

Comparative Evaluation of LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian Low-Resource Text Generation under Augmented Data Conditions

Aulia Akhrian Syahidi

Data Science and Artificial Intelligence, Thai-Nichi Institute of Technology, Thailand

aulia.sya@tni.ac.th

Received: November 6, 2025; Revised: June 14, 2026; Accepted: June 16, 2026; Published: June 30, 2026

ABSTRACT – Large Language Models (LLMs) have demonstrated remarkable capabilities in text generation, yet their effectiveness for Indonesian, a Low-Resource Language (LRL) that remains comparatively underrepresented in multilingual training resources and evaluation benchmarks, is still insufficiently explored. This study systematically compares LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian low-resource text generation under augmented data conditions, using an augmented dataset designed to mitigate data scarcity. Three complementary augmentation strategies, namely back-translation using Google Translate and IndoNMT, synonym substitution via mBERT-based lexical embeddings, and paraphrasing using IndoBERT/IndoGPT checkpoints, were employed to increase lexical and syntactic diversity while preserving semantic consistency. Quantitative results show that GPT-4-mini achieves higher BLEU (0.55 vs. 0.52), ROUGE-L (0.64 vs. 0.61), METEOR (0.57 vs. 0.54), lower Perplexity (11.5 vs. 12.8), and higher mBERTScore (0.90 vs. 0.88) compared to LLaMA 3.1-13B-Instruct, indicating stronger lexical and semantic alignment. Conversely, LLaMA 3.1-13B-Instruct exhibits greater lexical diversity (Distinct-2 = 0.34 vs. 0.31), suggesting a stronger ability to generate varied expressions while reducing repetitive output patterns. Human evaluation involving five independent native Indonesian raters confirms that GPT-4-mini excels in fluency (4.5 vs. 4.1), coherence (4.4 vs. 4.0), and relevance (4.4 vs. 4.3), while LLaMA 3.1-13B-Instruct slightly surpasses it in factual accuracy (4.5 vs. 4.2). These findings highlight the complementary strengths of the two models, with GPT-4-mini demonstrating superior fluency and semantic coherence, and LLaMA 3.1-13B-Instruct exhibiting stronger factual precision and lexical richness, while also demonstrating that carefully designed data augmentation can effectively enhance Indonesian low-resource text generation performance.

KEY WORDS - dataset augmentation, GPT-4-mini, Indonesian text generation, large language models, LLaMA 3.1-13B-Instruct, low-resource language

1. Introduction

The rapid advancement of Large Language Models (LLMs) such as GPT-4 and LLaMA has fundamentally transformed the field of Natural Language Generation (NLG) across multiple languages [1], [2]. These models have demonstrated remarkable capabilities in producing coherent, contextually appropriate, and instruction-following text, establishing new state-of-the-art benchmarks in High-Resource Languages (HRLs) such as English [3]. Despite these advances, their effectiveness in Low-Resource Languages (LRLs), including Indonesian, remains comparatively underexplored. This underrepresentation is particularly important because the linguistic complexity, morphological richness, and contextual variability of Indonesian present unique challenges for model generalization, semantic preservation, and factual consistency in large-scale generation tasks. Moreover, the increasing adoption of LLM-based applications in domains such as educational technology, digital governance, customer services, and automated content generation underscores the urgent need for accurate, culturally aligned, and linguistically robust models capable of supporting Indonesia's rapidly expanding digital ecosystem.

Dataset augmentation techniques [4], including synonym replacement, back-translation, and paraphrasing, have emerged as effective approaches for alleviating data scarcity and improving model generalization in LRL settings [5]. By enriching the diversity of training data, these methods not only increase lexical and syntactic variation but also enhance sentence fluency and reduce overfitting to narrow linguistic patterns. While extensive research has investigated augmentation and prompt engineering for English and other high-resource languages, most Indonesian NLP studies remain task-specific, focusing primarily on machine translation [6, 7], text classification [8], and sentiment analysis [9]. Although these studies demonstrate the value of augmentation for discriminative tasks, they rarely examine its influence on open-ended text generation or investigate how different instruction-tuned LLMs respond under identical augmented-data conditions. Furthermore, existing works commonly evaluate only a single model or rely on a single augmentation strategy, limiting a broader understanding of the interaction between dataset enrichment and multilingual generative capability. Consequently, a substantial research gap remains regarding how LLMs generate, adapt, and preserve semantic fidelity in Indonesian, particularly when multiple augmentation strategies are integrated within a unified evaluation framework [10], [11].

Although Indonesian is spoken by more than 280 million people [10], it remains comparatively

underrepresented in publicly available multilingual pretraining corpora and standardized NLG evaluation benchmarks relative to high-resource languages such as English or Chinese. The majority of large-scale language resources and benchmark datasets continue to be concentrated on a limited number of dominant languages, creating disparities in model adaptation, evaluation, and downstream application quality. As a result, advances achieved for HRLs cannot always be directly transferred to Indonesian without additional strategies to address linguistic and resource limitations. This imbalance provides a strong motivation for dedicated investigations into Indonesian as an LRL setting, particularly in the context of generative AI applications that require both linguistic diversity and factual reliability.

To bridge this gap, the present study systematically compares two leading instruction-tuned models, LLaMA 3.1-13B-Instruct and GPT-4-mini, for Indonesian low-resource text generation under augmented data conditions. Rather than proposing a novel augmentation technique, this research investigates whether a carefully designed dataset augmentation pipeline can serve as an effective strategy for mitigating data scarcity and improving generative performance in underrepresented languages. Three complementary augmentation strategies, namely back-translation, synonym substitution, and paraphrasing, are applied to enrich the training corpus, after which both models are evaluated under identical augmented-data conditions. Through comprehensive quantitative evaluations using BLEU, ROUGE-L, METEOR, Perplexity, mBERTScore, and lexical diversity metrics, complemented by qualitative human assessments of fluency, coherence, relevance, and factual accuracy, this study provides a multi-dimensional analysis of model behavior and the practical impact of dataset augmentation on Indonesian text generation.

The findings aim to advance understanding of the role of dataset augmentation in Indonesian low-resource text generation while offering practical insights for the development of more inclusive, accurate, and resource-efficient NLP systems for underrepresented languages. In addition, this study contributes empirical evidence regarding the complementary strengths of open-source and proprietary instruction-tuned LLMs under the same augmented-data conditions, providing a useful reference for future research and real-world deployment of generative AI technologies in Indonesian and other low-resource linguistic contexts.

2. Materials and Methods

This study was conducted in several stages, presented through a pipeline as shown in Figure 1. Figure 1 presents the overall methodological pipeline employed in this study. The workflow consists of multi-source dataset collection, text preprocessing and validation, dataset augmentation through back-translation, synonym substitution, and paraphrasing,

followed by the application of the same augmented corpus to both LLaMA 3.1-13B-Instruct and GPT-4-mini. The generated outputs are subsequently evaluated using complementary quantitative metrics and human-centered qualitative assessment, enabling a systematic comparison of model performance and the impact of dataset augmentation in Indonesian low-resource text generation.

