Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1549~1560
e-ISSN: 2548-6861

Evaluating Machine Translation Models and LLMs for Indonesian–Javanese Translation Across Speech Levels

Mahendra Bayu Prayoga [1]*, Bagas Restya Ermawan [2]*, Akmal Rafi Fadhillah [3]*,Mohammad Nizar Farizi [4]*, Ema Utami [5]*
* Teknik Informatika, Universitas Amikom Yogyakarta
m.bayu@students.amikom.ac.id [1], b.restya@students.amikom.ac.id [2], a.rafi@students.amikom.ac.id [3], ,m.nizar@students.amikom.ac.id [4] ,ema.u@amikom.ac.id [5]

Article Info

Article history:

Received 2026-01-21
Revised 2026-02-22
Accepted 2026-04-08

Keyword:

Javanese Translation,
NLLB-200,
M2M100,
Gemini,
Low-Resource Language.

ABSTRACT

Despite being one of the most widely spoken regional languages in Indonesia, Javanese remains underrepresented in modern machine translation systems, particularly with respect to its hierarchical speech-level system. This study presents a comprehensive benchmarking of machine translation approaches for low-resource Indonesian-to-Javanese translation with explicit consideration of Javanese speech-level registers, namely Ngoko, Krama, and Krama Alus. We evaluate the effectiveness of two multilingual neural machine translation models, NLLB-200 and M2M100, under both zero-shot and supervised fine-tuning settings using a parallel corpus of approximately 4,000 sentence pairs from the Unggah-Ungguh dataset. Translation quality is assessed using BLEU, chrF++, METEOR, and BERTScore on both register-specific and overall test sets constructed from a balanced evaluation set of 1,500 sentence pairs (500 per register). Experimental results show that supervised fine-tuning substantially improves translation performance, with fine-tuned M2M100 achieving the strongest results among neural machine translation models. In addition, instruction-based translation using the Gemini large language model demonstrates superior overall performance, particularly in semantic-oriented metrics, highlighting its effectiveness under controlled instruction-based conditions within the scope of this experimental configuration. Overall, this study provides a reproducible and extensible evaluation framework for sociolinguistically informed machine translation of regional languages.

I. INTRODUCTION

Language functions as a critical medium for information exchange in multilingual societies, where linguistic diversity constitutes both a cultural asset and a technological challenge, particularly in ensuring equitable access to digital services. Indonesia is one of the most linguistically diverse countries in the world, with more than 700 regional languages spoken across its archipelago. Among these languages, Javanese is the most widely spoken regional language and remains actively used in daily communication, cultural practices, and local governance.

Despite its substantial speaker population, Javanese remains significantly underrepresented in digital language technologies, reflecting a persistent resource imbalance that limits computational modeling and large-scale deployment. This condition places Javanese in the category of low-resource languages, where the availability of large-scale parallel corpora, standardized linguistic tools, and annotated training resources is extremely limited. As digital transformation continues to expand across public administration, education, and online services, the lack of reliable Javanese translation systems increasingly marginalizes regional language speakers from full participation in the digital ecosystem.

The Javanese language exhibits unique linguistic characteristics that further complicate automatic translation. It employs a hierarchical speech-level system consisting of

Ngoko (informal), Krama (polite), and Krama Alus (highly polite), where lexical choice, pronouns and verb forms vary according to social context and politeness. An effective translation system must therefore preserve not only semantic meaning but also pragmatic appropriateness and honorific consistency. These sociolinguistic properties make Indonesian–Javanese translation a particularly challenging task for computational models, especially under low-resource conditions where large-scale parallel corpora and linguistic resources are scarce.

Recent advances in neural machine translation (NMT) have demonstrated that deep learning models, particularly Transformer-based architectures, significantly outperform traditional rule-based and statistical approaches. Despite these advances, the performance of NMT systems remains highly dependent on the availability of large, high-quality parallel corpora, which remain scarce for most regional languages. Extensive research on low-resource NMT has shown that translation quality can be improved through various strategies, including the exploitation of source-side monolingual data [1], post-editing of machine-generated translations [2], Transformer-based re-ranking mechanisms [3], hybrid NMT–SMT frameworks [4], and incremental training schemes with multilingual knowledge transfer [5], [6].

Several studies have also demonstrated that fine-tuning multilingual pretrained models such as mBART50 and mT5 can substantially improve translation quality for low-resource language pairs [7], [8]. Knowledge distillation techniques have been proposed to enhance multilingual translation performance while reducing model complexity [9] and language disentanglement approaches have been introduced to improve unsupervised translation in extremely low-resource scenarios [10]. Moreover, the exploitation of monolingual corpora and data augmentation techniques has been shown to be effective in mitigating data scarcity [11].

Comprehensive surveys and systematic literature reviews in low-resource machine translation consistently emphasize the importance of multilingual modeling and transfer learning as the most promising directions for advancing translation performance in data-constrained environments [12], [13], [14], [15]. Multilingual neural machine translation models, such as mT5, mBART, NLLB, and M2M100, are trained on hundreds of languages simultaneously and leverage cross-lingual representations to enable knowledge sharing across language families. The FLORES benchmark has become a widely adopted evaluation dataset for multilingual translation, particularly for low-resource languages such as Nepali and Sinhala [16].

Nevertheless, most existing benchmarks and evaluation frameworks primarily focus on international languages or national languages with moderate to high resource availability. Regional languages in Indonesia, particularly Javanese, remain largely underrepresented in multilingual machine translation research. Furthermore, current benchmarks do not explicitly model sociolinguistic properties such as speech-level variation, which is a defining feature of the Javanese language and a critical component of pragmatic correctness in translation.

Recent work has shown that efficient fine-tuning strategies and prompt-based optimization techniques can further improve translation performance for low-resource languages using large language models [8], [17]. However, systematic benchmarking studies that evaluate the capability of multilingual translation models on Indonesian–Javanese translation across different speech registers are still lacking. This gap limits our understanding of how well modern multilingual models handle sociolinguistic variation and contextual politeness in regional languages.

Based on these limitations, this study identifies two major research problems. First, there is no standardized and reproducible benchmark for evaluating Indonesian–Javanese machine translation that explicitly accounts for Javanese speech-level variation. Second, there is a lack of comprehensive comparative studies that analyze the performance of multilingual neural machine translation models on Javanese under low-resource conditions. The objective of this study is to establish a comprehensive benchmarking framework for evaluating multilingual neural machine translation models on low-resource Indonesian–Javanese translation across three major speech registers Ngoko, Krama, and Krama Alus. This study further aims to analyze the impact of supervised fine-tuning on improving register-aware translation quality using a curated parallel corpus.

