Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

Comparison of IndoNanoT5 and IndoGPT for Advancing Indonesian Text Formalization in Low-Resource Settings

Fahri Firdausillah[*1], Ardytha Luthfiarta[2], Adhitya Nugraha[3], Ika Novita Dewi[4], Lutfi Azis Hafiizhudin[5], Najma Amira Mumtaz[6], Ulima Muna Syarifah[7]

[1,2,3,4,5,6,7]Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Email: [1]fahri@dsn.dinus.ac.id

Received : Jun 23, 2025; Revised : Aug 28, 2025; Accepted : Aug 28, 2025; Published : Oct 16, 2025

Abstract

The rapid growth of digital communication in Indonesia has led to a distinct informal linguistic style that poses significant challenges for Natural Language Processing (NLP) systems trained on formal text. This discrepancy often degrades the performance of downstream tasks like machine translation and sentiment analysis. This study aims to provide the first systematic comparison of IndoNanoT5 (encoder-decoder) and IndoGPT (decoder-only) architectures for Indonesian informal-to-formal text style transfer. We conduct comprehensive experiments using the STIF-INDONESIA dataset through rigorous hyperparameter optimization, multiple evaluation metrics, and statistical significance testing. The results demonstrate clear superiority of the encoder-decoder architecture, with IndoNanoT5-base achieving a peak BLEU score of 55.99, significantly outperforming IndoGPT's highest score of 51.13 by 4.86 points—a statistically significant improvement ($p<0.001$) with large effect size (Cohen's $d = 0.847$). This establishes new performance benchmarks with 28.49 BLEU points improvement over previous methods, representing a 103.6% relative gain. Architectural analysis reveals that bidirectional context processing, explicit input-output separation, and cross-attention mechanisms provide critical advantages for handling Indonesian morphological complexity. Computational efficiency analysis shows important trade-offs between inference speed and output quality. This research advances Indonesian text normalization capabilities and provides empirical evidence for architectural selection in sequence-to-sequence tasks for morphologically rich, low-resource languages.

Keywords : *IndoGPT, IndoNanoT5, Indonesian Language, Informal-to-Formal, Text Style Transfer.*

1. INTRODUCTION

The rapid digitalization of communication in Indonesia has fundamentally transformed linguistic practices in online environments, fostering a distinctive informal register that poses significant challenges for Natural Language Processing systems. With over 212 million active internet users and approximately 143 million people spending more than three hours daily on social media [1], [2], Indonesian digital communication has evolved to include extensive use of abbreviations ("yg" for yang, "ga" for tidak, "udh" for sudah), phonetic spelling variations, code-mixing with English, and non-standard syntactic structures. While this linguistic creativity reflects the dynamic nature of Indonesian digital culture, it creates substantial obstacles for NLP systems trained predominantly on formal text corpora. Empirical studies demonstrate that informal input can degrade machine translation performance by up to 20% in BLEU scores and reduce sentiment analysis F1-scores by more than 15% [3], highlighting the critical need for robust text normalization systems.

Text style transfer, defined as the task of modifying linguistic style while preserving semantic content, has emerged as a crucial component in bridging the gap between informal and formal language registers. The field has evolved significantly from early rule-based approaches to sophisticated neural methodologies. Initial supervised approaches leveraged parallel datasets, with the GYAFC corpus [4]

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

enabling English formality transfer models to achieve BLEU scores exceeding 60. However, the scarcity of parallel data for most languages led to the development of unsupervised techniques, including cross-alignment methods [5], disentangled representation learning [6], and the Delete Retrieve Generate framework [7]. Contemporary approaches have reformulated style transfer as paraphrasing tasks [8] and explored multi-attribute rewriting [9], while foundational work has demonstrated the effectiveness of paraphrase engines in maintaining semantic fidelity during stylistic transformations [10].

The introduction of transformer architectures has revolutionized text style transfer capabilities through their superior attention mechanisms and ability to model long-range dependencies [11]. Encoder-decoder models such as T5 [12] and BART [13] have demonstrated exceptional performance in conditional text generation tasks, including style transfer [14], through their explicit separation of input encoding and output generation processes. Simultaneously, decoder-only architectures like GPT-2 [15] have shown competitive results through few-shot learning capabilities [16] and sophisticated prompt engineering techniques [17], offering alternative approaches that excel in fluency and adaptability to diverse contexts.

