



Fine-Tuning Transformer Models for Indonesian Text Summarization and Translation: A Comparative Study of PEGASUS, T5, and Komodo-7B

Muhamad Irkham Khudzifa Al Rody *, Walid

Department of Mathematics, Faculty of Mathematics and Natural Science,
Universitas Negeri Semarang, Semarang, 50229, Indonesia

Article Info

Article History:

Received 11-06-2025

Accepted 20-04-2026

Published 30-04-2026

Keywords:

Fine-tuning, Komodo-7B,

Machine Translation,

PEGASUS, Text

Summarization, Transformer,

T5

Abstract

Indonesian is a low-resource language for natural language processing, and developers do not know which Transformer model to fine-tune for summarization and translation, with which hyperparameters, and at what cost. This study compares PEGASUS (pre-trained for abstractive summarization), T5 (a text-to-text model for both tasks), and Komodo-7B (an Indonesian-centric large language model fine-tuned with LoRA). The models were fine-tuned on IndoSum (about 19,000 articles) and an Indonesian–English corpus (about 45,000 sentence pairs), split 90:10, with two batch sizes and three learning rates, and compared by ROUGE, BLEU, qualitative analysis, and resource use. PEGASUS achieved the highest ROUGE-L (0.652) in summarization (learning rate $1e-4$, batch size 4). Komodo-7B achieved the highest BLEU (49.33) in translation but needed about six times the training time of T5 (BLEU 45.77), and PEGASUS failed at translation (best BLEU 11.56). A learning rate of $1e-3$ degraded most models. PEGASUS is recommended for summarization, Komodo-7B when translation quality has priority, and T5 when resources are limited.

How to cite:

Al Rody, M. I. K., & Walid. (2025). Fine-tuning Transformer models for Indonesian text summarization and translation: A comparative study of PEGASUS, T5, and Komodo-7B. *UNNES Journal of Mathematics*. 15(1): 1-13

1. Introduction

The exponential growth of digital data in the information age has led to a pressing need for intelligent systems capable of understanding, summarizing, and translating text automatically. Every day, millions of textual documents are generated from online media, academic publications, and institutional archives. This data surge poses a significant challenge in filtering and processing relevant information efficiently (M. Zhang et al., 2022). Natural Language Processing (NLP) has emerged as a key solution, especially through two critical tasks namely text summarization and machine translation (Syed et al., 2021).

Text summarization aims to generate concise and meaningful summaries from longer texts, while machine translation facilitates cross-lingual communication by automatically translating content from one language to another (T. Zhang et al., 2024; X. Zhang et al., 2023). Both tasks have become increasingly relevant across academic, professional, and commercial domains, particularly in a globally connected world where understanding multilingual content is crucial (Dhanda & Gupta, 2024).

However, most advancements in NLP have been centered around high-resource languages such as English, Mandarin, and Spanish. Despite being one of the top 11 most spoken languages worldwide, Bahasa Indonesia remains a low-resource language with limited annotated corpora, fewer pre-trained models, and a relatively small body of research. These limitations hinder the development of high-performing NLP systems for Indonesian (Cahyawijaya et al., 2021).

The concrete problem addressed here is practical. Three kinds of Transformer models are now available for Indonesian text generation: an encoder-decoder model designed for summarization (PEGASUS), a general text-to-text model (T5), and an Indonesian-centric decoder-only large language model (Komodo-7B). It is not known which of them should be fine-tuned for summarization and which for translation, with which learning rate and batch size, and at what computational cost. The urgency is that fine-tuning budgets of Indonesian research groups are limited, while the training time of a model with 7 billion parameters is many times that of models with about 220 million parameters (Section 3), so that the choice of model needs empirical evidence and not only benchmark results from English.

Previous studies applied RNN, LSTM, and GRU models to Indonesian summarization (Adelia et al., 2019; Ivanedra & Mustikasari, 2019; Saputra, 2021), but such models struggle with long-term dependencies and coherence (Khasanah & Hayaty, 2023). Transformer models have since been fine-tuned for Indonesian: Itsnaini et al. (2023) fine-tuned T5-base on Indonesian news and obtained ROUGE-1, ROUGE-2, and ROUGE-L of 0.68, 0.61, and 0.65, although the model abstracted poorly, and Khasanah and Hayaty (2023) used GPT-2. Comparisons of Transformer models come mostly from English: Gupta et al. (2022) found T5 more accurate than PEGASUS and BART on English news. For translation, fine-tuning large language models has been studied for adaptive translation (Moslem et al., 2023) and with QLoRA (X. Zhang et al., 2023), and large language models have been benchmarked for news summarization (T. Zhang et al., 2024). Komodo-7B (Owen et al., 2024) has not been systematically evaluated on summarization and translation, and no study compares it with PEGASUS and T5 under the same fine-tuning grid and resource measurements.

The three models were chosen because they represent three design choices. PEGASUS is pre-trained by gap-sentence generation, which resembles summarization (J. Zhang et al., 2019); T5 treats every task as text-to-text and can therefore serve both tasks (Raffel et al., 2019); and Komodo-7B is pre-trained on Indonesian and regional-language text (Owen et al., 2024). Fine-tuning was used because it adapts the weights to the task and the language, and LoRA was used for the 7-billion-parameter model because full fine-tuning would exceed the available memory. ROUGE and BLEU were used because they are the standard automatic metrics for summarization and translation, and qualitative analysis and resource measurements were added because the metrics alone do not show factual or semantic faithfulness or cost.