The contributions of this research are threefold. First, it introduces the first register-aware benchmark for Indonesian–Javanese machine translation. Second, it provides a systematic comparison of multilingual translation models under low-resource conditions using multiple evaluation metrics. Third, it offers a reproducible experimental framework that can serve as a strong baseline for future research on regional language translation in Indonesia. By establishing a standardized evaluation protocol and analyzing sociolinguistic sensitivity in multilingual translation models, this study is expected to contribute to the development of more inclusive, accessible, and linguistically informed machine translation technologies that support Indonesia's rich linguistic diversity and promote the preservation of regional languages through digital innovation.

II. METHODS

This section describes the experimental framework employed in this study. The methodology is designed to ensure reproducibility and follows standard benchmarking practices in machine translation research. The experimental pipeline consists of four main stages: dataset preparation, translation model selection, fine-tuning strategy, and evaluation protocol. All methods used in this study are based on established approaches in multilingual neural machine

translation, with necessary adaptations for the low-resource Indonesian–Javanese setting.

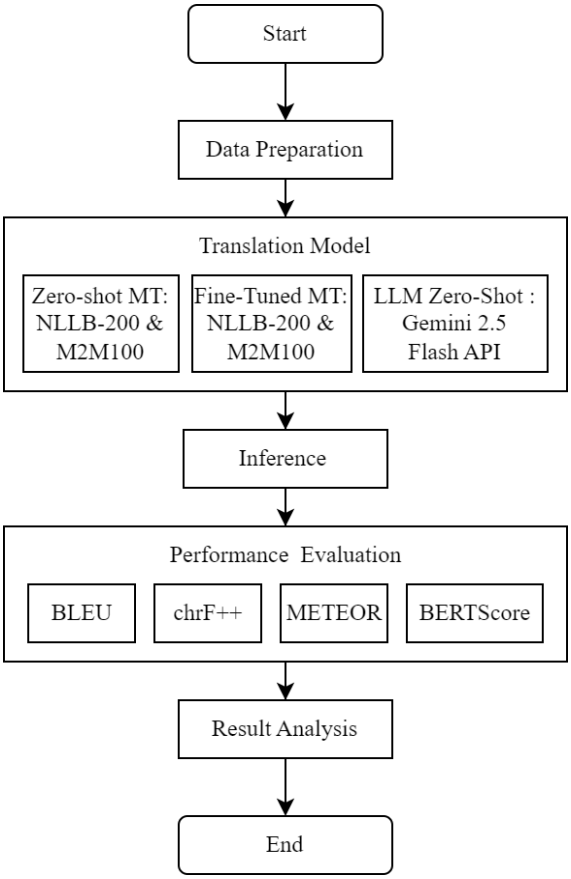

Figure 1. Research Workflow

As illustrated in Figure 1, the overall research workflow consists of dataset acquisition and preprocessing, baseline model evaluation, model fine-tuning, translation generation, and quantitative evaluation. This structured workflow is designed to ensure methodological consistency and facilitate independent replication under controlled experimental settings.

This methodology provides a comprehensive benchmarking framework for evaluating multilingual machine translation systems on low-resource Indonesian–Javanese translation. By combining zero-shot evaluation, supervised fine-tuning, register-aware evaluation, and multi-metric assessment, this framework enables systematic analysis of translation quality across linguistic and sociolinguistic dimensions [12], [13], [15], [16].

A. Dataset Preparation

The dataset used in this study, as shown in Table I, is derived from the Unggah-Ungguh Javanese Honorifics dataset, which is publicly available on the Hugging Face platform. In this dataset, the label column indicates the original honorific category, while the group column is used to map samples into the three evaluation registers considered in this study. Developed as part of the JavaneseHonorifics project, the dataset is specifically designed to capture the hierarchical speech-level system of the Javanese language and contains parallel Indonesian–Javanese sentence pairs annotated with politeness labels that reflect sociolinguistic formality and honorific usage.

TABLE I
DATASET

label	javanese sentence	group	indonesia sentence
0	Nggawaa jeruk kuwi!	1	Membawalah jeruk itu.
1	Panjenengan ngampil jeruk kuwi!	1	Membawalah jeruk itu.
2	Sampeyan mbekta jeram menika!	1	Membawalah jeruk itu.
3	Panjenengan ngampil jeram menika!	1	Membawalah jeruk itu.
…	…	…	…
2	Mila sampun anggenipun, lare alit yen kanginan lajeng anget.	542	Memang sudah sewajarnya, anak kecil jika kena angin tubuhnya lalu hangat.

The full dataset consists of approximately 4,000 Indonesian–Javanese sentence pairs. Each instance includes an Indonesian source sentence, a corresponding Javanese translation, and an honorific label. The labels represent four categories: Ngoko (label 0), Ngoko Alus (label 1), Krama (label 2), and Krama Alus (label 3); however, this study focuses on three primary registers (Ngoko, Krama, and Krama Alus) for evaluation. For benchmarking purposes, a curated evaluation subset was constructed by selecting 1,500 sentence pairs and grouping them into three registers, with 500 samples per register. Although the dataset size is modest compared to large-scale multilingual corpora, it reflects realistic data availability conditions for low-resource regional languages and provides a representative experimental setting for evaluating register-aware translation under constrained-resource scenarios.

Prior to training, basic preprocessing was performed to ensure corpus consistency. This included format verification, Unicode normalization, and removal of duplicate or empty entries. No additional linguistic normalization procedures, such as lowercasing, stemming, or punctuation filtering, were applied in order to preserve the original lexical forms and register-specific honorific expressions.

The corpus does not focus on a specific topical domain such as medical, legal, or technical content. Instead, it primarily consists of general-purpose sentences reflecting everyday communication and formal register usage. This broad yet non-specialized coverage allows the evaluation to emphasize register adaptation rather than domain-specific vocabulary, although it may limit conclusions regarding highly specialized translation scenarios.