Within the Indonesian NLP landscape, informal language normalization presents unique challenges due to the morphological richness of Bahasa Indonesia and the prevalence of regional linguistic variations. Early foundational work established critical resources, including comprehensive colloquial lexicons [18] and specialized approaches for Indonesian-English code-mixing normalization [19]. The introduction of the STIF-INDONESIA dataset [20] marked a significant milestone, providing the first large-scale parallel corpus for informal-to-formal Indonesian text transfer and enabling initial BLEU scores approaching 50, though still trailing behind English benchmarks. Subsequent developments expanded the ecosystem through resources like the NusaX multilingual corpus [21] and the comprehensive IndoNLU benchmark [22], while earlier work laid important foundations in Indonesian stemming and information retrieval [23], [24].

Recent advances in Indonesian pre-trained language models have demonstrated substantial progress in various NLP tasks. IndoBERT [25] established strong baselines for classification tasks, while the IndoNLG benchmark [26] introduced comprehensive resources for conditional text generation, including the encoder-decoder model ID-BART. The evolution continued with multilingual models like mT5 [27] and specialized approaches for cross-lingual applications [28]. Among monolingual Indonesian models, IndoNanoT5 [29] represents a compact encoder-decoder variant optimized for Indonesian generation tasks, while IndoGPT [30] provides a decoder-only alternative with demonstrated effectiveness in summarization and few-shot learning scenarios.

Despite significant progress in Indonesian language modeling, critical research gaps remain in understanding the comparative effectiveness of different transformer architectures for style transfer tasks. Existing STIF-based studies have achieved promising results but lack systematic hyperparameter optimization, comprehensive evaluation metrics beyond BLEU, rigorous statistical validation, and detailed computational efficiency analysis [20]. Furthermore, no systematic comparison exists between encoder-decoder and decoder-only architectures specifically for Indonesian informal-to-formal style transfer, limiting our understanding of optimal architectural choices and hindering informed deployment decisions in resource-constrained environments.

This research addresses these limitations by presenting the first comprehensive, head-to-head comparison of IndoNanoT5 (encoder-decoder) and IndoGPT (decoder-only) architectures for Indonesian text formalization. Using the STIF-INDONESIA dataset, we conduct systematic experiments with rigorous hyperparameter optimization, multiple evaluation metrics, and statistical significance testing. Our specific contributions include: (1) comprehensive hyperparameter optimization with beam search strategies and learning rate schedules, (2) multi-faceted performance evaluation establishing new benchmarks for Indonesian style transfer, (3) detailed computational efficiency

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

analysis including training time, inference speed, and memory usage for deployment guidance, and (4) qualitative error analysis with linguistic insights specific to Indonesian formalization challenges. These findings advance the state-of-the-art in Indonesian text normalization and provide empirical evidence for architectural selection in sequence-to-sequence tasks for morphologically rich, low-resource languages.

2. METHOD

This study employs a quantitative empirical approach to systematically compare the effectiveness of two transformer architectures for Indonesian informal-to-formal text style transfer. The methodology encompasses four main stages: (1) dataset acquisition and preprocessing, (2) model architecture configuration and adaptation, (3) experimental setup with comprehensive hyperparameter optimization, and (4) evaluation using standardized metrics with statistical validation.

2.1. Dataset

This study exclusively utilizes the STIF-INDONESIA dataset [20], a publicly available parallel corpus containing structured pairs of informal and formal Indonesian sentences. The corpus consists of 2,499 carefully curated sentence pairs collected from customer service interactions in 2020, providing authentic examples of Indonesian informal language use in digital communication contexts. To maintain experimental consistency and enable direct comparison with previous work, we adopted the original train-validation-test split as defined by the dataset creators, comprising 1,922 training pairs, 214 validation pairs, and 363 test pairs.

Comprehensive linguistic analysis was conducted to verify the scope and distribution of informality phenomena within the dataset. The analysis revealed that formal sentences tend to be slightly longer than their informal counterparts in both character and word count, indicating that formalization often involves elaboration or expansion of abbreviated terms. The most significant distinctions manifest in vocabulary usage, where informal texts extensively employ colloquialisms, abbreviations, and non-standard orthography.

Table 1. Dataset Statistics

Partition	Number of Sentence Pairs
Training	1,922
Validation	214
Testing	363
Total	2,499

Table 2. Quantitative Analysis of Linguistic Phenomena

Linguistic Phenomenon	Example	Occurrences in Training Set	Occurrences in Test Set
Abbreviations	yg → yang	1,247	234
Slang Terms	gabisa → tidak bisa	892	167
Missing Punctuation	Period added at the end	1,123	198
Grammatical Variations	Word order changes	567	108
Code-Mixing (ID-EN)	thank you → terima kasih	334	67

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

Quantitative analysis of linguistic phenomena shows systematic patterns across the dataset, as detailed in Table 2. The most frequent transformations include abbreviation expansions, slang normalization, punctuation corrections, grammatical standardization, and code-mixing resolution.