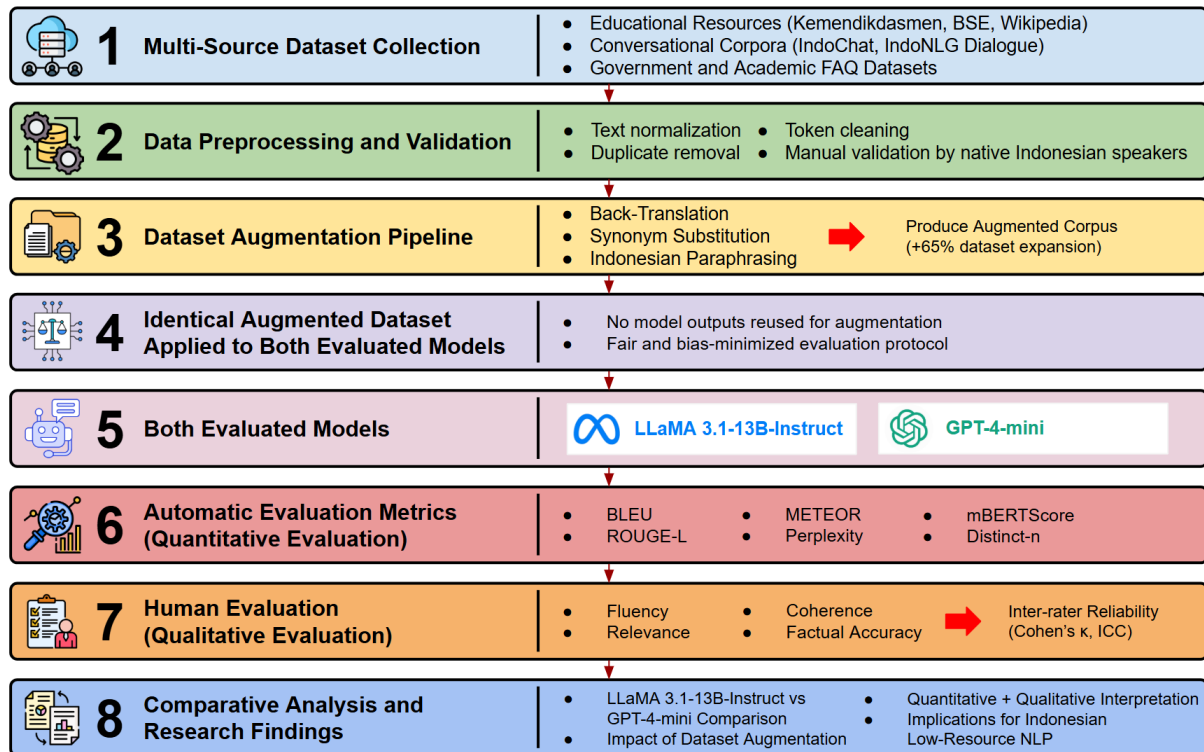


Figure 1. Overall research pipeline for the comparative evaluation of LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian low-resource text generation under augmented data conditions.

2.1 Dataset Preparation

The dataset used in this study was developed through a multi-source collection strategy to ensure both linguistic variety and contextual relevance for Indonesian text generation. The base dataset consisted of publicly available Indonesian texts obtained from several categories: (1) educational materials, including open-access Indonesian learning resources such as Kemendikdasmen Digital Library [12], Buku Sekolah Elektronik (BSE) [13], and Wikipedia Pendidikan; (2) conversational datasets, collected from open-domain chat corpora such as IndoChat and IndoNLG Dialogue Corpus [14, 15, 16], which contain natural conversational exchanges in Bahasa Indonesia; and (3) frequently asked questions (FAQs) from government and academic information portals, including Indonesia.go.id [17], Kampus Merdeka

FAQs, and university support pages that feature structured question-answer pairs. These sources were chosen to represent a wide linguistic spectrum, ranging from formal and instructional discourse commonly found in educational texts to more casual and interactive language used in dialogue and FAQ data. This diversity provides a rich foundation for evaluating the capabilities of NLG systems across different communication styles and contexts.

To address data scarcity typical of LRL, three data augmentation strategies were applied: synonym replacement, back-translation, and paraphrasing. The synonym replacement technique replaced selected words (verbs, adjectives, and nouns) with semantically equivalent alternatives based on the Kamus Besar Bahasa Indonesia (KBBI) lexical database [18], introducing lexical diversity while preserving sentence meaning. The back-translation technique translated each sentence from Indonesian into English and back into Indonesian

using a high-quality Neural Machine Translation (NMT) system (Google Translate API and IndoNMT) [19]. This method effectively generated paraphrased sentences with natural variations in structure and vocabulary. The paraphrasing approach employed transformer-based Indonesian paraphrasing checkpoints (IndoBERT [20, 21]/IndoGPT [14, 22]), which generated syntactically diverse yet semantically consistent variants of the original sentences.

To ensure methodological fairness, the dataset augmentation pipeline was designed to be entirely independent of the comparative evaluation stage. The augmented corpus was constructed using established augmentation techniques, namely back-translation, mBERT-based synonym substitution, and paraphrasing based on Indonesian language checkpoints (IndoBERT/IndoGPT), prior to any model evaluation. The resulting augmented dataset was then applied identically to both LLaMA 3.1-13B-Instruct and GPT-4-mini during the inference-based experiments. No outputs generated by either of the evaluated models were incorporated into the augmentation corpus or reused during testing, thereby minimizing the risk of evaluation bias and ensuring a fair and consistent comparative setting.

The final dataset combined the original and augmented data, resulting in a balanced corpus suitable for robust evaluation. Each augmentation technique contributed approximately one-third of the total dataset expansion, enabling comparative analysis of augmentation effectiveness. Text preprocessing was conducted to ensure consistency, including normalization (lowercasing and punctuation standardization), removal of duplicated entries, and token cleaning to eliminate extraneous symbols. A manual validation phase by native Indonesian speakers was also performed to confirm grammatical accuracy and contextual relevance. This carefully curated dataset provided a comprehensive foundation for evaluating LLaMA 3.1-13B-Instruct and GPT-4-mini on Indonesian text generation tasks under augmented data conditions.

2.2 Model Configuration

This study employed two state-of-the-art LLMs for Indonesian text generation: LLaMA 3.1-13B-Instruct and GPT-4-mini, both representing instruction-tuned variants designed to handle multilingual and context-rich NLG tasks.

The LLaMA 3.1-13B-Instruct model, developed by Meta AI, comprises approximately 13 billion parameters and was accessed through the Hugging Face Transformers API. The model was utilized in inference mode with a temperature setting of 0.7, a balanced configuration that allows for both diversity and coherence in text generation. The

checkpoint “meta-llama/Llama-3.1-13B-Instruct” was used, running under a modest computational setup consisting of a single NVIDIA T4 GPU (16 GB VRAM) environment provided via Google Colab Pro. This setup, while limited compared to high-end research infrastructure, was sufficient to perform controlled inference tasks and ensured reproducibility within a lightweight configuration suitable for academic evaluation.