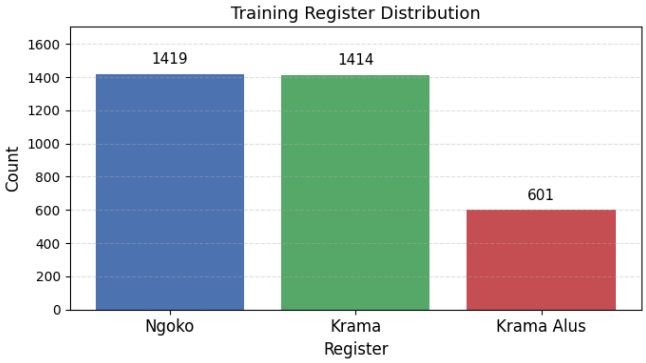

Figure 2. Training Register Distribution

The distribution of registers within the training data is shown in Figure 2, indicating an imbalance favoring Ngoko and Krama, while Krama Alus represents a minority portion of the dataset. For model training, a total of 3,434 sentence pairs with available Javanese translations were used.

All data were stored in comma-separated value (CSV) format with standardized column names, including indonesian_sentence as the source text and javanese_sentence as the reference translation. No additional linguistic normalization procedures (e.g., lowercasing, token-level filtering, or punctuation standardization) were applied beyond format validation and Unicode consistency checks, in order to preserve the original sociolinguistic characteristics of the corpus. This expanded and balanced evaluation set is intended to provide more stable and reliable estimates for automatic evaluation metrics and to support a more robust benchmarking analysis across speech registers.

B. Translation Models

This study evaluates two widely used multilingual neural machine translation models as baselines: NLLB-200-distill-600M and M2M100 418M. M2M100 is a many-to-many multilingual Transformer model trained on over 100 languages and supports direct translation between any language pair without relying on an English pivot. This property makes it suitable for evaluating direct Indonesian–Javanese translation in a low-resource setting [6], [15].

NLLB-200 (No Language Left Behind) is a multilingual encoder–decoder model trained on 200 languages using large-scale curated parallel corpora. In this study, the distilled 600M-parameter variant of NLLB-200 is used due to its favorable balance between translation quality and computational efficiency. Both models support Indonesian (ind_Latn) and Javanese (jav_Latn) language codes [9].

The baseline evaluation is performed in a zero-shot setting, where no task-specific fine-tuning is applied. This setup allows measurement of the generalization capability of pretrained multilingual models on low-resource regional languages [14].

C. Fine-Tuning Strategy

To investigate the impact of supervised domain adaptation, both the NLLB-200 and M2M100 models are fine-tuned using the Unggah-Ungguh dataset. The fine-tuning process follows a standard sequence-to-sequence training paradigm commonly used in neural machine translation [6], [7].

For fine-tuning, the dataset is split into training and validation sets using a 90:10 ratio. The model is trained for three epochs using the AdamW optimizer with a learning rate of 2×10^{-4} and a batch size of 4. The maximum sequence length is set to 128 tokens for both source and target sentences. Tokenization is performed using the original tokenizer associated with each respective model, and the models are optimized using negative log-likelihood loss between the predicted output and the human reference translations. The balanced evaluation set of 1,500 sentence pairs was held out exclusively for testing and was not used during training or validation to prevent data leakage.

All fine-tuning experiments were implemented using the Hugging Face Transformers library with a PyTorch backend. Experiments were conducted on a single GPU device to ensure computational consistency across models. A fixed random seed was applied to reduce variability and improve reproducibility. No early stopping mechanism was employed; model selection was performed based on validation loss after three training epochs.

D. Prompting Strategy for LLM-based Translation

The Gemini large language model [18] was used with its default generation settings, and no explicit control over decoding parameters such as temperature, top-p sampling, or random seed was applied. As the translations were generated through an interactive interface, deterministic generation could not be strictly enforced. Unlike neural machine translation models, Gemini was explicitly instructed with the target speech register and processed multiple sentences within a single prompt.

As a result, the Gemini-based translation setup differs fundamentally from the automated inference pipeline used by neural machine translation models. While NMT systems generate translations in a fully automatic manner without explicit register control in the input, Gemini benefits from direct register specification and interactive instruction-based generation.

Preliminary experiments were conducted using the Gemini API (Gemini 2.5 Flash) however, due to token and usage limitations, the final translations used for evaluation were generated through controlled instruction-based prompting via the official Gemini interface. The model was not fine-tuned or updated during the experiment.

Execution time for the Gemini model was manually recorded by measuring the duration from prompt submission to completion of a single batch prompt containing multiple sentences. This timing reflects interactive usage under manual prompting conditions and is therefore reported as an approximate value. As such, this execution time is not directly comparable to the automated inference time reported for neural machine translation models.

Each prompt explicitly specified the target language (Javanese) and the desired speech register (Ngoko, Krama, or Krama Alus). To ensure consistency and prevent undesired explanations or meta-commentary, strict output constraints were enforced. The model was instructed to generate translations only, following a fixed enumerated format without additional explanations or notes.

For each speech register, the evaluation sentences were provided in controlled batch prompts, covering a total of 500 Indonesian sentences per register. No in-context examples were included in the prompts, ensuring that all LLM-based translations were generated under a zero-shot instruction setting [8], [17]. A representative prompt template used for the instruction-based translation experiments is provided below.

```
Instruksi:

Tugas Anda adalah menerjemahkan 500 kalimat berikut
dari bahasa Indonesia ke bahasa Jawa tingkat tutur
Krama Alus (tingkat tutur yang sangat halus untuk
lawan bicara atau subjek yang sangat dihormati).

Jangan memberi penjelasan, jangan menambahkan catatan
apa pun.

Format jawaban seperti berikut:

(1) hasil
(2) hasil
...
(500) hasil

Kalimat:
(1) …
(2) …
…
(500) …
```

The same prompt structure was applied for Ngoko and Krama registers by modifying only the specified speech level in the instruction. The generated outputs were manually extracted and stored in CSV format for subsequent automatic evaluation using standard machine translation metrics.

Due to these differences in input control, generation strategy, and execution mode, the Gemini-based translation results in this study are reported as a reference performance rather than as a fully comparable alternative to automated neural machine translation systems.

To enhance procedural transparency, all prompts were constructed using a fixed template with identical structural constraints across registers, differing only in the specified target speech level. Each batch was executed once under consistent conditions, and no iterative refinement or output filtering was applied. While deterministic replication cannot be fully guaranteed due to the interactive generation environment, the controlled prompt design ensures procedural consistency throughout the evaluation.