The novelty of this study is threefold: (1) the first systematic comparison, to the authors' knowledge, of Komodo-7B with PEGASUS and T5 on both Indonesian summarization and Indonesian-English translation; (2) a common fine-tuning grid (two batch sizes and three learning rates) that shows how sensitive each model is to its hyperparameters; and (3) the joint reporting of task quality and computational cost (training time, RAM, and GPU memory), which allows model selection under resource constraints.

This study aims to compare PEGASUS, T5, and Komodo-7B after fine-tuning for Indonesian text summarization and Indonesian-English machine translation. The research questions are: (1) which model is the most accurate on each task according to ROUGE, BLEU, and qualitative analysis; (2) how do the learning rate and the batch size affect each model; (3) what are the computational costs of the models; and (4) which model and configuration should be recommended for each task and resource setting?

2. Method

2.1. Text Summarization

Text summarization is the process of condensing a longer text into a shorter version while preserving its essential meaning and core information. This technique enables users to grasp the key content without reading the entire document. In this study, we focus on abstractive summarization, a method that generates novel sentences rather than extracting existing ones. It leverages deep learning models to comprehend context and semantics, allowing the generation of concise, accurate, and human-readable summaries (Gupta et al., 2022).

2.2. Machine Translation

Machine Translation (MT) refers to the automated process of translating text or speech from one language to another using computational models. It is a critical component in cross-linguistic information access, enabling multilingual communication and document comprehension. In this research, we investigate MT in the Indonesian-English language

pair, a low-resource translation setting that poses unique challenges in syntactic structure and lexical diversity (Moslem et al., 2023).

2.3. Transformer Architecture and Models

The Transformer is a deep learning architecture introduced by Vaswani et al. (2017) through their seminal work "Attention Is All You Need". Unlike traditional Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models, Transformers process input data in parallel and utilize self-attention mechanisms to capture relationships between all words in a sequence, regardless of their distance. This enables better handling of long-term dependencies and efficient training.

The core computation in self-attention is defined as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Where Q, K and V denote the query, key, and value matrices respectively, and d_k is the dimensionality of the key vectors. Transformers apply multi-head attention, allowing the model to jointly attend to information from different representation subspaces (Vaswani et al., 2017). This study compares the following three Transformer-based models:

PEGASUS A sequence-to-sequence model designed specifically for abstractive summarization. It utilizes a pre-training strategy based on gap-sentence generation, training the model to predict masked sentences that serve as pseudo-summaries (J. Zhang et al., 2019).

T5 (Text-to-Text Transfer Transformer) A unified framework that reformulates every NLP task as a text-to-text problem. This model simplifies the architecture and training pipeline, making it applicable to a wide range of tasks including summarization and translation (Raffel et al., 2019).

Komodo-7B An Indonesian-centric large language model (LLM) with 7 billion parameters. It is pre-trained on Indonesian and regional language corpora, making it particularly well-suited for tasks involving Indonesian semantics, grammar, and morphology (Owen et al., 2024).

2.4. Experimental Setup

This research adopts an experimental methodology to evaluate the fine-tuning effectiveness of Transformer models on Indonesian-language tasks. The models, PEGASUS, T5, and Komodo-7B are fine-tuned on two Indonesian corpora: IndoSum for summarization and Indonesian-MT-Data for translation.

The two tasks are defined as sequence generation problems. In summarization, the input x is a news article and the output y is a summary of its main content. In translation, the input x is an Indonesian sentence and the output y is its English translation. For each model and task, the parameters θ were adapted by minimizing the loss in Equation (2) on the training data. A grid of two batch sizes (2 and 4) and three learning rates (1e-5, 1e-4, and 1e-3) was used for every model and task, which gives six configurations and 36 fine-tuning runs, so that the effect of each hyperparameter on each model could be compared.

Dataset Preparation

This study uses two publicly available datasets in Indonesian: IndoSum for text summarization and Indonesian-MT-Data for machine translation. IndoSum, proposed by (Kurniawan & Louvan, 2018), contains over 19,000 Indonesian news articles paired with human-written summaries. The dataset is provided in JSONLines format, with each entry including full article text and its corresponding abstractive summary.

Indonesian-MT-Data, developed by Wahyu Guntara et al. (2020), consists of approximately 45,000 English–Indonesian sentence pairs collected from various domains including informal dialogues and formal text. It is available in CSV format with two primary columns: src (source sentence) and tgt (target translation). Both datasets are split into 90% for training and 10% for testing using stratified random sampling to ensure generalization and replicability.

Data Preprocessing

Preprocessing was conducted to clean and standardize the text input. For text summarization, JSONLines were converted to tabular form, and non-alphanumeric characters, excess whitespace, and HTML tags were removed. The summary and article texts were normalized and verified to avoid duplicates or null entries.

In the translation dataset, the preprocessing involved validating sentence pairs, removing anomalies, and standardizing punctuation and spacing. The goal of this step was to ensure all input was clean, consistent, and in a format suitable for tokenization and fine-tuning.

Model Architecture

The three Transformer-based models used in this study PEGASUS, T5, and Komodo-7B are pretrained language models with different architectures. All are based on the Transformer framework introduced by Vaswani et al. (2017), which leverages self-attention for parallel sequence modeling.