The GPT-4-mini model, accessed through the OpenAI API, is a compact and efficient variant of the GPT-4 family designed to provide strong multilingual generation capabilities with relatively low computational overhead. The model was selected because of its efficiency and demonstrated effectiveness across a wide range of multilingual NLG tasks, including LRL settings such as Indonesian. To ensure methodological consistency and a fair comparison, the temperature value was fixed at 0.7 for both models, providing a balanced trade-off between output diversity and generation stability. The remaining generation parameters were retained at their default API settings (`top_p` = 1.0, `frequency_penalty` = 0, and `presence_penalty` = 0), while the maximum output length was limited to 512 tokens to ensure comparability across experimental conditions. All API calls were conducted through the OpenAI Python SDK in a stable online environment, ensuring reproducibility of results.

To ensure fairness and methodological consistency, standardized prompts were applied to both models. Each model received identical prompts written in formal Indonesian, designed to elicit coherent, contextually appropriate, and factually consistent responses. The prompt templates covered multiple linguistic scenarios, including descriptive, explanatory, and conversational text generation tasks. Both models were evaluated using the same augmented test corpus generated during the dataset preparation stage, ensuring that any observed performance differences reflected model-specific behavior rather than variations in input data or prompt design.

All model outputs were logged automatically and version-controlled for traceability. By employing a controlled inference configuration, standardized prompting, and an identical augmented evaluation corpus, this study ensured that the comparison between LLaMA 3.1-13B-Instruct and GPT-4-mini accurately reflected their generative capabilities for Indonesian low-resource text generation under augmented data conditions.

2.3 Quantitative Evaluation Metrics

The quantitative evaluation employed multiple complementary metrics to assess both surface-level and semantic quality of generated Indonesian texts. First, **Bilingual Evaluation**

Understudy (BLEU) metric measures the lexical similarity between generated and reference texts by calculating the n-gram overlap and applying a brevity penalty to discourage overly short outputs [23], [24], [25]. It computes the geometric mean of the modified n-gram precision across 1–N grams, where each precision P_n represents the ratio of matched n-grams (clipped by the maximum frequency in the reference) to the total n-grams in the candidate. The brevity penalty (BP) adjusts the score downward for outputs shorter than the reference, ensuring a balance between precision and completeness, BP is defined by Equation (1). Intuitively, BLEU assesses how well the generated text reproduces the words and short phrases found in the reference, making it suitable for evaluating lexical and phrase-level correspondence. To obtain the BLEU score, it is defined by Equation (2).

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \quad (1)$$

Where

c = the total number of tokens in the candidate.

r = effective number of tokens in the reference (usually the length of the closest reference).

$$BLEU = BP \cdot \exp(\sum_{n=1}^N W_n \log P_n) \quad (2)$$

Where

P_n = n-gram precision, which is the proportion of generated n-grams that appear in the reference.

W_n = the weight assigned to each n-gram level, commonly uniform ($W_n = 1/n$).

N = maximum n-gram length, typically set to 4.

BP = to punish sentences that are too short.

A higher BLEU score (approaching 1) signifies stronger lexical similarity and precision between the generated text and the reference, indicating that the output closely mirrors the reference in terms of word selection, phrasing, and overall language alignment.

Second, **Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L)** measures the Longest Common Subsequence (LCS) between the generated and reference texts, emphasizing semantic alignment and word order preservation [23], [26], [27]. This method is sensitive to the overall sentence structure rather than exact word matches [28]. ROUGE-L is defined by Equation (3).

$$ROUGE - L = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}}$$

Where

R_{LCS} = represents the recall of LCS.

P_{LCS} = represents the precision of LCS.

β = a parameter used to balance recall and precision, and is typically set to 1.2.

A higher ROUGE-L indicates stronger alignment between the structure and meaning of the generated and reference feedback.

Third, **Metric for Evaluation of Translation with Explicit Ordering (METEOR)** evaluates translation quality by combining unigram precision and recall while accounting for synonymy and morphological variations [23], [29], [30]. In contrast to BLEU, which emphasizes precision, METEOR penalizes both excessive and insufficient word generation, providing a more balanced assessment of output quality. The METEOR score is calculated using Equation (4).

$$METEOR = (1 - Penalty) \cdot F_{mean} \quad (4)$$

Where F_{mean} is defined as Equation (5), as follows:

$$F_{mean} = \frac{10 \cdot P \cdot R}{P + 9 \cdot R} \quad (5)$$

Where

P = unigram-based precision for matching words.

R = unigram-based recall.

$Penalty$ = fragmentation penalty is applied to penalize answers where the matching words appear in a different order from the reference.

F_{mean} = harmonic mean that assigns nine times more weight to R than to P .

A higher METEOR score indicates stronger semantic and lexical similarity, even when different but synonymous words are used.

Fourth, **Perplexity** measures how well a model predicts the actual sequence of tokens, where lower values indicate greater confidence in correct token prediction [31], [32]. It reflects the model's ability to assign high probabilities to the correct tokens and is widely used to assess language model quality [33]. The exponential transformation of the average negative log-likelihood converts it into a more interpretable scale, representing the model's effective "branching factor," or the average number of equally probable word choices considered at each prediction step. Thus, a lower perplexity value indicates higher confidence and fluency in generation, while a higher value reflects greater

uncertainty and weaker predictive performance. To obtain the perplexity score, it is defined by Equation (6).

$$Perplexity = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log p(w_i | w < i)\right) \quad (6)$$

Where

- N = the total number of tokens (words or subwords) in the text being evaluated.
- w_i = represents the i^{th} word in the sequence.
- $w < i$ = denotes the sequence of all previous words before w_i .
- $p(w_i | w < i)$ = the conditional probability assigned by the model to the word w_i given its preceding context.
- $\log$ = typically in natural base (ln), and the negative average log probability measures the model's uncertainty.

In this study, perplexity was computed over all generated sentences for both LLaMA 3.1-13B-Instruct and GPT-4-mini to assess their linguistic modeling efficiency under Indonesian language conditions.

Fifth, **Multilingual BERT (mBERT)** is a variant of the BERT model designed to generate contextual text representations effectively across multiple languages [34]. The model was pre-trained on 104 languages from Wikipedia, including Indonesian, and has proven effective for various multilingual tasks such as cross-lingual natural language inference, named entity recognition, and dependency parsing [35]. In this study, mBERT is utilized through the Multilingual BERTScore (mBERTScore) metric to evaluate the semantic similarity between the generated and reference texts. By comparing deep contextual representations, mBERTScore enables a more accurate assessment of meaning preservation, synonym usage, and paraphrasing quality, particularly in LRL contexts. The mBERTScore value is computed using Equation (7).

$$mBERTScore = \frac{1}{|X|} \sum_{x_i \in X} \max_{y_j \in Y} \cos(e(x_i), e(y_j)) \quad (7)$$

Where

- X = $\{x_1, x_2, \dots, x_m\}$ is the set of tokens in the generated text.
- Y = $\{y_1, y_2, \dots, y_n\}$ is the set of tokens in the reference text.
- $e(x_i)$ = the embedding vectors of token x_i , obtained from the mBERT encoder.