E. Performance Evaluation

Translation quality is evaluated using four complementary automatic evaluation metrics: BLEU, chrF++, METEOR, and BERTScore. BLEU measures n-gram overlap between the model output and the reference translation and serves as a standard lexical-level evaluation metric [14]. chrF++ is a character-level F-score metric that is particularly suitable for morphologically rich languages such as Javanese, where word-level tokenization can be inconsistent [15]. METEOR incorporates stemming and synonym matching, allowing for more flexible alignment between candidate and reference translations [13]. BERTScore is a neural similarity metric that computes contextual embedding similarity between the translated output and the reference sentence, providing a semantic-level evaluation [12]. The evaluation pipeline is implemented using the SacreBLEU, NLTK, and BERTScore libraries. Metrics are computed separately for each speech register and for the combined test set.

III. Results And Discussion

This section presents the quantitative results of the machine translation experiments. The evaluation compares zero-shot pretrained multilingual models, fine-tuned multilingual models, and an instruction-based large language model across different Javanese speech registers. Translation quality is assessed using BLEU, chrF++, METEOR, and BERTScore. Results are reported at both the overall level and the register level based on a balanced evaluation set of 1,500 sentence pairs to provide stable and representative performance estimates.

A. Zero Shot Translation Performance

Zero-shot evaluation measures the baseline translation capability of pretrained multilingual models without any task-specific adaptation. Two models are evaluated in this setting: NLLB-200-distill-600M and M2M100 418M. Both models are applied directly to the evaluation dataset without fine-tuning.

TABLE II
ZERO SHOT TRANSLATION PERFORMANCE (OVERALL)

Model	BLEU	chrF++	METEOR	BERTScore
Zero-Shot NLLB- 200	6.892074	27.717525	0.167336	0.766909
Zero-Shot M2M100	2.978750	21.143721	0.091273	0.711595

Table II illustrates that NLLB-200 consistently outperforms M2M100 across all evaluation metrics in the zero-shot setting. The comparison presented in Table II further indicates a clear performance gap in favor of NLLB-200, particularly in lexical-level metrics such as BLEU and chrF++. Using the expanded and balanced evaluation set of 1,500 sentence pairs, NLLB-200 attains a BLEU score of 6.89, compared to 2.98 for M2M100, indicating stronger n-

gram overlap with the reference translations. A similar trend is observed for chrF++, where NLLB-200 achieves 27.72 compared to 21.14 for M2M100, suggesting better character-level similarity and morphological consistency.

Semantic-oriented metrics also favor NLLB-200. The METEOR score of 0.167 and BERTScore (F1) of 0.767 indicate better alignment with reference translations compared to M2M100, which records 0.091 METEOR and 0.712 BERTScore. Collectively, these findings position NLLB-200 as a comparatively stronger zero-shot baseline, although overall register-sensitive robustness remains limited.

TABLE III
ZERO SHOT TRANSLATION PERFORMANCE BY SPEECH REGISTER

Model	Register	BLEU	chrF++	METEOR	BERTScore
Zero-Shot NLLB - 200	Ngoko	13.538974	39.976397	0.299611	0.799928
Zero-Shot NLLB - 200	Krama	2.835005	22.824944	0.092893	0.750026
Zero-Shot NLLB - 200	Krama Alus	3.017679	23.055520	0.109505	0.750774
Zero-Shot M2M100	Ngoko	3.711496	23.326703	0.112530	0.717937
Zero-Shot M2M100	Krama	2.361248	19.877232	0.064350	0.705154
Zero-Shot M2M100	Krama Alus	2.777530	20.652560	0.096939	0.711695

A register-level breakdown (Table III) reveals substantial performance variation across speech levels, highlighting the sensitivity of zero-shot models to sociolinguistic formality. In the Ngoko register, NLLB-200 achieves the strongest performance, with a BLEU score of 13.54 and a chrF++ score of 39.98, substantially outperforming M2M100, which records 3.71 BLEU and 23.33 chrF++. The relatively higher scores in the Ngoko register suggest that informal language forms are more readily captured by zero-shot pretrained models.

In contrast, translation quality decreases markedly for the Krama and Krama Alus registers. For NLLB-200, BLEU scores drop to 2.84 for Krama and 3.02 for Krama Alus, indicating increased difficulty in modeling higher politeness levels without supervised adaptation. Similar degradation is observed for M2M100, which achieves BLEU scores below 3.0 across both formal registers.

Despite the decline in lexical and character-level metrics, BERTScore values for NLLB-200 remain relatively stable across registers, ranging from 0.75 to 0.80. This suggests that semantic similarity is partially preserved even when surface-level realizations become less accurate. Overall, these results indicate that zero-shot pretrained models struggle to consistently capture sociolinguistic distinctions, particularly for higher politeness registers such as Krama and Krama Alus.

B. Fine-Tuned Translation Performance

To provide additional insight into the fine-tuning process, this section first reports the training dynamics of the neural machine translation models. Training and validation loss values are presented to illustrate optimization behavior, convergence stability, and potential overfitting during supervised fine-tuning. These results complement the translation quality metrics by showing how model performance evolves throughout the training process.

TABLE IV
TRAINING AND VALIDATION LOSS OF NLLB-200 FINE-TUNING

Epoch	Train Loss	Valid Loss
1	0.6606	0.1956
2	0.1374	0.1710
3	0.0882	0.1632

As shown in Table IV, both training and validation loss for the NLLB-200 model decrease steadily across epochs. This consistent downward trend indicates stable optimization and effective learning from the parallel Indonesian–Javanese training data. The relatively small gap between training and validation loss suggests that the model does not exhibit severe overfitting during fine-tuning.

TABLE V
TRAINING AND VALIDATION LOSS OF M2M100 FINE-TUNING

Epoch	Train Loss	Valid Loss
1	0.6363	0.3517
2	0.2508	0.2714
3	0.1664	0.2362

Table V presents the training and validation loss values observed during the fine-tuning of the M2M100 model. Similar to NLLB-200, both loss curves show a consistent decline across epochs, indicating effective convergence. Notably, the stable validation loss trend supports the substantial translation quality improvements achieved by

M2M100 after fine-tuning, as reported in the subsequent evaluation results.

In addition to loss dynamics, the execution time required for the fine-tuning process was also recorded. Fine-tuning the NLLB-200 model for three epochs required approximately 34 minutes, while fine-tuning M2M100 completed in approximately 23 minutes and 56 seconds under the same experimental setup. These durations represent end-to-end fine-tuning execution time and are reported to provide contextual information on computational cost rather than as indicators of translation performance.