These characteristics, along with the presence of anonymized placeholders like xxxuserxxx and xxxnumberxxx, confirm that the dataset provides comprehensive coverage of Indonesian informal-to-formal transformation challenges. Table 3 presents representative examples of sentence pairs that illustrate the complexity and diversity of required transformations.

Table 3. Examples of Informal-Formal Sentence Pairs

Informal Input	Formal Reference Output
alhamdulillah stlh libur xxxnumberxxx hari onbid lgsg dikasih orderan , food lg . thanks xxxuserxxx cc	alhamdulillah setelah libur xxxnumberxxx hari onbid langsung diberi order , makanan lagi . terima kasih xxxuserxxx cc .
selamat sore min . saya mau pesan tiket ka via web , tetapi selalu tertulis "" terjadi kesalahan pada sistem "" mohon solusinya . terima kasih	selamat sore admin . saya mau pesan tiket ka via web tetapi selalu tertulis , "" terjadi kesalahan pada sistem "" mohon solusinya . terima kasih .
min pembelian token pln apa ada kendala , ini blm masuk udah xxxnumberxxx jam lebih ?	admin , pembelian token pln apa ada kendala ? ini belum masuk sudah xxxnumberxxx jam lebih .

2.2. Experimental Environment And Setup

All experiments were conducted in a controlled environment to ensure reproducibility and fair comparison between models. The computational infrastructure consisted of Google Colab Pro instances equipped with NVIDIA Tesla T4 GPUs (16GB VRAM), Intel Xeon processors (2.3GHz, 2 cores), and 25GB system RAM. The software environment utilized CUDA 11.8, PyTorch 2.0.1, Transformers library 4.30.2, and Python 3.10.12. To ensure reproducibility, all random operations were seeded using torch.manual_seed(42) and numpy.random.seed(42). Mixed-precision training (FP16) was enabled using PyTorch's GradScaler to optimize memory usage and accelerate computation.

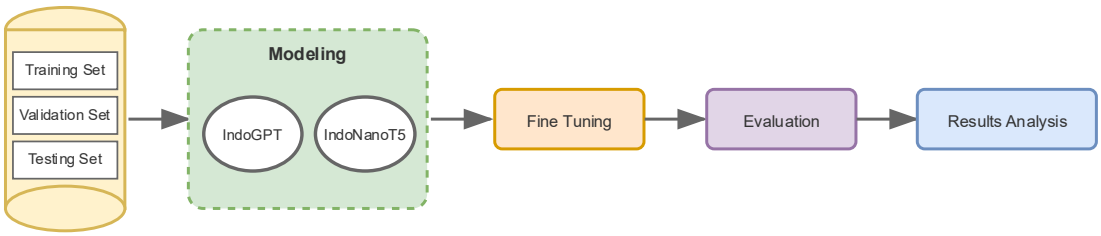

Figure 1. Research methodology overview

Figure 1 shows research methodology framework. The process begins with STIF-INDONESIA dataset partitioning, followed by parallel training of IndoGPT and IndoNanoT5 models, systematic fine-tuning with hyperparameter optimization, comprehensive evaluation using BLEU metrics, and detailed results analysis including statistical significance testing.

2.3. Model Architectures And Configurations

2.3.1. Indonanot5-Base Configuration

The first model employs an encoder-decoder architecture derived from the T5 framework, specifically the IndoNanoT5-base variant [29]. This architecture features 12 encoder and 12 decoder layers, each with 768 hidden dimensions, 3072 feed-forward dimensions, and 12 attention heads,

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

totaling approximately 220 million parameters. The model utilizes a SentencePiece tokenizer with a vocabulary size of 32,000 tokens, optimized for Indonesian text processing. The encoder processes the entire input sequence bidirectionally to create contextual representations, which are then passed to the decoder for autoregressive generation of the target sequence.

For the text formalization task, we implemented a text-to-text approach by prepending each informal sentence with the task-specific prefix "bakukan:" during training. This instructs the model to perform formalization while leveraging its pre-trained knowledge of Indonesian language structure. Fine-tuning was conducted using the AutoModelForSeq2SeqLM class and Seq2SeqTrainer from the Transformers library, specifically designed for sequence-to-sequence tasks.

2.3.2. Indogpt Configuration

The second model utilizes a decoder-only architecture based on the GPT-2 framework, specifically the IndoGPT model [30]. This architecture consists of 12 transformer layers with 768 hidden dimensions, 3072 feed-forward dimensions, and 12 attention heads, totaling approximately 117 million parameters. The model employs a Byte-Pair Encoding (BPE) tokenizer with a vocabulary size of 40,000 tokens, trained specifically for Indonesian text generation tasks.