$e(y_j)$ = the embedding vectors of token y_j , obtained from the mBERT encoder.

$\cos(e(x_i), e(y_j))$ = denotes the cosine similarity between the two contextual embeddings.

The score for each token x_i is determined by its maximum similarity to any token in Y , followed by averaging all tokens in X .

Sixth, to evaluate the lexical diversity of the generated Indonesian text, this study employs the **Distinct-n** metric, a simple yet effective measure of variation across model outputs. The intuition behind this metric is to quantify how frequently a model repeats identical n-grams (sequences of n words) in its generated texts. High repetition (low diversity) often reflects degeneracy, where the model produces repetitive or formulaic sentences, whereas high Distinct-n values indicate more varied and creative text generation [36], [37]. The Distinct-n value is computed using Equation (8).

$$Distinct - n = \frac{|U|}{|T|} \quad (8)$$

Where

- n = size of the word sequence (e.g., 1 for unigram, 2 for bigram).
- U = number of distinct n-grams appearing at least once in the generated text (unique n-grams in all outputs).
- T = total number of n-grams (including repetitions) across all outputs (total n-grams in all outputs).

For example, Distinct-1 measures the proportion of unique individual words (unigrams), while Distinct-2 evaluates unique two-word combinations (bigrams). The resulting score ranges between 0 and 1, where higher values indicate greater lexical diversity [38].

In the context of this study, Distinct-n serves as a complementary metric to BLEU, ROUGE, and mBERTScore, focusing not on similarity but on variety and richness of expression in Indonesian language generation. A model with a high diversity score demonstrates better capability to generate non-repetitive, contextually flexible sentences, which is crucial in creative or dialogic NLG tasks. However, it is important to note that Distinct-n does not account for semantic coherence or factual correctness; a model might achieve high diversity by producing linguistically varied but semantically nonsensical outputs.

Therefore, in this research, Distinct-n is used jointly with fluency and semantic similarity metrics to provide a balanced evaluation of both

linguistic richness and content quality for LLaMA 3.1-13B-Instruct and GPT-4-mini in Indonesian text generation.

For quantitative evaluation, the generated outputs were compared against reference texts using BLEU, ROUGE-L, METEOR, Perplexity, mBERTScore, and Distinct-n metrics computed via the NLTK and Evaluate (Hugging Face) libraries in Python.

2.4 Qualitative Evaluation

In addition to the quantitative metrics, a human-based qualitative evaluation was conducted to assess the linguistic and contextual quality of model outputs that could not be fully captured by automated scores [39], [40], [41]. This evaluation aimed to provide a deeper understanding of how effectively each model generates coherent, meaningful, and factually [42, 43] accurate Indonesian text.

A subset of 200 text outputs was randomly sampled from the complete test results, ensuring balanced representation across various prompt categories such as descriptive, explanatory, and conversational tasks. The evaluation was performed by a group of five independent human raters, all of whom were native Indonesian speakers with academic backgrounds in linguistics, language technology, or data science. The use of five evaluators was intended to improve the statistical robustness of the assessment and reduce ambiguity in majority-based judgments. Prior to the evaluation, the raters participated in a calibration session to

establish a common understanding of the scoring criteria and reduce subjective interpretation bias.

Each generated output was assessed independently based on four key evaluation criteria, adapted from established NLG evaluation frameworks [44]:

1. **Fluency** - measures the grammaticality, naturalness, and readability of the generated text in Indonesian. Texts with minimal grammatical errors, smooth word choice, and native-like phrasing received higher scores.
2. **Relevance** - assesses how well the generated response aligns with the input prompt or question. Outputs that directly address the prompt with relevant information were rated higher than those containing tangential or generic responses.
3. **Coherence** - evaluates the logical consistency and organization of ideas across sentences. High-scoring outputs maintain smooth transitions and clear sentence-to-sentence connections without abrupt topic shifts.
4. **Factual Accuracy** - determines the correctness of factual content within the text. This includes verifying names, dates, events, and domain-specific knowledge where applicable. Factual errors or hallucinations led to lower ratings.

Each criterion was scored using a five-point Likert scale, where 1 = very poor, 2 = poor, 3 = fair, 4 = good, and 5 = excellent. The scale was designed to capture fine-grained differences in quality across the models [45]. An example of the evaluation rubric is presented in Table 1.

Table 1. Human evaluation rubric for generated text quality

Score	Fluency	Relevance	Coherence	Factual Accuracy
5	Perfectly natural and error-free Indonesian	Fully addresses prompt	Logically structured and consistent	Factually accurate
4	Minor grammatical or lexical issues	Mostly relevant	Mostly coherent	Minor factual imprecision
3	Noticeable linguistic errors	Partially relevant	Some inconsistencies	Some factual inaccuracy
2	Frequent grammatical or word choice errors	Largely irrelevant	Disjointed ideas	Multiple factual errors
1	Unintelligible or incorrect Indonesian	Not relevant	Illogical	Factually incorrect

To minimize evaluation bias, all samples were randomized and anonymized, meaning raters were not informed about which model produced each

output. The final human evaluation dataset contained a total of $200 \times 5 = 1,000$ individual rating entries

across all evaluators, providing a sufficiently robust basis for inter-rater agreement analysis.

To ensure consistency and reliability in human judgments, inter-rater reliability was calculated using Cohen's Kappa (κ) for pairwise agreement and Intraclass Correlation Coefficient (ICC) for group-level consistency.

Cohen's Kappa (κ) measures the level of agreement between raters beyond chance for categorical or ordinal data [46, 47], calculated using Equation (9).

$$k = \frac{p_o - p_e}{1 - p_e} \quad (9)$$

Where

p_o = the observed proportion of agreement.

p_e = the expected agreement by chance.

κ values above 0.75 are considered excellent, between 0.4-0.75 as fair to good, and below 0.4 as poor agreement.

Intraclass Correlation Coefficient (ICC) was applied to measure rating consistency among multiple evaluators on a continuous scale (Likert 1-5) [48]. The general form of ICC is defined by Equation (10).

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W} \quad (10)$$

Where

MS_B = the mean square between raters.

MS_W = the mean square within raters.

k = the number of raters.

ICC values range from 0 to 1, with higher values indicating stronger inter-rater consistency [49].

Finally, the mean and standard deviation of the scores across all criteria were computed for each model to produce a comprehensive qualitative performance profile. Correlation analysis between human ratings and quantitative metrics was also conducted to assess the extent to which automated metrics reflect human perceptions of output quality. This human-centered evaluation procedure provided a critical interpretive layer to the study, complementing the quantitative metrics and ensuring a holistic understanding of model performance in Indonesian NLG.

For qualitative evaluation, a stratified subset of 200 outputs was extracted and anonymized before being independently rated by human evaluators. The full dataset, model outputs, and evaluation scripts were version-controlled using Git to ensure experimental traceability and replicability.