To evaluate the effect of supervised domain adaptation, both NLLB-200 and M2M100 are fine-tuned using parallel Indonesian–Javanese data from the Unggah-Ungguh dataset. The fine-tuned models are evaluated on the same test set as the zero-shot models to ensure direct comparability.

TABLE VI
FINE-TUNED TRANSLATION PERFORMANCE (OVERALL)

Model	BLEU	chrF++	METEOR	BERTScore
Fine-Tuned NLLB- 200	6.892074	27.717525	0.167336	0.766909
Fine-Tuned M2M100	32.296017	56.392052	0.472486	0.857684

Table VI summarizes the overall performance of fine-tuned models, demonstrating measurable improvements compared to their zero-shot counterparts. Supervised fine-tuning results in substantial performance improvements for M2M100, while improvements for NLLB-200 remain limited under the current experimental configuration. The fine-tuned NLLB-200 model achieves a BLEU score of 6.89 and a chrF++ score of 27.72, accompanied by METEOR and BERTScore values of 0.167 and 0.767, respectively. These results indicate moderate gains in both lexical accuracy and semantic alignment compared to the zero-shot setting.

In contrast, M2M100 exhibits a substantial performance improvement after fine-tuning. The BLEU score increases to 32.30, and chrF++ reaches 56.39, reflecting strong word-level and character-level accuracy. Semantic-oriented metrics also show notable gains, with METEOR rising to 0.472 and BERTScore to 0.858. These results constitute the strongest performance observed among the evaluated neural machine translation models within the scope of this study, highlighting the effectiveness of supervised fine-tuning for register-aware Indonesian–Javanese translation.

TABLE VII
FINE-TUNED TRANSLATION PERFORMANCE BY SPEECH REGISTER

Model	Register	BLEU	chrF++	METEOR	BERTScore
Fine-Tuned NLLB-200	Ngoko	13.538974	39.976397	0.299611	0.799928
Fine-Tuned NLLB-200	Krama	2.835005	22.824944	0.092893	0.750026
Fine-Tuned NLLB-200	Krama Alus	3.017679	23.055520	0.109505	0.750774
Fine-Tuned M2M100	Ngoko	22.242452	46.826400	0.340635	0.825540
Fine-Tuned M2M100	Krama	40.863060	63.901178	0.555222	0.883097
Fine-Tuned M2M100	Krama Alus	33.384426	56.740658	0.521601	0.864416

The register-specific results in Table VII further demonstrate the differential impact of supervised fine-tuning across speech levels. For the NLLB-200 model, the Ngoko register benefits the most from fine-tuning, achieving a BLEU score of 13.54 and a chrF++ score of 39.98. In contrast, performance on the Krama and Krama Alus registers remains substantially lower, with BLEU scores of 2.84 and 3.02, respectively. This indicates limited adaptation of NLLB-200 to highly formal speech levels, despite moderate gains in semantic-oriented metrics such as BERTScore.

In comparison, M2M100 demonstrates consistently strong performance across all registers after fine-tuning. The model achieves a BLEU score of 22.24 for Ngoko, 40.86 for Krama, and 33.38 for Krama Alus, accompanied by high chrF++ scores exceeding 46 across registers. Notably, the Krama register yields the highest performance, with a chrF++ score of 63.90 and a BERTScore of 0.883, indicating effective learning of formal lexical and morphological patterns.

While fine-tuning substantially improves most evaluation metrics, performance variations across registers remain evident. In particular, lower METEOR scores for the Krama Alus register suggest sensitivity of semantic evaluation metrics to lexical diversity and formal variation, especially when evaluated against a single reference translation. Overall, these results highlight the importance of register-aware training data for achieving robust translation quality across sociolinguistically distinct speech levels.

C. LLM-Based Translation Performance (Gemini)

The LLM-based translation system, Gemini, is evaluated using an instruction-based prompting strategy under a zero-shot setting. Translation is guided solely by explicit prompts specifying the target language and speech register, without any parameter updates or fine-tuning.

TABLE VIII
GEMINI TRANSLATION PERFORMANCE (OVERALL)

Model	BLEU	chrF++	METEOR	BERTScore
Gemini	37.488089	67.077348	0.566769	0.880644

Table VIII presents the overall performance of the instruction-based LLM, highlighting its strong automatic

evaluation scores across metrics. Gemini achieves a BLEU score of 37.49 and a chrF++ score of 67.08, demonstrating higher automatic evaluation scores than the evaluated neural machine translation models under the specified instruction-based experimental setting. The METEOR score (0.567) and BERTScore (0.881) further indicate strong semantic alignment with the reference translations.

TABLE IX
GEMINI TRANSLATION PERFORMANCE BY SPEECH REGISTER

Register	BLEU	chrF++	METEOR	BERTScore
Ngoko	42.953733	69.950838	0.623067	0.882991
Krama	24.921587	61.720111	0.476104	0.861171
Krama Alus	42.849018	70.520171	0.601136	0.897770

A register-level analysis (Table IX) indicates that Gemini maintains relatively stable performance across speech levels, although metric variations remain observable. The Ngoko register achieves the highest BLEU score (42.95), while Krama Alus also attains a comparably high BLEU score (42.85). Although performance in the Krama register is lower (24.92 BLEU), it remains competitive relative to fine-tuned neural models.

Notably, BERTScore values remain consistently high across all registers, ranging from 0.861 to 0.898, indicating stable semantic similarity regardless of politeness level. These findings indicate that explicit instruction-based prompting may facilitate more effective register adaptation in large language models compared to fully automated multilingual NMT pipelines [8], [17].

D. Overall Comparison of Translation Approaches

Table X provides a consolidated cross-paradigm comparison, enabling direct evaluation of performance trade-offs across modeling strategies. Translation quality is evaluated using BLEU, chrF++, METEOR, and BERTScore. The reported time values reflect the execution duration required to generate translations under each respective setup. It should be noted that execution time and translation performance for the instruction-based large language model are not directly comparable to automated neural machine translation systems.