Table 1 Architectural Comparison of the Evaluated Transformer Models

Component	PEGASUS	T5	Komodo-7B
Architecture Type	Encoder-decoder	Encoder-decoder	Decoder Only
Number of Layers	12	12	32
Attention Heads	12	12	32
Embedding Dimension	768	768	4096
Parameter Count	223 Million	220 Million	7 Billion
Positional Encoding	Relative	Relative	Rotary
Reference	Zhang et al.	Raffel et al.	Owen et al.

Fine-Tuning Procedure

Fine-tuning is a process of adapting a pretrained model to a specific downstream task by optimizing its weights over a labeled dataset. Formally, the objective is to minimize the expected loss L over a task-specific data distribution D_{target}

$$\theta^* = \operatorname{argmin}_{\theta} E_{(x,y) \sim D_{target}} [L(f_{\theta}(x), y)] \quad (2)$$

This expression denotes the optimization of model parameters θ to minimize the expected loss L over the target data distribution D_{target} . Here, $f_{\theta}(x)$ represents the model's prediction, and y is the true label. The goal is to adapt pre-trained parameters to the specific task domain, yielding optimal parameters θ^* that enhance model performance.

All models were fine-tuned using the HuggingFace Transformers and PyTorch frameworks. PEGASUS and T5 underwent full fine-tuning, whereas Komodo-7B was fine-tuned using LoRA (Low-Rank Adaptation) to optimize a subset of weights, minimizing resource usage.

Table 2 Hyperparameter Configurations Used in the Experiments

Model	Architecture	Configuration	Batch size	Learning rate	Epoch	Optimizer
PEGASUS	Seq2seq	Full	2, 4	1e-5, 1e-4, 1e-3	10	AdamW
T5	Seq2seq	Full	2, 4	1e-5, 1e-4, 1e-3	10	AdamW
Komodo-7B	Decoder-Only	LORA	2, 4	1e-5, 1e-4, 1e-3	10	AdamW

Evaluation metrics

To assess the effectiveness of fine-tuned models in text summarization and machine translation tasks, this study employs two widely adopted automatic evaluation metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy). These metrics quantify the similarity between the model-generated outputs and reference texts, offering a standardized framework to evaluate the fluency, adequacy, and fidelity of generated content.

For text summarization, ROUGE-1, ROUGE-2, and ROUGE-L are utilized. ROUGE-1 and ROUGE-2 compute unigram and bigram overlaps, respectively, while ROUGE-L captures the longest common subsequence (LCS), which reflects sentence-level structural similarity. Each ROUGE score is reported using Precision, Recall, and F1-score. Given a generated summary S and reference summary R , the F1-score for ROUGE-N is calculated as:

$$ROUGE - N_{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

$$ROUGE - N_{Precision} = \frac{Overlapping\ n - grams}{Total\ n - grams\ in\ S} \quad (4)$$

$$ROUGE - N_{Recall} = \frac{Overlapping\ n - grams}{Total\ n - grams\ in\ R} \quad (5)$$

BLEU evaluates the n-gram precision of the generated translation relative to one or more reference translations, incorporating a brevity penalty to discourage overly short hypotheses. It is defined as the geometric mean of modified n-gram precisions, typically for $n = 1$ to 4, and penalized by the brevity of the translation:

$$BLEU = BP \cdot e^{\sum_{n=1}^N w_n \log p_n} \quad BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c \leq r \end{cases} \quad (6)$$

where p_n denotes the modified precision for n-grams, w_n are uniform weights (usually $\frac{1}{4}$), c is the length of the candidate translation, and r is the length of the reference translation (X. Zhang et al., 2023). Both ROUGE and BLEU provide complementary insights into the quality of sequence generation models, with ROUGE emphasizing content coverage and BLEU emphasizing syntactic fidelity.

3. Results and Discussion

3.1.1. Text Summarization Results

This section presents the evaluation results of three pre-trained Transformer-based models PEGASUS, T5, and Komodo-7B in performing abstractive text summarization in Indonesian. The evaluation comprises both quantitative metrics, such as loss values and ROUGE scores, and qualitative assessments based on generated summaries.

Table 3 Performance Comparison of Models on the Text Summarization Task

Model	Batch Size	Learning Rate	Train Time (s)	Val Loss	RAM (GB)	GPU (GB)	Epochs
PEGASUS	2	1,00E-05	2626.86	0.6509	3.3	6.6	10
		1,00E-04	1331.30	0.6084	3.3	6.6	10
		1,00E-03	1052.44	1.2441	3.3	6.6	4
	4	1,00E-05	2007.57	0.6617	3.5	8.6	10
		1,00E-04	1367.71	0.6064	3.5	8.6	10
		1,00E-03	795.90	0.8730	3.5	8.6	10
T5	2	1,00E-05	1007.45	0.6589	3.1	4.9	10
		1,00E-04	773.73	0.7970	3.1	4.9	10
		1,00E-03	1328.80	1.2378	3.1	4.9	10
	4	1,00E-05	660.49	0.6461	4.2	5.8	10
		1,00E-04	779.28	0.7258	4.2	5.8	10
		1,00E-03	963.94	1.2351	4.2	5.8	10
Komodo-7B	2	1,00E-05	23100.96	1.6164	4.1	5.1	10
		1,00E-04	25696.28	1.6089	4.1	5.1	10
		1,00E-03	8450.61	1.6965	4.1	5.1	4
	4	1,00E-05	26447.95	1.6258	5.4	8.1	10
		1,00E-04	21443.35	1.6177	5.4	8.1	10
		1,00E-03	24656.26	1.6500	5.4	8.1	10

Table 3 reports the training statistics across various hyperparameter configurations. T5 and PEGASUS, both encoder-decoder models, demonstrate reasonable training time and memory usage, with PEGASUS requiring more GPU memory due to its more complex architecture. Komodo-7B, a decoder-only large language model with over 7 billion parameters, incurs significantly higher training time and memory costs, reflecting the computational demands of fine-tuning such a model.