3. Results and Discussion

3.1 Quantitative Results

To assess the comparative performance of both models, a comprehensive quantitative evaluation was conducted using multiple automatic metrics. The quantitative evaluation results are summarized in Table 2. GPT-4-mini consistently outperformed LLaMA 3.1-13B-Instruct across most automatic metrics, achieving higher BLEU (0.55 vs 0.52), ROUGE-L (0.64 vs 0.61), METEOR (0.57 vs 0.54), lower Perplexity (11.5 vs 12.8), and higher mBERTScore (0.90 vs 0.88). However, LLaMA 3.1-13B-Instruct demonstrated superior lexical diversity, with higher Distinct-1 and Distinct-2 values, indicating richer vocabulary use and lower repetition.

Table 2. Quantitative performance comparison of LLaMA 3.1-13B-Instruct and GPT-4-mini

Metric	Description	LLaMA 3.1-13B-Instruct	GPT-4-mini
BLEU ↑	Lexical overlap with reference	0.52	0.55
ROUGE-L ↑	Longest common subsequence recall	0.61	0.64
METEOR ↑	Precision/recall with synonym and stemming	0.54	0.57
Perplexity ↓	Model confidence in next-token prediction	12.8	11.5
mBERTScore ↑	Semantic similarity (embedding-based)	0.88	0.90
Distinct-1 ↑	Unigram diversity	0.42	0.39
Distinct-2 ↑	Bigram diversity	0.34	0.31

(↑ = higher is better; ↓ = lower is better)

These results indicate that GPT-4-mini generates text with greater alignment and semantic coherence to reference outputs, reflected in its higher BLEU, ROUGE-L, METEOR, and mBERTScore. Its lower Perplexity value also shows higher confidence in predicting tokens under Indonesian syntax, suggesting stronger language modeling capability despite being a compact model.

Conversely, LLaMA 3.1-13B-Instruct generated more lexically diverse outputs, reflecting its instruction-tuning design that emphasizes creative phrasing and open-ended text generation. Its higher Distinct-n scores suggest a greater capacity to reduce repetitive expressions, which is an important characteristic for maintaining output diversity in open-domain NLG applications.

From a linguistic perspective, GPT-4-mini's advantage in semantic metrics suggests better understanding of contextual relationships, possibly due to its larger multilingual training distribution and stronger fine-tuning alignment. LLaMA 3.1, however, retains an advantage in expressive variability, which can be beneficial for creative or conversational Indonesian tasks.

3.2 Qualitative Results

Human evaluation results across four criteria, namely fluency, relevance, coherence, and factual accuracy, are presented in Table 3. The mean scores were calculated based on the assessment of 200 samples by five independent human evaluators.

Table 3. Human evaluation rubric for generated text quality

Criterion	LLaMA 3.1-13B-Instruct	GPT-4-mini
Fluency	4.1 ± 0.4	4.5 ± 0.3
Relevance	4.3 ± 0.3	4.4 ± 0.3
Coherence	4.0 ± 0.5	4.4 ± 0.3
Factual Accuracy	4.5 ± 0.2	4.2 ± 0.3

Inter-rater reliability was high, with Cohen's $\kappa = 0.82$ and ICC = 0.88, indicating excellent agreement across evaluators.

Overall, GPT-4-mini achieved higher ratings in fluency, relevance, and coherence, reflecting its ability to produce smooth, logically connected, and contextually appropriate Indonesian sentences. LLaMA 3.1-13B-Instruct, while slightly less fluent, excelled in factual accuracy, suggesting a more grounded response style that prioritizes information correctness over stylistic flow.

These findings align with the quantitative metrics: GPT-4-mini's stronger BLEU, ROUGE, and mBERTScore correspond with higher human-perceived fluency and coherence, while LLaMA's higher Distinct-n correlates with more varied lexical choices observed by evaluators.

3.3 Data Augmentation Simulation and Results

To enhance the robustness of Indonesian low-resource text generation, a controlled simulation of data augmentation was performed prior to the comparative evaluation stage. The augmentation process aimed to increase lexical and syntactic diversity while preserving the semantic integrity of the original corpus. Three complementary

augmentation strategies were employed: (1) back-translation using English and Malay as pivot languages through Google Translate and IndoNMT, (2) synonym substitution based on pretrained mBERT lexical embeddings, and (3) paraphrasing using Indonesian transformer-based checkpoints (IndoBERT/IndoGPT), which generate syntactically diverse yet semantically consistent sentence variants. The dataset focused on Indonesian history, national profile, and general knowledge topics, sourced from publicly available educational and informational portals. This augmentation strategy was intended to enrich linguistic variability while maintaining factual consistency and contextual relevance.

The augmentation pipeline was conducted independently of the model evaluation stage. The augmented corpus was created entirely before experimentation and was subsequently applied identically to both LLaMA 3.1-13B-Instruct and GPT-4-mini during the inference-based evaluation. No outputs generated by either evaluated model were incorporated into the augmentation corpus or reused during testing, thereby minimizing the risk of evaluation bias and ensuring a fair comparative setting. The augmented corpus increased the dataset size by approximately 65%, generating around

18,000 additional sentence pairs from the original 27,500 samples. Table 4 presents representative

examples of sentence transformations from the original dataset to their augmented variants.

Table 4. Examples of sentence transformations

Original Sentence	Augmented Sentence
Proklamasi kemerdekaan Indonesia terjadi pada 17 Agustus 1945.	Indonesia memproklamasikan kemerdekaannya pada tanggal 17 Agustus tahun 1945.
Pulau Jawa merupakan pusat pemerintahan dan ekonomi Indonesia.	Pusat kegiatan ekonomi dan pemerintahan Indonesia berada di Pulau Jawa.
Soekarno adalah presiden pertama Republik Indonesia.	Presiden pertama Republik Indonesia adalah Ir. Soekarno.
Lagu kebangsaan Indonesia berjudul “Indonesia Raya”.	“Indonesia Raya” merupakan lagu kebangsaan negara Indonesia.
Candi Borobudur terletak di provinsi Jawa Tengah.	Di provinsi Jawa Tengah terdapat Candi Borobudur, salah satu warisan budaya dunia.

As shown in Table 4, the augmented sentences preserve the semantic content of the original texts while introducing structural and lexical variations. This process effectively broadens the linguistic coverage of the dataset, enabling the evaluation of model performance under more diverse expression patterns without altering the underlying meaning.

To further illustrate the practical effect of the augmented-data setting, representative outputs generated by LLaMA 3.1-13B-Instruct and GPT-4-mini are presented in Table 5. The examples demonstrate how the enriched dataset supports more fluent, contextually appropriate, and lexically diverse generation while maintaining the factual content of the original prompts.