TABLE X
OVERALL COMPARISON OF TRANSLATION APPROACHES

Model Category	Model	BLEU	chrF++	METEOR	BERTScore	Time
Zero-Shot MT	NLLB- 200	6.892074	27.717525	0.167336	0.766909	6 min 10 sec
	M2M100	2.978750	21.143721	0.091273	0.711595	7 min 54 sec
Fine-Tuned MT	NLLB- 200	6.892074	27.717525	0.167336	0.766909	6 min 7 sec
	M2M100	32.296017	56.392052	0.472486	0.857684	8 min 58 sec
Zero-Shot LLMs	Gemini	37.488089	67.077348	0.566769	0.880644	≈ 1 min 15 sec

Among zero-shot neural machine translation models, NLLB-200 consistently outperforms M2M100 across all evaluation metrics, achieving a BLEU score of 6.89 and a BERTScore of 0.767, compared to 2.98 BLEU and 0.712 BERTScore for M2M100. Although both models complete inference within several minutes, their relatively low translation scores indicate limited effectiveness in capturing register-sensitive linguistic features under low-resource conditions.

Supervised fine-tuning leads to substantial performance improvements for neural machine translation models, particularly for M2M100. Fine-tuned NLLB-200 shows limited gains, retaining a BLEU score of 6.89 and a BERTScore of 0.767 with an execution time of approximately six minutes. In contrast, fine-tuned M2M100 demonstrates a dramatic increase in translation quality, achieving a BLEU score of 32.30, a chrF++ score of 56.39, and a BERTScore of 0.858. This improvement is accompanied by a longer execution time of approximately nine minutes, reflecting the additional computational cost associated with higher translation accuracy.

The instruction-based large language model, Gemini, achieves the highest automatic metric scores within the evaluated experimental configuration, with a BLEU score of 37.49, a chrF++ score of 67.08, a METEOR score of 0.567, and a BERTScore of 0.881. The reported execution time of approximately one minute corresponds to a single manual batch prompt containing multiple sentences. Due to its interactive and interface-based usage, this timing reflects manual execution rather than automated inference and should therefore be interpreted separately from neural machine translation pipelines. Consequently, Gemini results are reported as a reference or upper-bound performance rather than as a fully operational alternative to automated neural machine translation systems.

Overall, the results in Table X highlight clear trade-offs across translation paradigms. Zero-shot multilingual models offer limited translation quality despite moderate execution times, fine-tuned neural machine translation models provide strong lexical and morphological accuracy at increased computational cost, and instruction-based large language models deliver superior semantic quality under explicit register control. These findings provide practical guidance for selecting translation approaches based on quality requirements, computational resources, and deployment constraints.

Beyond translation quality, practical factors such as computational efficiency and deployment cost must also be

considered. Fine-tuned neural machine translation models, once trained, can be deployed within fully automated pipelines with predictable inference time and without reliance on external service providers. In contrast, instruction-based large language models typically operate through interactive or API-based environments, which may introduce latency, usage limits, or monetary costs depending on deployment configurations.

Although instruction-based LLMs demonstrate strong semantic performance, their integration into large-scale production systems requires careful assessment of scalability and operational constraints. Consequently, selecting between fine-tuned neural models and instruction-based LLMs involves balancing translation accuracy with efficiency, scalability, and long-term deployment feasibility.

A closer examination of the evaluation metrics reveals important distinctions between lexical-level and semantic-level performance indicators. In several instances, models with relatively low BLEU scores still obtain moderately stable BERTScore values, particularly in zero-shot settings. This pattern suggests that while surface-level n-gram overlap remains limited, contextual semantic similarity may still be partially preserved. Conversely, the substantial improvements observed in BLEU and chrF++ after supervised fine-tuning, particularly for M2M100, indicate enhanced lexical and morphological alignment and improved modeling of register-sensitive vocabulary.

The consistently high BERTScore values achieved by the instruction-based LLM further emphasize strong contextual alignment with reference translations, even when surface-level realizations vary. This divergence across evaluation metrics underscores the necessity of multi-metric assessment in low-resource, register-aware translation tasks, where lexical fidelity, morphological precision, and semantic adequacy do not always improve proportionally.

While statistical and rule-based translation approaches have historically contributed to low-resource language processing, the present study focuses specifically on contemporary multilingual neural and instruction-based paradigms. Incorporating traditional baselines would require distinct handcrafted linguistic resources and system configurations that fall outside the scope of the current benchmarking framework. Future research may explore comparative evaluation across neural and rule-based paradigms to further contextualize register-aware translation performance.

E. Zero-Shot Multilingual Models in a Low-Resource Setting

The zero-shot results demonstrate that pretrained multilingual models possess a baseline capability to translate Indonesian into Javanese without task-specific adaptation. Among the evaluated models, NLLB-200 consistently outperforms M2M100 across all evaluation metrics, indicating stronger cross-lingual generalization in the absence of supervision.

Nevertheless, the overall performance of both models remains relatively limited, particularly in lexical and character-level metrics. This suggests that large-scale multilingual pretraining alone is insufficient to capture the linguistic complexity of Javanese, especially its hierarchical politeness system [12]. Register-wise results further highlight this limitation, as both models exhibit performance degradation when translating into Krama and Krama Alus registers, even under a balanced and expanded evaluation set.

These findings suggest that while zero-shot multilingual models offer a functional baseline, their capacity to encode hierarchical sociolinguistic constraints remains insufficient without supervised adaptation. As a result, zero-shot translation may be inadequate for applications requiring precise control over politeness and honorific usage.

F. Impact of Supervised Fine-Tuning on Register Sensitivity

Supervised fine-tuning leads to substantial improvements in translation quality for both NLLB-200 and M2M100. The observed gains across BLEU, chrF++, METEOR, and BERTScore confirm that domain adaptation plays a crucial role in improving translation performance for low-resource language pairs [5], [7].

The improvement is particularly pronounced for M2M100, which exhibits a dramatic increase in both lexical accuracy and semantic similarity after fine-tuning. This suggests that, despite weaker zero-shot performance, M2M100 benefits significantly from supervised exposure to parallel Indonesian–Javanese data. Fine-tuning enables the model to better learn register-specific lexical choices and morphological patterns that are not reliably acquired during multilingual pretraining.

In contrast, the impact of supervised fine-tuning on NLLB-200 appears limited under the current experimental configuration, as the fine-tuned model does not show measurable improvement over its zero-shot performance. This outcome suggests that the relatively small supervised dataset may be insufficient to substantially adapt the pretrained multilingual representations. It further highlights that the effectiveness of fine-tuning is model-dependent and influenced by underlying architectural and pretraining characteristics.

G. Register-Level Translation Behavior

The register-wise analysis provides important insights into model behavior across different levels of formality. Across different modeling approaches, translation performance varies by speech register, with Ngoko generally yielding higher scores in zero-shot settings, while fine-tuned and instruction-based models demonstrate improved performance on more formal registers. This trend persists despite the use of a balanced evaluation set and reflects both linguistic complexity and differences in how sociolinguistic rules are represented during training.