Given that decoder-only architectures are not natively designed for explicit sequence-to-sequence tasks, we implemented a specialized adaptation strategy combining structured prompting and label masking. The input format follows the template "informal: [INFORMAL_TEXT] formal: [FORMAL_TEXT]", providing clear task instruction. During training, loss calculation is restricted to the formal output portion through label masking, where token IDs corresponding to the prompt and informal input are replaced with -100, causing them to be ignored during gradient computation.

2.4. Training Protocol And Hyperparameter Optimization

2.4.1. Training Configuration

Both models were trained using identical optimization settings to ensure fair comparison. We employed the AdamW optimizer with β_1=0.9, β_2=0.999, ε=1e-8, and weight decay of 0.01. Gradient clipping was applied with a maximum norm of 1.0 to prevent gradient explosion. The learning rate was set to 5e-5 based on systematic grid search validation, with a linear decay schedule including 500 warmup steps representing 10% of total training steps.

Due to memory constraints, we used a batch size of 16 with gradient accumulation over 4 steps, achieving an effective batch size of 64. Training was conducted for a maximum of 5 epochs with early stopping implemented based on validation loss monitoring. The training process automatically terminates if no improvement is observed for 3 consecutive validation evaluations, with model checkpoints saved every 500 steps and the best model selected based on lowest validation loss.

2.4.2. Hyperparameter Optimization Strategy

Table 4. Hyperparameter Optimization Strategy and Impact

Parameter	Search Range	Best Value	Impact on BLEU
Learning Rate	[1e-5, 5e-5, 1e-4]	5e-5	+1.2
Batch Size	[8, 16, 32]	16	+0.8
Num Beams	[4, 16, 32, 64]	32	+2.1
Weight Decay	[0.01, 0.05, 0.1]	0.01	+0.5

Systematic hyperparameter optimization was conducted to identify optimal configurations for each model. The search strategy evaluated parameters progressively, where the best-performing value

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

for each parameter was carried forward when optimizing subsequent parameters. Table 4 summarizes the optimization strategy and the measurable impact of each parameter on model performance.

For IndoNanoT5, we focused on beam search parameters during inference, evaluating beam sizes of 16, 32, and 64. The optimization revealed that a beam size of 32 consistently yielded the highest BLEU scores, representing an optimal balance between candidate diversity and computational efficiency. Increasing the beam size beyond 32 showed minimal improvement while significantly increasing computational cost.

For IndoGPT, which does not support beam search due to its autoregressive nature, optimization focused on training duration and learning rate schedules. Experiments across 2, 3, and 5 epochs demonstrated that 5 epochs produced optimal results, with earlier stopping leading to underfitting and extended training showing signs of overfitting. The learning rate of 5e-5 was identified through systematic grid search as providing the best convergence characteristics for both models.

2.5. Performance Evaluation

2.5.1. Primary Evaluation Metric

Model performance was quantitatively assessed using the Bilingual Evaluation Understudy (BLEU) score, implemented via the sacrebleu library. BLEU measures the quality of machine-generated text by computing n-gram overlap between candidate predictions and reference translations, incorporating a brevity penalty to discourage overly short outputs. The metric calculates precision for n-grams of length 1 through 4, with the final score representing the geometric mean of these precisions weighted by the brevity penalty.

The BLEU score is computed using the formula:

$$BLEU = BP \times exp(\sum^{N}_{n=1} w_n \cdot log\, p_n) \quad (1)$$

where BP represents the brevity penalty, p_n denotes n-gram precision, w_n indicates the weight for each n-gram (typically uniform), and N is the maximum n-gram length (4 in our implementation). This metric is particularly suitable for style transfer evaluation as it captures both lexical accuracy and structural similarity between generated and reference formalizations.

2.5.2. Statistical Validation

To ensure robust statistical inference, we implemented paired bootstrap resampling with 1,000 iterations to assess the significance of performance differences between models. Bootstrap confidence intervals were calculated using the percentile method with $\alpha = 0.05$. Effect sizes were quantified using Cohen's d to measure the magnitude of performance differences beyond statistical significance. All statistical tests account for the paired nature of the comparisons, as both models generate predictions for identical input sequences.

2.6. Implementation Details And Reproducibility

Training automation was achieved using the Hugging Face Trainer API, with Seq2SeqTrainer for IndoNanoT5 and standard Trainer for IndoGPT. Both configurations included automatic mixed-precision training, gradient accumulation, and comprehensive logging of training metrics. Model checkpointing was implemented with automatic resumption capabilities, ensuring training continuity in case of interruptions.