Learning rate significantly affects model convergence. All three models experience performance degradation at a high learning rate (1e-3), where loss values increase substantially. This is particularly notable in T5, which displays unstable behavior under aggressive learning rates. Optimal performance is generally observed at a learning rate of 1e-4.

Table 4. ROUGE Score Comparison Across All Models and Hyperparameter Settings

Model	Batch Size	Learning Rate	ROUGE-1	ROUGE-2	ROUGE-L
PEGASUS	2	1,00E-05	0.684860	0.608405	0.647602
		1,00E-04	0.684880	0.606927	0.648289
		1,00E-03	0.527114	0.390539	0.484989
	4	1,00E-05	0.685494	0.610452	0.649223
		1,00E-04	0.687789	0.612850	0.652134

T5	2	1,00E-03	0.642110	0.541974	0.601330
		1,00E-05	0.597623	0.527360	0.568994
		1,00E-04	0.592580	0.526070	0.565931
		1,00E-03	0.525347	0.454117	0.499314
	4	1,00E-05	0.593520	0.522719	0.561997
		1,00E-04	0.581220	0.507788	0.548662
		1,00E-03	0.541347	0.474752	0.515321
Komodo-7B	2	1,00E-05	0.656700	0.576700	0.576700
		1,00E-04	0.665600	0.583800	0.583800
		1,00E-03	0.523900	0.462300	0.462300
		1,00E-05	0.658600	0.580100	0.580100
	4	1,00E-04	0.669600	0.587100	0.587100
		1,00E-03	0.540800	0.484100	0.484100

Table 4 presents ROUGE-1, ROUGE-2, and ROUGE-L scores for each model. PEGASUS consistently outperforms the other models across all metrics, achieving the highest scores with a configuration of batch size 4 and learning rate 1e-4: **ROUGE-1 (0.6878)**, **ROUGE-2 (0.6129)**, and **ROUGE-L (0.6521)**. These results highlight PEGASUS's architectural strengths and pretraining strategy tailored for summarization tasks.

Komodo-7B, despite being a decoder-only model not specialized for sequence-to-sequence generation, achieves respectable scores with ROUGE-1 reaching up to 0.6696. This suggests that with proper fine-tuning, even large general-purpose LLMs can perform well on summarization tasks, especially when trained on domain-relevant data.

T5 performs moderately, with the best ROUGE scores attained at learning rate 1e-5. However, the model shows sensitivity to hyperparameter changes, particularly learning rate, where increasing it leads to notable declines in summary quality.

The ROUGE-2 scores of 0.61 for PEGASUS are very high for abstractive summarization, which indicates that the generated summaries largely reuse word sequences of the article. This is consistent with the qualitative sample in Table 5, where the outputs of PEGASUS and T5 repeat the first sentences of the article, and with the characteristics of IndoSum, whose summaries are less abstractive than those of the larger Liputan6 dataset (Koto et al., 2020). It is also consistent with Itsnaini et al. (2023), who reported ROUGE-2 of 0.61 with T5 on Indonesian news but found that the model abstracted poorly. The ROUGE scores in Table 4 therefore measure largely extractive overlap, and a Lead-3 baseline should be reported to put them in perspective.

The effect of the learning rate is clear for all models. At a learning rate of 1e-3, the ROUGE-L of PEGASUS with batch size 2 falls to 0.485 from 0.648 at 1e-4, that of T5 falls to 0.499 from 0.569 at 1e-5, and the ROUGE-1 of Komodo-7B falls to 0.524 from 0.666 at 1e-4. The batch size matters mainly at the high learning rate: PEGASUS reaches ROUGE-L of 0.601 with batch size 4 against 0.485 with batch size 2, while at 1e-4 and 1e-5 the two batch sizes differ by less than 0.005. The best learning rate also depends on the model: 1e-4 for PEGASUS and Komodo-7B and 1e-5 for T5, so that no single configuration is the best for all models.

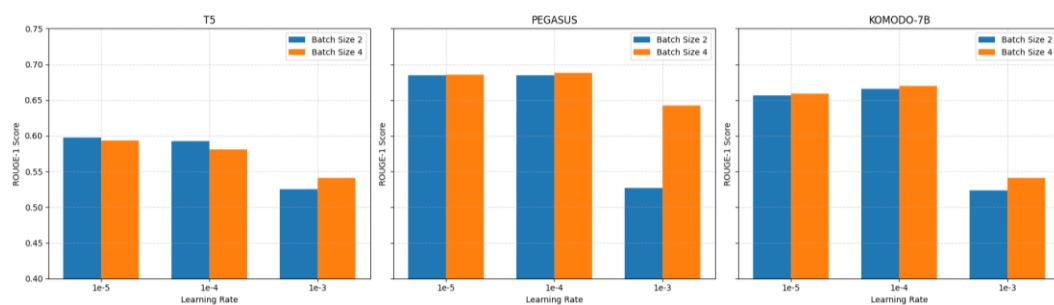


Figure 1. ROUGE-1 Score Trends for Each Model Across Different Hyperparameter Combinations

Figure 1 shows that the best setting is a learning rate of 1e-4 for PEGASUS (batch size 4) and Komodo-7B, whereas T5 is the most accurate at 1e-5. A learning rate of 1e-3 is the worst setting for all three models.