For fine-tuned models, improvements in Krama and Krama Alus registers indicate that labeled training data enables the

models to partially internalize sociolinguistic rules governing politeness and honorific usage. However, performance disparities across registers persist, highlighting the inherent difficulty of modeling formal speech levels in Javanese [13], [15].

These findings underscore the importance of register-aware evaluation for languages with hierarchical speech systems. Evaluating translation quality solely at the aggregate level may obscure significant performance differences that emerge when sociolinguistic variation is taken into account.

The observed variation across registers can be further explained by inherent linguistic differences in speech-level structure. Ngoko, as the informal register, typically employs more direct lexical forms and fewer honorific substitutions, making it comparatively easier for models to generate accurate translations. In contrast, Krama and particularly Krama Alus involve specialized honorific vocabulary, hierarchical lexical replacement, and context-sensitive politeness markers. These forms often require selecting entirely different lexical items rather than performing direct word-level translation, increasing the risk of register mismatch.

Consequently, translation in higher politeness registers demands not only semantic equivalence but also precise sociolinguistic alignment, which remains challenging for both zero-shot and supervised models, especially in longer or syntactically complex sentences.

It is important to note that the training data distribution across speech registers is not perfectly balanced, which may contribute to the observed performance differences, particularly for formal registers such as Krama and Krama Alus. Nevertheless, the evaluation set used in this study is explicitly balanced across registers, ensuring that the reported results primarily reflect model behavior rather than evaluation bias.

Beyond quantitative performance differences, qualitative inspection of model outputs reveals several recurring error patterns. In zero-shot settings, models frequently generate register mismatches, where informal lexical forms appear in contexts requiring Krama or Krama Alus politeness levels. This suggests limited sensitivity to hierarchical honorific constraints without supervised adaptation.

TABLE XI
ILLUSTRATIVE TRANSLATION EXAMPLES ACROSS MODELS

Register	Indonesia	Reference Translation	Zero-Shot NLLB - 200	Zero-Shot M2M100	Fine-Tuned NLLB-200	Fine-Tuned M2M100	Gemini
Ngoko	Tembang ini berisi petunjuk yang mengingatkan para pemuda.	Tembang iki isi pepeling tumrap para mudha.	Lagu iki ngandhut petunjuk kanggo ngélingaké para nom-noman.	Tembang ini berisi petunjuk yang mengingatkan para pemuda.	Lagu iki ngandhut petunjuk kanggo ngélingaké para nom-noman.	Tembang iki isi pepeling tumrap para mudha.	Tembang iki isine pituduh sing ngelingake para nom-noman.
Krama	Merah-merah itu baju atau serbet?	Abrit-abrit menika rasukan menapa serbet?	Werna abang kuwi klambi utawa serbet?	Warna merah punika baju utawa serbet?	Werna abang kuwi klambi utawa serbet?	Abang-abang kae klambi apa serbet?	Abrit-abrit menika rasukan punapa serbet?
Krama Alus	Bu Dewi memetik bunga di depan rumah.	Bu Dewi mundhut sekar ing ngajeng dalem.	Dewi mau njupuk kembang ing ngarep omah.	Piyambakipun nggunakake bunga ing depan rumah.	Dewi mau njupuk kembang ing ngarep omah.	Bu Dewi mundhut sekar ing ngajeng dalem.	Bu Dewi mundhut sekar ing ngajeng dalem.

Table XI presents representative translation outputs across models and speech registers to complement the quantitative evaluation results. In the Ngoko register, zero-shot M2M100 exhibits lexical inconsistency and partial retention of Indonesian structures, while fine-tuned M2M100 and Gemini produce more fluent and register-aligned translations. For the Krama register, zero-shot models frequently generate informal lexical forms or incorrect honorific substitutions, whereas fine-tuned M2M100 and Gemini demonstrate improved politeness alignment. In the Krama Alus example, the zero-shot M2M100 output shows clear register mismatch by using informal vocabulary, while fine-tuned M2M100 and Gemini more accurately reflect formal honorific usage. These qualitative differences reinforce the register-level quantitative findings reported in Tables III, VII, and IX.

Fine-tuned models demonstrate improved lexical accuracy; however, occasional inconsistencies in politeness markers and verb forms remain, particularly in longer sentences containing multiple honorific elements. The instruction-based LLM shows stronger register alignment under explicit prompting, although minor stylistic variations and lexical substitutions are still observed. These qualitative observations support the quantitative findings and further illustrate the complexity of modeling sociolinguistic variation in low-resource translation.

H. Instruction-Based LLM Translation

The instruction-based translation results indicate that the Gemini large language model achieves the highest automatic metric scores within the evaluated experimental configuration. Unlike neural machine translation models,

Gemini relies entirely on explicit prompts to infer both the translation task and the desired speech register, without parameter updates or supervised training. As such, Gemini is positioned in this study as a reference or upper-bound performance under explicit instruction-based conditions rather than as a fully automated translation system.

The consistently high BERTScore values across registers indicate strong semantic alignment with reference translations, suggesting that Gemini effectively preserves sentence-level meaning regardless of politeness level. Additionally, the relatively stable performance across Ngoko, Krama, and Krama Alus registers highlights the effectiveness of explicit instruction-based prompting in guiding register usage.

Although Gemini demonstrates stronger overall automatic metric scores, performance variations across registers remain observable. This suggests that, although LLMs exhibit strong semantic understanding, surface-level realization of formal linguistic structures remains challenging without explicit training data. Furthermore, this performance advantage is achieved under interactive, instruction-driven settings and does not directly reflect deployment conditions of fully automated neural machine translation systems. It should also be noted that this study relies solely on automatic evaluation metrics and does not include human evaluation, which may limit the assessment of sociolinguistic appropriateness and politeness nuances that are critical in Javanese translation.

I. Comparison Across Translation Paradigms

The comparative analysis reveals clear differences between translation paradigms. Zero-shot multilingual models provide a useful baseline but lack reliable register control [14]. Fine-tuned neural machine translation models achieve strong lexical and morphological accuracy, particularly when sufficient supervised data is available. Instruction-based LLM translation offers superior semantic coherence and flexibility, especially in low-resource scenarios where labeled data is limited.