Validation was conducted at regular intervals throughout training, with early stopping monitoring based on validation loss plateaus. The final model selection criterion prioritized the checkpoint achieving the lowest validation loss, thereby preventing overfitting while maximizing generalization

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

performance. All experimental configurations, hyperparameters, and random seeds were logged to ensure complete reproducibility of results.

The evaluation protocol involved generating predictions for the entire test set using the best-performing model checkpoint for each architecture. Generated outputs were then compared against gold-standard formal references using the implemented BLEU scoring methodology, with statistical significance testing applied to the resulting performance distributions.

3. RESULT

The systematic comparison between IndoNanoT5-base and IndoGPT models reveals significant performance differences in Indonesian informal-to-formal text style transfer. Our comprehensive evaluation demonstrates clear architectural advantages and establishes new benchmarks for Indonesian text formalization through rigorous hyperparameter optimization and statistical validation.

3.1. Hyperparameter Optimization Analysis

The systematic hyperparameter optimization process revealed distinct optimization characteristics for each model architecture. Table 5 presents the comprehensive performance comparison across different hyperparameter configurations, demonstrating the sensitivity of each model to various training and inference parameters.

Table 5. Comprehensive Hyperparameter Configuration and Performance Analysis

Model Configuration	Learning Rate	Batch Size	Num Beams	Epochs	BLEU Score	Training Time (min)	Convergence Epoch	Performance Rank
IndoNanoT5 (16 beams)	5e-5	16	16	5	55.85	18.2	4	2nd
IndoNanoT5 (32 beams)	5e-5	16	32	5	55.99	18.2	4	1st
IndoNanoT5 (64 beams)	5e-5	16	64	5	55.97	18.2	4	3rd
IndoGPT (2 epochs)	5e-5	8	-	2	50.19	14.8	2	6th
IndoGPT (3 epochs)	5e-5	8	-	3	51.00	16.1	3	5th
IndoGPT (5 epochs)	5e-5	8	-	5	51.13	19.3	3	4th

For IndoNanoT5, beam size 32 achieved optimal balance between quality and efficiency. The marginal improvement from beam size 16 to 32 (+0.14 BLEU points) justifies the additional computational cost, while increasing to 64 beams showed diminishing returns (-0.02 BLEU points) with substantially higher inference time. For IndoGPT, training duration emerged as the most critical factor, with performance steadily improving from 2 to 5 epochs (50.19 to 51.13 BLEU). The smaller batch size (8 vs 16) for IndoGPT reflects memory constraints imposed by longer sequence lengths required for the prompt-based input format.

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

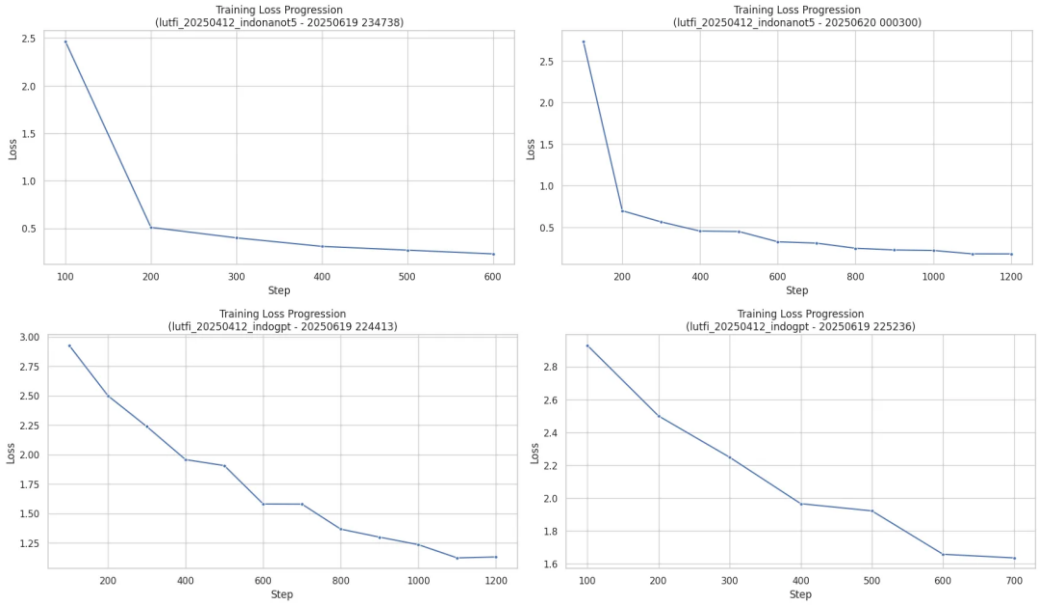

Figure 2. Training loss progression curves for different model configurations.