To complement the quantitative evaluation, qualitative analysis was conducted on generated summaries from two representative input samples.

Table 5. Qualitative Comparison of Generated Summaries on a Random Sample

News Text	
	Jakarta CNN Indonesia Aktor senior yang sempat memenangkan penghargaan Oscar Michael Douglas dituduh melakukan tindakan pelecehan seksual terhadap seorang wanita Dikutip dari Reuters pada Sabtu 20 1 sang wanita mengaku kalau suami dari aktris Catherine Zeta Jones itu.....Mengenai gerakan MeToo yang merangkum pengakuan korban pelecehan seksual Douglas menyatakan dukungannya dan mengatakan kalau dukungannya dilatarbelakangi oleh rasa cintanya akan Zeta Jones dan anak mereka
Reference Summary	Aktor senior yang sempat memenangkan penghargaan Oscar Michael Douglas dituduh melakukan tindakan pelecehan seksual terhadap seorang wanita Sang wanita mengaku kalau Douglas melakukan pelecehan seksual dalam bentuk verbal dan psikis selama mereka bekerja bersama pada tahun 1980 an Douglas yang saat ini telah berusia 73 tahun sejak beberapa minggu yang lalu telah mengelak tuduhan pelecehan seksual yang ditujukan kepadanya
PEGASUS Summary	Aktor senior yang sempat memenangkan penghargaan Oscar Michael Douglas dituduh melakukan tindakan pelecehan seksual terhadap seorang wanita Dikutip dari Reuters pada Sabtu 20 1 sang wanita mengaku kalau suami dari aktris Catherine Zeta Jones itu melakukan pelecehan seksual dalam bentuk verbal dan psikis selama mereka bekerja bersama pada tahun 1980 an Ia merasa dirinya penguasa dunia dan bisa melecehkan saya tanpa bisa mendapat akibatnya kata Susan Brady wartawan dan penulis dalam wawancaranya di Today Show
T5 Summary	Aktor senior yang sempat memenangkan penghargaan Oscar Michael Douglas dituduh melakukan tindakan pelecehan seksual terhadap seorang wanita Dikutip dari Reuters pada Sabtu 20 1 sang wanita mengaku kalau suami dari aktris Catherine Zeta Jones itu melakukan pelecehan seksual dalam bentuk verbal dan psikis selama mereka bekerja bersama pada tahun 1980 an Ia merasa dirinya penguasa dunia dan bisa melecehkan saya tanpa bisa mendapat akibatnya kata Susan Brady wartawan dan penulis dalam wawancaranya di Today Show NBC News Douglas yang saat ini telah berusia 73 tahun sejak beberapa minggu yang lalu telah mengelak tuduhan pelecehan seksual yang ditujukan kepadanya Tuduhan terhadap Douglas semakin meramalkan daftar pria dalam dunia hiburan yang mendapat tuduhan serupa Sebelumnya produser film Harvey Weinstein aktor Kevin Spacey
Komodo-7B Summary	Aktor senior yang sempat memenangkan penghargaan Oscar Michael Douglas dituduh melakukan tindakan pelecehan seksual terhadap seorang wanita Dikutip dari Reuters pada Sabtu 20 1 sang wanita mengaku kalau suami dari aktris Catherine Zeta Jones itu melakukan pelecehan seksual dalam bentuk verbal dan psikis selama mereka bekerja bersama pada tahun 1980 an

In the sample in Table 5, the reference summary has three sentences: the accusation, the verbal and psychological harassment in the 1980s, and Douglas's denial. The PEGASUS summary reproduces the first two sentences and adds a quotation of the accuser, but omits the denial. The T5 summary includes the denial but continues with sentences about other men accused in the entertainment industry and ends in the middle of a sentence, so it is longer and less focused than the reference. The Komodo-7B summary is the shortest and consists of the first two sentences only, so it is accurate but incomplete. Only one sample is shown; a larger sample with human judgment would be needed to generalize.

These qualitative results agree with the ROUGE scores in that PEGASUS gives readable summaries close to the reference, Komodo-7B is accurate but terse, and T5 is competitive but less focused.

Overall, PEGASUS emerges as the most effective model for abstractive summarization in Indonesian, both in terms of ROUGE metrics and qualitative coherence. The optimal configuration for this model is batch size 4 with learning rate 1e-4. The findings support the importance of selecting model architecture and fine-tuning strategy based on task specificity, especially in low-resource language settings.

3.1.2. Machine Translation Results

This section presents the performance evaluation of machine translation models in translating Indonesian text into English. Three Transformer-based models T5, PEGASUS, and Komodo-7B were fine-tuned and assessed using BLEU

scores and training loss as primary indicators. In addition to quantitative metrics, qualitative analysis was conducted to examine the semantic adequacy of generated translations.