This cross-paradigm evaluation further indicates that no single translation paradigm universally dominates across lexical, morphological, and semantic dimensions; instead, performance trade-offs emerge depending on supervision level and register control mechanisms. Consequently, the choice of translation strategy should be guided by application requirements, such as the need for lexical fidelity, semantic robustness, or sociolinguistic adaptability.

J. Implications and Future Directions

The results of this study highlight the importance of combining quantitative evaluation with register-aware analysis when studying machine translation for languages with complex sociolinguistic systems. Fine-tuning remains a critical strategy for improving translation quality in low-resource settings, while instruction-based LLMs offer a promising alternative when training data is scarce.

Future work may explore hybrid approaches that integrate neural machine translation models with instruction-guided generation, as well as larger-scale human evaluation focusing on sociolinguistic appropriateness [6], [8]. Additionally, extending the evaluation to broader datasets and automated LLM prompting frameworks could further enhance the robustness of register-aware machine translation systems.

IV. CONCLUSION

This study presents a comprehensive and reproducible benchmarking framework for evaluating machine translation approaches for Indonesian–Javanese translation with explicit consideration of speech-level registers. The experimental results demonstrate that zero-shot multilingual models provide only limited baseline performance and struggle to consistently capture the hierarchical sociolinguistic distinctions inherent in Javanese, particularly in higher politeness registers.

Supervised fine-tuning substantially improves translation quality, especially for register-sensitive forms, with fine-tuned M2M100 achieving the strongest performance among the evaluated neural machine translation models. These findings highlight the importance of domain adaptation in low-resource, register-aware translation settings.

Instruction-based translation using a large language model exhibits strong overall performance, particularly in semantic-oriented metrics. However, this performance is achieved under explicit instruction-driven and interactive conditions and is therefore reported as a reference or upper-bound result rather than as a fully automated alternative to neural machine translation systems.

Overall, the empirical evidence reveals complementary strengths across translation paradigms and reinforces the necessity of integrating register-aware evaluation and sociolinguistic considerations into low-resource machine translation research. Such an approach supports the development of linguistically informed translation systems for regional languages with complex speech-level structures such as Javanese. Beyond technical evaluation, these findings also have potential implications for educational applications, digital preservation of regional languages, and culturally sensitive multilingual public services in linguistically diverse societies.

REFERENCES

[1] A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data," *Appl. Sci.*, vol. 13, no. 2, p. 1201, Jan. 2023, doi: 10.3390/app13021201.

[2] D. Rakhimova, A. Karibayeva, and A. Turarbek, "The Task of Post-Editing Machine Translation for the Low-Resource Language," *Appl. Sci.*, vol. 14, no. 2, p. 486, Jan. 2024, doi: 10.3390/app14020486.

[3] A. Javed *et al*., "Transformer-Based Re-Ranking Model for Enhancing Contextual and Syntactic Translation in Low-Resource

Neural Machine Translation," *Electronics*, vol. 14, no. 2, p. 243, Jan. 2025, doi: 10.3390/electronics14020243.

[4] K. Bhuvaneswari and M. Varalakshmi, "Efficient incremental training using a novel NMT-SMT hybrid framework for translation of low-resource languages," *Front. Artif. Intell.*, vol. 7, p. 1381290, Sep. 2024, doi: 10.3389/frai.2024.1381290.

[5] K. Huang, P. Li, J. Ma, T. Yao, and Y. Liu, "Knowledge Transfer in Incremental Learning for Multilingual Neural Machine Translation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 15286–15304. doi: 10.18653/v1/2023.acl-long.852.

[6] M. K. Pasupuleti, "Multilingual NLP for Low-Resource Languages Using Transfer Learning," *Int. J. Acad. Ind. Res. Innov.*, vol. 05, no. 05, pp. 452–461, May 2025, doi: 10.62311/nesx/rphcr7.

[7] Z. Kozhirbayev, "Enhancing Neural Machine Translation with Fine-Tuned mBART50 Pre-Trained Model: An Examination with Low-Resource Translation Pairs," *Ingénierie Systèmes Inf.*, vol. 29, no. 3, pp. 831–838, Jun. 2024, doi: 10.18280/isi.290304.

[8] P. W. Khoboko, V. Marivate, and J. Sefara, "Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models," *Mach. Learn. Appl.*, vol. 20, p. 100649, Jun. 2025, doi: 10.1016/j.mlwa.2025.100649.

[9] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T.-Y. Liu, "Multilingual Neural Machine Translation with Knowledge Distillation," 2019, *arXiv*. doi: 10.48550/ARXIV.1902.10461.

[10] X.-P. Nguyen, S. Joty, W. Kui, and A. T. Aw, "Refining Low-Resource Unsupervised Translation by Language Disentanglement of Multilingual Model," 2022, *arXiv*. doi: 10.48550/ARXIV.2205.15544.

[11] J. Pang *et al.*, "Rethinking the Exploitation of Monolingual Data for Low-Resource Neural Machine Translation," *Comput. Linguist.*, vol. 50, no. 1, pp. 25–47, Mar. 2024, doi: 10.1162/coli_a_00496.

[12] S. Shi, X. Wu, R. Su, and H. Huang, "Low-resource Neural Machine Translation: Methods and Trends," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 5, pp. 1–22, Sep. 2022, doi: 10.1145/3524300.

[13] T. O. Tafa *et al.*, "Machine Translation Performance for Low-Resource Languages: A Systematic Literature Review," *IEEE Access*, vol. 13, pp. 72486–72505, 2025, doi: 10.1109/ACCESS.2025.3562918.

[14] S. Ranathunga, E.-S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural Machine Translation for Low-Resource Languages: A Survey," 2021, *arXiv*. doi: 10.48550/ARXIV.2106.15115.

[15] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, and A. Birch, "Survey of Low-Resource Machine Translation," *Comput. Linguist.*, vol. 48, no. 3, pp. 673–732, Sep. 2022, doi: 10.1162/coli_a_00446.

[16] F. Guzmán *et al.*, "The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 6097–6110. doi: 10.18653/v1/D19-1632.

[17] S. H. Asefa and Y. Assabie, "Transformer-Based Amharic-to-English Machine Translation With Character Embedding and Combined Regularization Techniques," *IEEE Access*, vol. 13, pp. 1090–1105, 2025, doi: 10.1109/ACCESS.2024.3521985.

[18] G. Team *et al.*, "Gemini: A Family of Highly Capable Multimodal Models," May 09, 2025, *arXiv*: arXiv:2312.11805. doi: 10.48550/arXiv.2312.11805. 85.