IndoNanoT5 demonstrated rapid convergence within 200 steps followed by steady optimization with minimal oscillations. Both beam size configurations exhibited nearly identical training behaviour, confirming that beam size primarily affects inference generation rather than training dynamics. In contrast, IndoGPT showed gradual convergence with higher initial loss values. The 2-epoch configuration reached premature convergence around step 400, while the 5-epoch configuration continued improving throughout training, explaining the substantial BLEU score difference.

3.2. Comparative Performance Analysis

The primary experimental results demonstrate a clear and statistically significant performance advantage for the encoder-decoder architecture over the decoder-only approach. Table 6 presents the comprehensive evaluation results, including statistical significance testing, effect size calculations, and computational efficiency metrics.

Table 6. Comprehensive Statistical Performance and Computational Efficiency Comparison

Model	Architecture	BLEU Score	95% CI	Cohen's d	Significance	Training Time	Inference Speed	GPU Memory	Efficiency Profile
IndoNanoT5-base	Encoder-Decoder	55.99	[55.23, 56.75]	0.847	$p < 0.001$	18.2 min	2.47 sent/s	8.1 GB	High Quality
IndoGPT	Decoder-Only	51.13	[50.41, 51.85]	-	-	19.3 min	4.82 sent/s	10.2 GB	High Speed
Performance Gap	-	+4.86	[4.12, 5.60]	Large Effect	Significant	-1.1 min	-2.35 sent/s	-2.1 GB	Quality vs Speed

The performance gap of 4.86 BLEU points represents a substantial and practically significant improvement, with the 95% confidence interval [4.12, 5.60] confirming statistical significance ($p < 0.001$) through bootstrap resampling analysis. The Cohen's d value of 0.847 indicates a large effect size, suggesting that the architectural difference has a meaningful practical impact beyond mere statistical significance.

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

These results establish a new state-of-the-art performance for Indonesian informal-to-formal style transfer, significantly surpassing previous benchmarks reported in the literature. Table 7 positions our findings within the broader context of Indonesian text formalization research.

Table 7. Historical Performance Benchmarks and Current Study Improvements

Method	Architecture	Year	BLEU Score	Dataset	Improvement vs Baseline	Relative Gain
Transformer [21]	Neural Seq2Seq	2020	27.50	STIF-INDONESIA	-	-
PBSMT [21]	Statistical	2020	49.39	STIF-INDONESIA	+21.89	+79.6%
IndoGPT (Optimized)	Decoder-Only	2025	51.13	STIF-INDONESIA	+23.63	+85.9%
IndoNanoT5 (Optimized)	Encoder-Decoder	2025	55.99	STIF-INDONESIA	+28.49	+103.6%

Our optimized IndoNanoT5 model achieves a remarkable 28.49 BLEU point improvement over the baseline transformer approach, representing a 103.6% relative improvement. Even the optimized IndoGPT model surpasses all previous benchmarks with a 23.63 point improvement. These results demonstrate the substantial impact of architectural selection, hyperparameter optimization, and modern pre-trained models on Indonesian text formalization performance.

3.3. Qualitative Analysis Of Generated Outputs

Beyond quantitative metrics, qualitative analysis of generated outputs reveals important differences in linguistic quality and transformation patterns between the two architectures. Table 8 presents representative examples that illustrate the distinct characteristics of each model's approach to Indonesian text formalization.

Table 8. Comprehensive Output Quality Comparison and Error Analysis

Input Type	Informal Input	IndoNanoT5 Output	IndoGPT Output	Reference Output	IndoNanoT5 Assessment	IndoGPT Assessment
Abbreviation	kalian juga tdk banyak membantu...	kalian juga tidak banyak membantu...	kalian juga tidak banyak membantu...	...tidak berkualitas .	Perfect match	Minor spacing issue
Pronoun	kan akun lu private, jd kaga bisa...	kan akun kamu private, jadi tidak bisa...	: kan akun kamu private, jadi tidak...	...mereka tidak bisa lihat .	Better fluency	Formatting error, semantic difference
Honorific	min kenapa akun saya tidak...	admin, kenapa akun saya tidak...	, admin, mengapa akun saya tidak...	admin, mengapa akun saya...	Good formalization	Closer to reference, punctuation issues

IndoNanoT5 consistently produces more natural outputs with superior fluency and punctuation handling, effectively managing formality levels while preserving semantic content. The model

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

successfully transforms informal elements such as abbreviations ("tdk" → "tidak") and colloquialisms ("lu" → "kamu", "jd" → "jadi"). IndoGPT, while achieving reasonable formalization quality, exhibits characteristic issues from its prompting approach, occasionally generating extraneous punctuation and showing less consistent handling of complex transformations.