Table 6. Performance Comparison of Models on the Machine Translation Task

Model	Batch Size	Learning Rate	Train Time (s)	Val Loss	RAM (GB)	GPU (GB)	Epochs
PEGASUS	2	1,00E-05	2532.85	0.3886	3.2	6.5	10
		1,00E-04	1764.25	0.3245	3.2	6.5	10
		1,00E-03	1532.08	0.4579	3.2	6.5	6
	4	1,00E-05	1968.67	0.4184	3.7	8.4	10
		1,00E-04	1943.98	0.3397	3.7	8.4	10
		1,00E-03	1966.65	0.4280	3.7	8.4	10
T5	2	1,00E-05	1628.89	0.1591	3.4	4.9	10
		1,00E-04	1490.53	0.2167	3.4	4.9	10
		1,00E-03	2047.72	0.5579	3.4	4.9	10
	4	1,00E-05	1233.78	0.1545	4.4	5.1	10
		1,00E-04	916.68	0.1827	4.4	5.1	10
		1,00E-03	1438.92	0.5007	4.4	5.1	10
Komodo-7B	2	1,00E-05	7991.38	0.4710	4.2	4.5	10
		1,00E-04	8495.68	0.4543	4.2	4.5	10
		1,00E-03	7702.43	0.5381	4.2	4.5	10
	4	1,00E-05	7746.75	0.4777	4.7	4.8	10
		1,00E-04	8205.60	0.4580	4.7	4.8	10
		1,00E-03	7656.89	0.5013	4.7	4.8	10

As summarized in Table 6, Komodo-7B required significantly longer training times and consumed more computational resources compared to T5 and PEGASUS, which is consistent with its large parameter size (7 billion). However, Komodo-7B showed stable training loss across various learning rates, particularly at 1e-4 and 1e-5, suggesting robustness in optimization. Conversely, PEGASUS exhibited higher and inconsistent loss values, indicating instability during fine-tuning, likely due to its architecture originally designed for summarization tasks.

Table 7. BLEU Scores for Machine Translation Across Models and Hyperparameter Variants

Model	Batch Size	Learning Rate	BLEU
PEGASUS	2	1,00E-05	7.16
		1,00E-04	11.56
		1,00E-03	3.29
	4	1,00E-05	5.30
		1,00E-04	10.04
		1,00E-03	4.57
T5	2	1,00E-05	42.48
		1,00E-04	33.33
		1,00E-03	4.04
	4	1,00E-05	45.77
		1,00E-04	36.23
		1,00E-03	7.83
Komodo-7B	2	1,00E-05	47.34
		1,00E-04	47.94
		1,00E-03	49.33
	4	1,00E-05	47.46
		1,00E-04	48.26
		1,00E-03	48.58

Table 7 reports the BLEU scores achieved by each model under different hyperparameter configurations. Komodo-7B outperformed all other models, achieving a peak BLEU score of **49.33** with batch size = 2 and learning rate = 1e-3. Its performance remained consistently high across learning rates, indicating both scalability and adaptability in handling translation tasks.

T5 ranked second, attaining a maximum BLEU score of **45.77** under batch size = 4 and learning rate = 1e-5. While slightly less accurate than Komodo-7B, T5 demonstrated competitive performance, especially considering its relatively smaller architecture and resource efficiency.

In contrast, PEGASUS consistently produced the lowest BLEU scores across all configurations, with the best result being **11.56**. This underperformance reinforces the notion that PEGASUS, optimized for abstractive summarization, is not well-suited for sequence-to-sequence translation tasks in low-resource settings like Indonesian-English translation.

The large gap between PEGASUS and the other models (best BLEU of 11.56 against 45.77 for T5 and 49.33 for Komodo-7B) is plausible because PEGASUS is pre-trained to generate masked sentences and has no translation objective, whereas T5 was pre-trained with a text-to-text objective that includes translation, and Komodo-7B is a language model pre-trained on Indonesian text. The qualitative sample supports this: PEGASUS translated “air” as “alcohol” in Table 8, which shows that it does not handle the Indonesian input reliably. This explanation depends on the pre-training data of each checkpoint, which should be stated.

The learning rate behaves differently for translation than for summarization. T5 degrades steadily with the learning rate (BLEU of 45.77, 36.23, and 7.83 at 1e-5, 1e-4, and 1e-3 with batch size 4), whereas Komodo-7B, fine-tuned with LoRA, is almost insensitive, with BLEU between 47.34 and 49.33 in all six configurations and the best value at the highest learning rate. A plausible reason is that LoRA trains few parameters, which tolerates and often needs higher learning rates; the same learning rate of 1e-3 lowers the ROUGE of Komodo-7B in summarization, so the best learning rate depends on the task as well. The 3.56-point advantage of Komodo-7B over T5 comes from single runs and should be confirmed with repeated runs.

Table 8. Qualitative Comparison of Translations Generated by Each Model

Model	Sample 1	Sample 2
Input	Saya minum air satu gelas lagi	Di atas meja ada buku dan buku tulis
Label	I will drink one more glass of water	There are a book and a notebook on the desk
PEGASUS Output	I drink one glass of alcohol every morning	There is a notebook on the desk
T5 Output	I will drink another cup of water	There are books and notebooks on the desk
Komodo-7B Output	I will drink one more glass of water	There is a book and a notebook on the desk

Overall, Komodo-7B demonstrated the best performance in both quantitative and qualitative assessments, making it the most effective model for Indonesian-English machine translation. T5 remains a viable option in resource-constrained environments, whereas PEGASUS is not recommended for translation tasks due to structural limitations and inconsistent output quality.

