tolong kalian pikirkan sendiri. Yang mana yang lebih baik? Kami mengikuti perintah bapak-bapak, atau kami mengikuti perintah Tuhan Allah?	Pikir sa. Mana yang lebe bae? Kotong iko bapa-bapa pung parenta, ko kotong iko Tuhan pung parenta?	3	2
4987	ekor kuda	kuda pung ekor	4	4
4988	Saya belum begitu paham dengan maksudnya. Bisakah engkau menerangkannya kembali kepada saya?	Beta balóm mangarti deng dia pung maksud bagini. Bosong bisa carita lai sang beta?	2	3
4989	Anak kecil harus tidur lebih awal.	ana kici harus tidor dolo.	2	4
4990	Ibu merasa sakit, hendak melahirkan	mama su sakit mau melahirkan...	3	3
4991	He! Jalanlah cepat! Jangan lambat seperti pengantin. Nanti kita ketinggalan kendaraan. Akhirnya kita hanya duduk termangu-mangu saja.	He! Jalan capat! Jang lambat sama ke pengantin, nanti kotong katinggal kendaraan, akirnya kotong dudu ragu-ragu sa.	3	2
4992	Waktu dia baru tiba dari Jakarta, dia pergi ke Kolhua untuk menegur-sapa pamannya	baru datang dari Jakarta dia pi Kolhua untuk tegur dia pung om.	3	2
4993	Jangan duduk mengangkang. Nanti anak laki-laki bisa melihat celana dalammu.	Jang dudu mangangkang, nanti ana laki laki liat lu pung calana dalam.	4	2
4995	Marilah kita bermain sembunyi, yuk? Engkau yang mencari dan aku yang bersembunyi, ya?	Mari kotong maen sembunyi, a? Lu yang cari, beta yang sambunyi, a?	4	2

Table 8: Examples of Fluency and Adequacy Input, Translation, and Evaluation Data

Variasi	Prompt
Formal & Langsung	Teks dalam bahasa {**SOURCE**}: {**INPUT**}\nTerjemahkan ke dalam bahasa {**TARGET**}:
Pertanyaan	Apa terjemahan dari teks ini dalam bahasa {**TARGET**}?\nTeks: {**INPUT**}
Konteks	{**INPUT**} adalah teks dalam bahasa {**SOURCE**}.\nSilakan terjemahkan ke bahasa {**TARGET**}.
Spesifik	Mohon bantu menerjemahkan teks berikut dari bahasa {**SOURCE**} ke bahasa {**TARGET**}.\nTeks: {**INPUT**}
Konteks Pengguna	Jika Anda membaca teks berikut dalam bahasa {**SOURCE**}: {**INPUT**} \nBagaimana Anda akan menyampaikannya dalam bahasa {**TARGET**}?

Table 9: 5 Examples of Instruction Template Variations

Table 10: Efficiency Metrics of Lius Model Variants

Aspect	Lius-Small-MT	Lius-Base-MT	Lius-Large-MT
Parameters	350M	580M	1.2B
Storage (GB)	1.13	1.95	3.4
Training Time (hh:mm:ss)	01:09:00	03:22:00	06:35:00
Memory Usage (%)	28	39	78
Inference Time (s)	14.21	22.24	72.35

Input:	Berikut adalah kategori kata dalam bahasa Melayu Kupang: Label 1: kala, siri, cop Label 2: saprei, Mabok sonde jadi-jadi, alas paru Label 3: fiik, binatang, babika Label 4: tiga-tiga, tigapulutiga, tiga Label 5: bet, rabe, be Teks berikut membutuhkan terjemahan dari bahasa Indonesia ke bahasa Melayu Kupang: Aku makan kue tiga potong. Terjemahan:
Output:	Beta makan kue tiga potong.

Table 11: Example of List-Group-Label Based Instruction Data

Input:	Anda memiliki kalimat berikut dalam bahasa Melayu Kupang: - Beta tar paksa lu ko iko sang beta. Ma kalo terpaksa, na iko sa - Dia rasa tar enak dalam hati, tagal dia pung ana pung jahat Terjemahkan teks berikut dari bahasa Indonesia ke bahasa Melayu Kupang. Teks: Saya juga, untuk pertama kalinya (saya menggunakannya) ketika saya berada di Surabaya tapi saya tidak jatuh ke bawah 0. Terjemahan:
Output:	beta memang pertama nae di Surabaya tapi son jato 0.