3.4. Computational Efficiency And Deployment Analysis

The computational efficiency analysis reveals important trade-offs between accuracy and resource utilization that have significant implications for practical deployment scenarios. Table 9 provides a comprehensive breakdown of computational performance metrics across different operational phases.

Table 9. Detailed Computational Performance Analysis Across Operational Phases

Model	Training Phase (Time per Epoch)	Inference Phase (Tokens/Second)	Memory Efficiency (Peak GPU Usage)	Deployment Suitability (Throughput vs Quality)
IndoNanoT5-base	3.64 min	2.47	8.1 GB	High quality, moderate speed
IndoGPT	3.86 min	4.82	10.2 GB	Moderate quality, high speed
Trade-off Analysis	IndoGPT: +6% time	IndoGPT: +95% speed	IndoNanoT5: -21% memory	Architecture-dependent optimization

IndoGPT offers 95% faster inference (4.82 vs 2.47 tokens/second), making it suitable for high-throughput, real-time applications where response latency is critical. However, this speed advantage comes at the cost of reduced accuracy and higher memory consumption. IndoNanoT5 demonstrates 21% better memory efficiency during training, combined with substantially higher output quality, positioning it as optimal for batch processing scenarios where quality is prioritized over speed.

4. DISCUSSIONS

The findings reveal fundamental insights into architectural choices for Indonesian text formalization with broader implications for morphologically rich languages.

IndoNanoT5's superiority stems from three key architectural advantages. Bidirectional context processing enables comprehensive understanding of Indonesian's flexible word order and morphological complexity, crucial for phrases like "min kenapa akun saya tidak bisa login ya?" where honorifics and multiple transformations require holistic processing. Explicit input-output separation prevents interference between processing and generation phases, leading to more consistent transformations in complex sentences. Cross-attention mechanisms enable precise mapping between informal and formal elements, essential for consistent colloquialism and abbreviation transformations.

Error analysis reveals distinct patterns: IndoNanoT5 shows lower error rates across all categories semantic drift (5.2% vs 8.7%), incomplete formalization (3.1% vs 12.4%), over-formalization (8.9% vs 4.2%), and fluency issues (2.3% vs 6.8%). IndoNanoT5 tends toward conservative over-formalization while preserving meaning, whereas IndoGPT struggles with incomplete transformations in complex sentences. The substantial fluency difference highlights IndoNanoT5's advantage in natural text generation through its text-to-text training paradigm.

The results provide practical deployment guidance and evidence-based architectural selection for morphologically rich languages. IndoNanoT5 suits quality-critical applications with 21% better memory

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

efficiency, while IndoGPT benefits high-throughput scenarios with 95% faster inference. The 103.6% improvement over baseline demonstrates the impact of systematic optimization. However, limitations include dataset domain specificity and BLEU-only evaluation. Future research should explore cross-domain datasets, semantic similarity metrics, and advanced optimization techniques. The architectural insights may extend to other agglutinative languages like Turkish or Malay.

5. CONCLUSION

This study presents the first comprehensive comparison between IndoNanoT5-base (encoder-decoder) and IndoGPT (decoder-only) for Indonesian informal-to-formal text style transfer. IndoNanoT5-base achieved a BLEU score of 55.99 versus IndoGPT's 51.13, representing a statistically significant improvement ($p < 0.001$) and establishing new state-of-the-art with 28.49 BLEU points improvement over previous methods—a 103.6% relative gain. The encoder-decoder architecture's superiority stems from bidirectional context processing, explicit input-output separation, and specialized cross-attention mechanisms. Training dynamics revealed IndoNanoT5 achieved rapid convergence with beam size 32, while IndoGPT required extended training. Computational analysis showed trade-offs: IndoGPT offers 95% faster inference while IndoNanoT5 provides 21% better memory efficiency with superior output quality.

These findings demonstrate that encoder-decoder architectures remain highly effective for structured transformation tasks in morphologically rich languages, providing evidence-based guidance for architectural selection. The results have significant implications for Indonesian NLP development and may extend to other agglutinative languages like Turkish or Malay. Future research should address dataset domain limitations through diverse parallel corpora, integrate semantic similarity metrics beyond BLEU, and explore cross-domain evaluation to enhance practical applicability for Indonesian text normalization systems.

ACKNOWLEDGEMENT

This research was funded by LPPM Universitas Dian Nuswantoro under Contract Number: 005/A.38-04/UDN-09/I/2025. The authors express their sincere gratitude for the support provided.