Table 5. Representative model outputs under augmented data conditions

Prompt / Original Information	Augmented Context Example	LLaMA 3.1-13B-Instruct Output	GPT-4-mini Output
Proklamasi kemerdekaan Indonesia terjadi pada 17 Agustus 1945.	Indonesia memproklamasikan kemerdekaannya pada tanggal 17 Agustus tahun 1945.	Proklamasi kemerdekaan Indonesia dilaksanakan pada 17 Agustus 1945 dan menjadi tonggak berdirinya Republik Indonesia.	Indonesia secara resmi memproklamasikan kemerdekaannya pada 17 Agustus 1945, menandai lahirnya negara yang berdaulat.
Soekarno adalah presiden pertama Republik Indonesia.	Presiden pertama Republik Indonesia adalah Ir. Soekarno.	Ir. Soekarno dikenal sebagai presiden pertama Republik Indonesia dan salah satu tokoh utama perjuangan kemerdekaan.	Soekarno merupakan presiden pertama Republik Indonesia yang berperan penting dalam proses kemerdekaan bangsa.
Candi Borobudur terletak di provinsi Jawa Tengah.	Di provinsi Jawa Tengah terdapat Candi Borobudur, salah satu warisan budaya dunia.	Candi Borobudur berada di Jawa Tengah dan dikenal sebagai salah satu peninggalan Buddha	Terletak di provinsi Jawa Tengah, Candi Borobudur merupakan salah satu situs warisan budaya dunia yang

		terbesar di dunia.	terkenal di Indonesia.
--	--	--------------------	------------------------

As illustrated in Table 5, both models successfully preserve the factual content of the original prompts while benefiting from the richer contextual information introduced through dataset augmentation. GPT-4-mini generally produces more fluent and naturally connected descriptions, whereas LLaMA 3.1-13B-Instruct tends to generate outputs with greater lexical variety and explicit factual emphasis. These representative examples complement the quantitative findings by demonstrating how augmented data conditions improve the overall quality and diversity of Indonesian low-resource text generation.

To assess the impact of augmented data conditions, both LLaMA 3.1-13B-Instruct and GPT-4-mini were evaluated using the original dataset and the augmented dataset under the same experimental protocol. Their outputs were compared using BLEU,

ROUGE-L, METEOR, Perplexity, mBERTScore, and Distinct-n metrics. As summarized in Table 6, the augmented dataset produced consistent improvements across multiple evaluation metrics for both models. GPT-4-mini demonstrated a 4.2% increase in BLEU and a 7.1% reduction in Perplexity, indicating stronger lexical alignment and more stable token prediction. Meanwhile, LLaMA exhibited notable gains in lexical diversity, with Distinct-2 increasing from 0.29 to 0.34, suggesting a greater capacity to generate varied expressions while reducing repetitive output patterns. The stronger Distinct-1 and Distinct-2 scores also indicate that LLaMA 3.1-13B-Instruct is less susceptible to the degeneration phenomenon commonly observed in neural text generation, where models repeatedly generate identical or highly similar expressions across generated outputs.

Table 6. Effect of data augmentation on model performance

Model	Dataset Type	BLEU	ROUGE-L	METEOR	Perplexity (↓)	mBERTScore	Distinct-2 (↑)
LLaMA 3.1-13B-Instruct	Original	0.50	0.59	0.52	13.8	0.86	0.29
	Augmented	0.52	0.61	0.54	12.8	0.88	0.34
GPT-4-mini	Original	0.53	0.61	0.55	12.4	0.88	0.28
	Augmented	0.55	0.64	0.57	11.5	0.90	0.31

These results confirm that data augmentation contributes to improving both syntactic fluency and semantic coverage in Indonesian low-resource contexts. While GPT-4-mini benefits more from augmented contextual richness, LLaMA appears to leverage the expanded lexical variety, producing outputs with greater factual stability and vocabulary breadth. The complementary impact of augmentation thus enhances the overall generalization capability of both models in handling Indonesian text generation tasks.

3.4 Discussion

Impact of Dataset Augmentation - Both models benefited from dataset augmentation, which improved linguistic robustness and reduced overfitting to narrow lexical patterns. The integration of three complementary augmentation strategies, namely back-translation, synonym substitution, and paraphrasing, provided distinct yet mutually

reinforcing advantages. Back-translation introduced alternative syntactic structures through multilingual transformations, increasing sentence variation while preserving semantic content. Synonym substitution, supported by mBERT-based lexical embeddings, enriched lexical diversity by replacing contextually appropriate words with semantically similar alternatives. Meanwhile, paraphrasing using Indonesian transformer-based checkpoints (IndoBERT/IndoGPT) exposed the models to structurally diverse sentence formulations, promoting better generalization across different prompt styles while maintaining semantic consistency. Collectively, these augmentation strategies expanded the expressive range and linguistic coverage of the Indonesian corpus, contributing to improved syntactic fluency, lexical richness, and semantic robustness under low-resource text generation conditions, in line with prior findings on augmentation for LRL modeling [5], [14], [21].

Comparative Performance Analysis - The complementary strengths observed, namely GPT-4-

mini's fluency and LLaMA's factual accuracy, likely reflect differences in their underlying architectures, alignment strategies, and multilingual training paradigms. GPT-4-mini appears to generate smoother and more contextually balanced outputs, whereas LLaMA 3.1-13B-Instruct tends to prioritize content precision and factual grounding, occasionally at the expense of stylistic smoothness. Perplexity analysis supports this interpretation: GPT-4-mini's lower perplexity (11.5) indicates stronger probabilistic confidence in token prediction, whereas LLaMA's higher perplexity (12.8) suggests slightly greater uncertainty when handling longer or more complex syntactic structures. These findings point to a practical trade-off between probabilistic fluency and factual reliability, a pattern that has also been reported in multilingual text generation research.

Semantic and Lexical Trends - Semantic evaluation using mBERTScore confirmed GPT-4-mini's advantage in capturing embedding-level similarity to reference texts, suggesting stronger contextual alignment. At the same time, the higher Distinct-1 and Distinct-2 scores indicate that LLaMA 3.1-13B-Instruct is less susceptible to the degeneration phenomenon commonly observed in neural text generation, where models repeatedly generate identical or highly similar expressions across outputs. This balance between semantic precision and lexical flexibility is consistent with previous studies [36], [37], which report that models optimized for fluency and alignment often generate less diverse outputs, whereas models with broader lexical representations tend to produce greater variation while sacrificing a small degree of stylistic naturalness. In the present study, GPT-4-mini demonstrated stronger semantic cohesion, while LLaMA generated more lexically diverse and factually grounded responses.

Human–Machine Correlation - Correlation analysis between human ratings and automated metrics reinforces the robustness of the evaluation. Pearson's correlation coefficient between human-rated fluency and automated fluency-related metrics (BLEU, ROUGE-L, METEOR) reached $r = 0.82$, indicating a strong positive relationship between quantitative scores and human perception of linguistic naturalness. Conversely, factual accuracy exhibited a weaker correlation with mBERTScore ($r = 0.48$), underscoring a limitation of embedding-based similarity metrics for verifying factual correctness. This gap highlights the importance of incorporating human evaluation when assessing factual quality, especially in generative tasks involving informational or historical content.

Implications - These findings indicate that both GPT-4-mini and LLaMA 3.1-13B-Instruct are effective models for Indonesian low-resource text generation under augmented data conditions, each

exhibiting distinct advantages. GPT-4-mini's strengths in fluency, coherence, and contextual naturalness make it well suited for educational, conversational, and other user-facing applications that require smooth and context-aware responses. In contrast, LLaMA's stronger factual grounding and lexical diversity make it particularly suitable for information-oriented or knowledge-intensive generation tasks, such as historical narration or educational content creation. More broadly, the results demonstrate that a carefully designed dataset augmentation pipeline can serve as an effective and practical strategy for mitigating data scarcity, improving linguistic coverage, and enhancing LLM performance in LRL settings.