3.1.3. Comparative Analysis

This section presents a comparative analysis of the three Transformer-based models PEGASUS, T5, and Komodo-7B based on their best performance configurations. The evaluation considers two main dimensions: task performance (ROUGE and BLEU scores) and computational efficiency (training time, GPU, and memory usage).

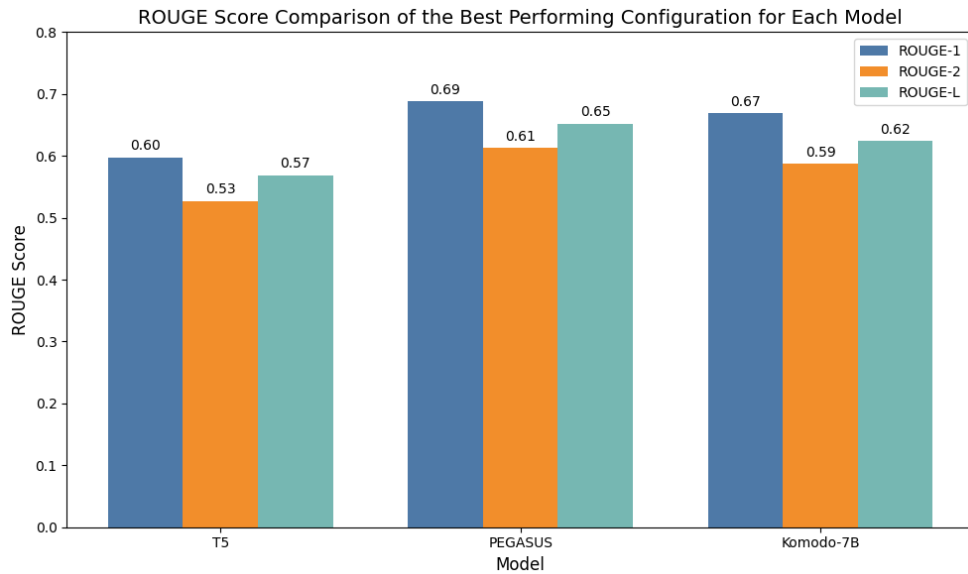


Figure 2. ROUGE Score Comparison of the Best Performing Configuration for Each Model

Figure 2 presents the comparison of ROUGE-1, ROUGE-2, and ROUGE-L scores across all models on the text summarization task. PEGASUS achieved the highest scores consistently with a learning rate of $1e-4$ and batch size of 4, outperforming T5 and Komodo-7B. The superior performance of PEGASUS can be attributed to its pretraining mechanism based on gap-sentence prediction, which enhances the model's ability to abstract key information and generate coherent summaries.

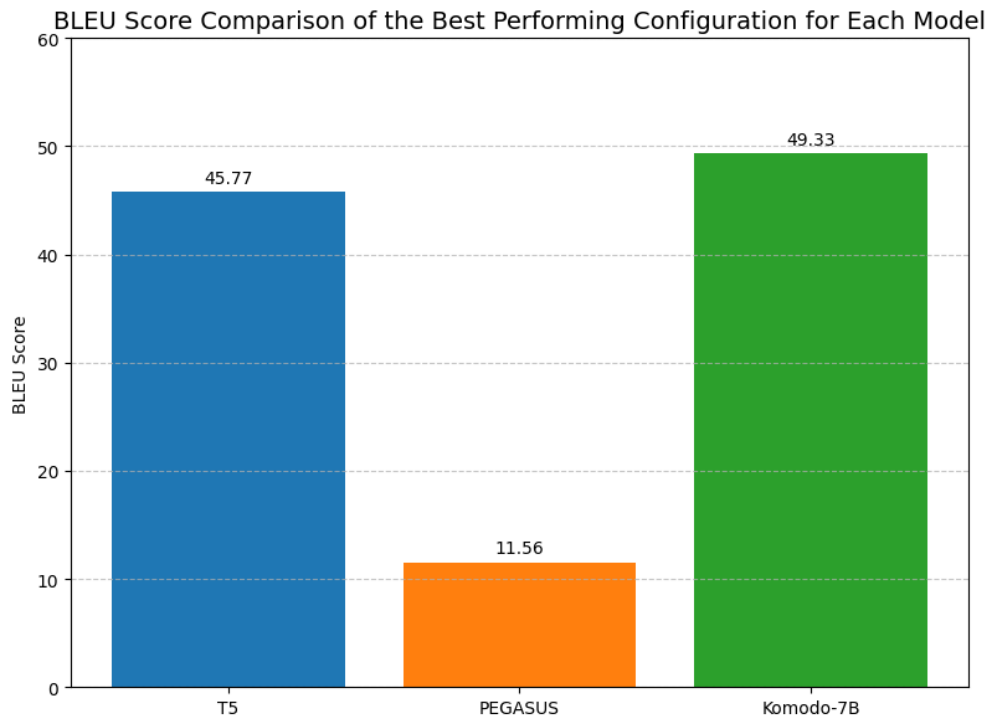


Figure 3. BLEU Score Comparison of the Best Performing Configuration for Each Model

On the other hand, Figure 3 shows BLEU score comparisons on the machine translation task. Komodo-7B clearly outperforms the other models with a BLEU score of 49.33 under the configuration of learning rate $1e-3$ and batch size 2. This superior performance reflects its architectural capacity and its training on a large-scale Indonesian corpus. T5 remains competitive with a score of 45.77, while PEGASUS performed sub-optimally, suggesting its limitation for direct translation tasks.