Table 12: Example of Context-Based Instruction Data

Input:	Sinonim dalam bahasa Indonesia dan Melayu Kupang diberikan sebagai berikut: - peras:pres - lebih:lebe - mentah:bahati Jika teks dalam bahasa Indonesia adalah: Apa kau bisa memakan tiram mentah?, apa artinya dalam bahasa Melayu Kupang?
Output:	Lu bisa makan kerang manta ko?

Table 13: Example of Semantics-Based Instruction Data

Input:	Perhatikan kalimat berikut dalam bahasa Melayu Kupang: Oma bo'I pung gigi dong su rongko buang samua. Kalo dia mau makan siri-pinang dia musti batumbuk dia pung siri-pinang deng hahabok Teks ini memiliki makna penting dalam bahasa Indonesia: Karena seperti biasa tempat seperti itu adalah suci, kan? Tolong terjemahkan ke bahasa Melayu Kupang.
Output:	biasa kalo tempat-tempat begitu tu yang kramat a?

Table 14: Example of Phonetic Instruction Data

Target	**Context**	**Semantic**	**Phonetic**	**List-Group-Label**	**Instructional**
Adu, be pung paru saki! Be parmisi pi bal-akang do.	aduh! beta pung perut saket. Mamisi, beta mau pi kamar kecil dolo.	aduh! beta pung perut saket. Mamisi, beta mau pi kamar kecil dolo.	adu! beta pung perut saket. Mamisi, beta mau pi kamar kecil dolo.	aduh! beta pung perut saket banget. Maaf, beta mau pi kamar kecil dolo.	Aduuu! Beta pung paru talalu saki. Tolong, beta pi kamar kici dolo.
Botong sangka dia pung badan ba'isi, tagal su ampa bulan sonde datang bulan.	Beta pikir dia hamil, karna su empat bulan sonde haid.	kotong kira dia hamil, karna su empat bulan dia son haid.	kotong kira dia hamil, karna su empat bulan dia tida haid.	Beta pikir dia hamil, karna su ampa bulan sonde haid.	Kotong kira dia hamil, te su am-pat bulan son haid.

Table 15: Example of translation result across different instruction.

Dataset	# of Texts	Avg. Words	Max. Words	Min. Words
Tapaleuk News	99	704.42	1,078	227
Jakarta Field Station	66	6.24	13,184	879
Taxi1500-RawData	18,452	43.31	284	1
Poem & Pantun	37	56.02	181	18

Table 16: Statistics of monolingual dataset in Kupang Malay.

Task	Input	Target	Output
Sentiment Analysis	*mung kamare mambu rokok lan banyu kamar mburi trocoh neng kamar karo AC kurang adem. Ngapurane komplen supoyo dijogo kualitase*	negative	negative
Question Answering - Does the text contain an element of surprise?	*simkuring nu isuk2 rek nyimpen pakean ka cucian jeung ninggali sirah orok bijil ti panto gudang (ieu sigana simkuring nganyimpang, soalna simkuring sieun pisan waktos eta)*	no	no
Topic Modeling (*fear*, *disgusted*, *sad*, *happy*, *angry*, *surprise*, *shame*)	*Alhamdulillah ammetai persib, rannui atia, tinroa tong sannang. MVP anne bangngia anjo [USERNAME] yang pa'kullei a'cini' persib na tena a'pau le'ba siallo susai a'komunikasi kah kasibukanganna ade bedur*	sad	sad

Table 17: Examples of multitasking performance on the `Lius` Model.

Parameter	mT5 small	mT5 base	mT5 large
max_input_length	128	128	128
max_output_length	52	52	52
batch_size	32	16	16
bfp16	True	True	True
lr	3e-5	3e-5	3e-5

Table 18: Hyperparameter settings for Cendol mT5 models (small to large).