In summary, the comparative evaluation reveals a complementary relationship between the two models. GPT-4-mini excels in generating contextually fluent, coherent, and semantically aligned text, whereas LLaMA 3.1-13B-Instruct demonstrates stronger factual reliability and greater lexical diversity. Collectively, these findings reinforce the value of dataset augmentation as a practical mechanism for enhancing Indonesian low-resource text generation, improving both linguistic coverage and model generalization under data-scarce conditions.

4. Conclusion and Future Works

This study presented a comprehensive comparative evaluation of LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian low-resource text generation under augmented data conditions, using an augmented dataset designed to mitigate the limitations of data scarcity. Quantitative analyses revealed that GPT-4-mini achieved superior results in BLEU (0.55 vs 0.52), ROUGE-L (0.64 vs 0.61), METEOR (0.57 vs 0.54), Perplexity (11.5 vs 12.8), and mBERTScore (0.90 vs 0.88), indicating stronger lexical alignment, semantic coherence, and predictive fluency. Meanwhile, LLaMA 3.1 exhibited higher lexical diversity (Distinct-2 = 0.34 vs 0.31) and greater factual precision according to human evaluation. The qualitative assessment by five independent human raters further emphasized GPT-4-mini's advantages in fluency (4.5 vs 4.1), coherence (4.4 vs 4.0), and relevance (4.4 vs 4.3), while LLaMA demonstrated slightly stronger factual accuracy (4.5 vs 4.2). These results highlight the complementary strengths of the two models: GPT-4-mini's generative naturalness and contextual cohesion versus LLaMA's factual rigor and vocabulary diversity, both of which benefit significantly from the augmented-data setting, where complementary strategies including back-translation, synonym substitution, and paraphrasing enrich lexical and semantic diversity.

Future research should further investigate how advanced augmentation strategies and multilingual transfer learning from related Austronesian languages, such as Malay, Javanese, and Tagalog, can improve Indonesian low-resource text generation. Exploring instruction-tuning and alignment techniques, including Reinforcement Learning from Human Feedback (RLHF) and contrastive prompt optimization, may further enhance factual consistency and stylistic flexibility. In addition, extending augmentation beyond lexical and syntactic variation toward semantic and contextual enrichment, together with larger-scale human evaluation frameworks, could contribute to the development of more robust, inclusive, and resource-efficient LLMs for underrepresented languages.

References

- [1] Y. Zhao, W. Zhang, G. Chen, K. Kawaguchi, and L. Bing, "How do large language models handle multilingualism?," In Proc. 38th Conference on Neural Information Processing Systems, 2024, pp. 1-24.
- [2] L. Qin, Q. Chen, Y. Zhou, Z. Chen, Y. Li, L. Liao, M. Li, W. Che, and P. S. Yu, "A survey of multilingual large language models," *Patterns*, vol. 6, no. 1, pp. 1-30, 2025.
- [3] Y. Tu, A. Xue, and F. Shi, "Blessing of multilinguality: A systematic analysis of multilingual in-context learning," In Proc. 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 6213-6248.
- [4] M. A. Ibrahim, Faisal, Z. D. Sulistiya, and T. S. Y. Winarto, "Prompt-based data augmentation with large language models for Indonesian gender-based hate speech detection," *Journal of Computer Science*, vol. 20, no. 8, pp. 819-826, 2024.
- [5] R. Kimera, D. N. Heo, D. N. Rim, and H. Choi, "Data augmentation with back translation for low resource languages: A case of english and Luganda," In Proc. 8th International Conference on Natural Language Processing and Information Retrieval, 2025, pp. 142-148.
- [6] F. I. Maulana, Y. Heryadi, G. P. Kusuma, and W. Budiharto, "Data augmentation English-Indonesia-Madurese parallel corpus dataset using neural machine translation," *Data in Brief*, vol. 62, no. 1, pp. 1-8, 2025.
- [7] P. N. Hadiwinoto and D. P. Lestari, "Data augmentation on spontaneous Indonesian automatic speech recognition using statistical machine translation," In Proc. International Conference on Information Technology and Digital Applications, 2019, pp. 1-8.
- [8] F. Muftie and M. Haris, "IndoBERT based data augmentation for Indonesian text classification," In Proc. International Conference on Information Technology Research and Innovation, 2023, pp. 128-132.
- [9] A. H. Nasution, A. Onan, Y. Murakami, W. Monika, and A. Hanafiah, "Benchmarking open-source large language models for sentiment and emotion classification in Indonesian tweets," *IEEE Access*, vol. 13, no. 1, pp. 94009-94025, 2025.
- [10] A. S. Wijaya and A. S. Girsang, "Augmented-based Indonesian abstractive text summarization using pre-trained model mT5," *International Journal of Engineering Trends and Technology*, vol. 71, no. 11, pp. 190-200, 2023.
- [11] B. R. Irnawan and R. Adi, "Improving Indonesian informal to formal style transfer via pre-training unlabelled augmented data," In Proc. 6th International Conference of Computer and Informatics Engineering, 2023, pp. 25-29.
- [12] KEMENDIKDASMEN, "Perpustakaan digital," Ministry of Primary and Secondary Education Republic of Indonesia, 2025. [Online]. Available: <https://pustaka-digital.kemendikdasmen.go.id/>. [Accessed: Sept. 2, 2025].
- [13] MYEDISI, "Buku Sekolah Elektronik (BSE)," My Edisi, 2021. [Online]. Available: <https://www.myedisi.com/bse>. [Accessed: Sept. 3, 2025].
- [14] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," In Proc. Conference on Empirical Methods in Natural Language Processing, 2021, pp. 8875-8898.
- [15] B. Wilie, S. Cahyawijaya, and G. I. Winata, "IndoNLG-Indonesian natural language generation," github.com, 2024. [Online]. Available: <https://github.com/IndoNLP/indonlg>. [Accessed: Sept. 1, 2025].
- [16] INDODATASET, "Indonesian datasets: NLP datasets for Indonesian," github.com, 2023. [Online]. Available: https://github.com/Wikidepia/indonesian_dataset. [Accessed: Sept. 2, 2025].
- [17] INDONESIAGOID, "Portal informasi Indonesia," Indonesia Gov, 2025. [Online]. Available: <https://indonesia.go.id/>. [Accessed: Sept. 1, 2025].
- [18] KBBI, "Kamus Besar Bahasa Indonesia (KBBI) VI daring," Indonesian Dictionary, 2025. [Online]. Available: <https://kbbi.kemdikbud.go.id/>. [Accessed: Sept. 1, 2025].
- [19] M. Dwiastuti, "English-Indonesian neural machine translation for spoken language