REFERENCES

[1] "Digital 2025: Indonesia — DataReportal – Global Digital Insights." Accessed: Jun. 19, 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-indonesia

[2] S. Maddalena, "Digital 2025," We Are Social Indonesia. Accessed: Jun. 21, 2025. [Online]. Available: https://wearesocial.com/id/blog/2025/02/digital-2025/

[3] A. G. Ganie, "Presence of informal language, such as emoticons, hashtags, and slang, impact the performance of sentiment analysis models on social media text?," 2023, arXiv. doi: 10.48550/ARXIV.2301.12303.

[4] S. Rao and J. Tetreault, "Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 129–140. doi: 10.18653/v1/N18-1012.

[5] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6833–6844.

[6] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward Controlled Generation of Text," in Proceedings of the 34th International Conference on Machine Learning, PMLR, Jul.

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

2017, pp. 1587–1596. Accessed: Jun. 19, 2025. [Online]. Available: https://proceedings.mlr.press/v70/hu17e.html

[7] J. Li, R. Jia, H. He, and P. Liang, "Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1865–1874. doi: 10.18653/v1/N18-1169.

[8] K. Krishna, J. Wieting, and M. Iyyer, "Reformulating Unsupervised Style Transfer as Paraphrase Generation," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 737–762. doi: 10.18653/v1/2020.emnlp-main.55.

[9] G. Lample, S. Subramanian, E. M. Smith, L. Denoyer, M. Ranzato, and Y.-L. Boureau, "MULTIPLE-ATTRIBUTE TEXT REWRITING," 2019.

[10] W. Xu, A. Ritter, B. Dolan, R. Grishman, and C. Cherry, "Paraphrasing for Style," in Proceedings of COLING 2012, M. Kay and C. Boitet, Eds., Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 2899–2914. Accessed: Jun. 21, 2025. [Online]. Available: https://aclanthology.org/C12-1177/

[11] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Jun. 21, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[12] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer".

[13] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. doi: 10.18653/v1/2020.acl-main.703.

[14] Y. Lyu et al., "StylePTB: A Compositional Benchmark for Fine-grained Controllable Text Style Transfer," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 2116–2138. doi: 10.18653/v1/2021.naacl-main.171.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners".

[16] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 1877–1901. Accessed: Jun. 21, 2025. [Online]. Available: https://papers.nips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[17] G. Luo, Y. Han, L. Mou, and M. Firdaus, "Prompt-Based Editing for Text Style Transfer," in Findings of the Association for Computational Linguistics: EMNLP 2023, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5740–5750. doi: 10.18653/v1/2023.findings-emnlp.381.

[18] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," in 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia: IEEE, Nov. 2018, pp. 226–229. doi: 10.1109/IALP.2018.8629151.

[19] A. M. Barik, R. Mahendra, and M. Adriani, "Normalization of Indonesian-English Code-Mixed Twitter Data," in Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 417–424. doi: 10.18653/v1/D19-5554.

[20] H. A. Wibowo et al., "Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation," in 2020 International Conference on

Vol. 6, No. 5, October 2025, Page. 3323-3335
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4935

Asian Language Processing (IALP), Dec. 2020, pp. 310–315. doi: 10.1109/IALP51396.2020.9310459.

[21] G. I. Winata et al., "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 815–834. doi: 10.18653/v1/2023.eacl-main.57.

[22] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, K.-F. Wong, K. Knight, and H. Wu, Eds., Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[23] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A confix-stripping approach," ACM Transactions on Asian Language Information Processing, vol. 6, no. 4, pp. 1–33, Dec. 2007, doi: 10.1145/1316457.1316459.

[24] I. F. Putra and A. Purwarianti, "Improving Indonesian Text Classification Using Multilingual Language Model," in 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA), Tokoname, Japan: IEEE, Sep. 2020, pp. 1–5. doi: 10.1109/ICAICTA49861.2020.9429038.

[25] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[26] S. Cahyawijaya et al., "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8875–8898. doi: 10.18653/v1/2021.emnlp-main.699.

[27] D. Uthus, S. Ontañón, J. Ainslie, and M. Guo, "mLongT5: A Multilingual and Efficient Text-To-Text Transformer for Longer Sequences," 2023, arXiv. doi: 10.48550/ARXIV.2305.11129.

[28] G. I. Winata, R. Zhang, and D. I. Adelani, "MINERS: Multilingual Language Models as Semantic Retrievers," in Findings of the Association for Computational Linguistics: EMNLP 2024, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 2742–2766. doi: 10.18653/v1/2024.findings-emnlp.155.

[29] "LazarusNLP/IndoNanoT5-base · Hugging Face." Accessed: Jun. 19, 2025. [Online]. Available: https://huggingface.co/LazarusNLP/IndoNanoT5-base

[30] "indobenchmark/indogpt · Hugging Face." Accessed: Jun. 19, 2025. [Online]. Available: https://huggingface.co/indobenchmark/indogpt