- domains,” In Proc. 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 309-314.
- [20] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoBERT: Indonesian version of BERT model,” indolem.github.io, 2020. [Online]. Available: <https://indolem.github.io/Task/>. [Accessed: Sept. 2, 2025].
- [21] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “Liputan6: A large-scale Indonesian dataset for text summarization,” In Proc. 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 598-608.
- [22] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, “IndoGPT Model,” huggingface.co, 2021. [Online]. Available: <https://huggingface.co/indobenchmark/indogpt>. [Accessed: Sept. 2, 2025].
- [23] A. Onan and H. Alhumyani, “Knowledge-enhanced transformer graph summarization (KETGS): Integrating entity and discourse relations for advanced extractive text summarization,” *Mathematics*, vol. 12, no. 23, pp. 1-25, 2024.
- [24] A. A. Syahidi and K. Kiyokawa, “Automatic text generation in Banjar language using GPT-4 for low-resource language preservation,” In Proc. 7th IEEE Symposium on Computers and Informatics, 2025, pp. 238-243.
- [25] A. Gupta, A. Rastogi, H. Malhotra, and K. Rangarajan, “Comparative evaluation of large language models for translating radiology reports into Hindi,” *Indian Journal of Radiology and Imaging*, vol. 35, no. 1, pp. 88-96, 2024.
- [26] A. A. Citarella, M. Barbella, M. G. Ciobanu, F. D. Marco, L. D. Biasi, and G. Tortora, “Assessing the effectiveness of ROUGE as unbiased metric in extractive vs. abstractive summarization techniques,” *Journal of Computational Science*, vol. 87, no. 102571, pp. 1-17, 2025.
- [27] E. Oro, F. M. Granata, and M. Ruffolo, “A comprehensive evaluation of embedding models and LLMs for IR and QA across English and Italian,” *Big Data and Cognitive Computing*, vol. 9, no. 141, pp. 1-41, 2025.
- [28] A. A. Syahidi, K. Kiyokawa, and S. Nuchitprasitchai, “A fine-tuned GPT-4-based question answering system for e-government services using a custom-built dataset,” In Proc. 7th IEEE Symposium on Computers and Informatics, 2025, pp. 232-237.
- [29] M. Mahyoub, Y. Wang, and M. T. Khasawneh, “GPT-4o in radiology: In-context learning based automatic generation of radiology impressions,” *Natural Language Processing Journal*, vol. 11, no. 100145, pp. 1-10, 2025.
- [30] P. Netisopakul and U. Taoto, “Comparison of evaluation metrics for short story generation,” *IEEE Access*, vol. 11, no. 1, pp. 140253-140269, 2023.
- [31] P. S. García-Montero, P. Vizcaíno, I. G. Reyes-Chacón, and M. E. Morochó-Cayamcela, “AI for all: Reducing perplexity and boosting accuracy in normative texts with fine-tuned LLMs and RAG,” *IEEE Access*, vol. 13, no. 1, pp. 179759-179775, 2025.
- [32] L. Fang, Y. Wang, Z. Liu, C. Zhang, S. Jegelka, J. Gao, B. Ding, and Y. Wang, “What is wrong with perplexity for long-context language modeling?,” In Proc. 13th International Conference on Learning Representations, 2025, pp. 1-23.
- [33] J. Xu, H. Zhang, Y. Yang, L. Yang, Z. Cheng, J. Lyu, B. Liu, X. Zhou, A. Bacchelli, Y. K. Chiam, and T. K. Chiew, “One size does not fit all: Investigating efficacy of perplexity in detecting LLM-generated code,” *ACM Transactions on Software Engineering and Methodology*, vol. 1, no. 1, pp. 1-33, 2025.
- [34] A. Aziz, M. A. Hossain, and A. N. Chy, “CSECU-DSG at SemEval-2022 task 3: Investigating the taxonomic relationship between two arguments using fusion of multilingual transformer models,” In Proc. 16th International Workshop on Semantic Evaluation, 2022, pp. 255-259.
- [35] E. A. Chi, J. Hewitt, and C. D. Manning, “Finding universal grammatical relations in multilingual BERT,” In Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5564-5577.
- [36] S. Liu, S. Sabour, Y. Zheng, P. Ke, X. Zhu, and M. Huang, “Rethinking and refining the distinct metric,” In Proc. 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 762-770.
- [37] A. Shypula, S. Li, B. Zhang, V. Padmakumar, K. Yin, and O. Bastani, “Evaluating the diversity and quality of LLM generated content,” In Proc. International Conference on Learning Representations, 2025, pp. 1-18.
- [38] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, “A diversity-promoting objective function for neural conversation models,” In Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 110-119.
- [39] J. Kim and W. Jung, “CARE: A framework for correcting numerical hallucinations in LLM-generated financial texts,” In Proc. IEEE

- Conference on Artificial Intelligence, 2025, pp. 69-74.
- [40] M. Y. Mohammed, S. A. Ali, S. K. Ali, A. A. Majeed, and E. H. Mohamed, "Aftina: Enhancing stability and preventing hallucination in AI based islamic fatwa generation using LLMs and RAG," *Neural Computing and Applications*, vol. 37, no. 1, pp. 20957-20982, 2025.
- [41] I. Jahan, M. T. R. Laskar, C. Peng, and J. X. Huang, "A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks," *Computers in Biology and Medicine*, vol. 171, no. 108189, pp. 1-23, 2024.
- [42] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, P. Mathur, G. E. Cacciamani, C. Sun, Y. Peng, and Y. Wang, "A framework for human evaluation of large language models in healthcare derived from literature review," *npj - Digital Medicine*, vol. 7, no. 258, pp. 1-20, 2024.
- [43] Q. Li, L. Cui, L. Kong, and W. Bi, "Exploring the reliability of large language models as customized evaluators for diverse NLP tasks," In Proc. 31st International Conference on Computational Linguistics, 2025, pp. 10325-10344.
- [44] Z. Li, X. Xu, T. Shen, C. Xu, J-C Gu, Y. Lai, C. Tao, and S. Ma, "Leveraging large language models for NLG evaluation: Advances and challenges," In Proc. Conference on Empirical Methods in Natural Language Processing, 2024, pp. 16028-16045.
- [45] G. Tevet and J. Berant, "Evaluating the evaluation of diversity in natural language generation," In Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021, pp. 326-346.
- [46] B. Więckowska, K. B. Kubiak, P. Jóźwiak, W. Moryson, and B. Stawińska-Witoszyńska, "Cohen's kappa coefficient as a measure to assess classification improvement following the addition of a new marker to a regression model," *International Journal of Environmental Research and Public Health*, vol. 19, no. 10213, pp. 1-15, 2022.
- [47] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, and D. Hupkes, "Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges," In Proc. Fourth Workshop on Generation, Evaluation and Metrics, 2025, pp. 404-430.
- [48] V. Hackl, A. E. Müller, M. Granitzer, and M. Sailer M, "Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings," *Frontiers in Education*, vol. 8, no. 1272229, pp. 1-8, 2023.
- [49] D-W. Zhang, M. Boey, Y. Y. Tan, and A. H. S. Jia, "Evaluating large language models for criterion-based grading from agreement to consistency," *npj - Science of Learning*, vol. 9, no. 79, pp. 1-4, 2024.