In terms of resource usage, Komodo-7B needs much more time: 21,443 to 26,448 seconds for summarization (ten-epoch runs) and 7,657 to 8,496 seconds for translation, against 660 to 1,329 seconds (summarization) and 917 to 2,048 seconds (translation) for T5. For the best configurations, Komodo-7B needed about 6.2 times the training time of T5 for translation (7,702 against 1,234 seconds) to gain 3.56 BLEU points, and about 15.7 times the training time of PEGASUS for summarization (21,443 against 1,368 seconds), where it was less accurate. The GPU memory of Komodo-7B with LoRA (4.5 to 8.1 GB) is similar to that of the smaller models (4.9 to 8.6 GB), so the cost of Komodo-7B is mainly time and not memory. T5 offers the best balance between quality and cost, which makes it a practical choice for many applications.

3.1.4. Key Findings

This study presents a comprehensive comparative evaluation of three Transformer-based models PEGASUS, T5, and Komodo-7B on Indonesian text summarization and machine translation tasks. Experimental results demonstrate that model performance is significantly influenced by the choice of hyperparameters, particularly learning rate and batch size. PEGASUS, fine-tuned with a learning rate of $1e-4$ and batch size of 4, achieved the highest performance on the summarization task, reaching a ROUGE-L score of **0.6521**. This suggests that PEGASUS's architecture, originally designed for abstractive summarization, is well-suited for capturing and condensing contextual information in Indonesian-language documents.

For machine translation, Komodo-7B consistently outperformed both T5 and PEGASUS across all hyperparameter settings, with its best configuration yielding a BLEU score of **49.33**. Its strong performance can be attributed to its large parameter count and pretraining on Indonesian-centric corpora, enabling better semantic and syntactic alignment with target reference translations. In contrast, PEGASUS showed substantial performance degradation on translation, underlining its limitation in generalizing to tasks beyond summarization. T5, while not achieving the top score, remained a strong contender across both tasks due to its stable performance and computational efficiency.

Beyond quantitative metrics, qualitative evaluations further support these findings. PEGASUS generated summaries that were contextually rich and coherent, while Komodo-7B delivered highly accurate translations closely aligned with reference sentences. T5 provided competitive outputs, particularly in low-resource configurations, reinforcing its viability in practical NLP applications. Overall, the results emphasize the importance of task-specific fine-tuning and informed model selection when deploying Transformer architectures for Indonesian language processing. These insights are expected to aid researchers and practitioners in optimizing NLP systems for low-resource languages.

3.1.5. Discussion

The results show that the suitability of a model is determined by how well its pre-training matches the task, more than by its size. PEGASUS, whose pre-training imitates summarization, is the best summarizer but fails at translation; Komodo-7B, the largest and most Indonesian-oriented model, is the best translator but not the best summarizer; and T5, which was designed as a general text-to-text model, is second in both tasks at a small fraction of the cost. The comparison with Gupta et al. (2022), who found T5 more accurate than PEGASUS on English news, shows that the ranking of PEGASUS and T5 is not universal: on IndoSum, PEGASUS is better. T5 in this study (ROUGE-L of 0.569) is below the T5-base of Itsnaini et al. (2023) (0.65) on Indonesian news, which may come from differences in the data, checkpoints, or hyperparameters.

The novelty of the findings is that Komodo-7B, although it is an Indonesian-centric language model, is the best translator but is outperformed in summarization by the much smaller PEGASUS, and that its advantage in translation (3.56 BLEU points over T5) is bought with six times the training time. The findings also show that the best learning rate depends on both the model and the task, so that hyperparameters tuned for one task cannot be transferred to another.

These findings lead to the following recommendations. (1) For Indonesian summarization, fine-tune PEGASUS with a learning rate of $1e-4$ and a batch size of 4. (2) For Indonesian–English translation when quality has priority and about 7,000 to 8,500 seconds of training are acceptable, fine-tune Komodo-7B with LoRA; the learning rate can be chosen from a wide range. (3) When resources are limited, use T5 with a learning rate of $1e-5$, which gives a BLEU of 45.77 in about 1,200 seconds. (4) Avoid a learning rate of $1e-3$ for full fine-tuning of PEGASUS and T5. (5) Do not use PEGASUS for translation.

Several limitations should be noted. Each configuration appears to have been run once, no significance test or repeated run is reported, and the best configurations were selected on the evaluation data. The ROUGE scores of IndoSum reflect largely extractive overlap, the qualitative analysis uses one or two samples, and Komodo-7B was fine-tuned with LoRA whereas the other models were fully fine-tuned, so that the comparison combines the effect of the model and of the tuning method. Future research should repeat the runs, add Lead-3 and human evaluation, use the more abstractive Liputan6 dataset, and compare parameter-efficient tuning for all models.

4. Conclusion

The four research questions are answered as follows. First, PEGASUS is the most accurate model for Indonesian summarization (ROUGE-L of 0.652, batch size 4, learning rate $1e-4$), and Komodo-7B is the most accurate for Indonesian–English translation (BLEU of 49.33), followed by T5 (45.77); PEGASUS performs poorly in translation (best BLEU of 11.56). Second, a learning rate of $1e-3$ degrades PEGASUS, T5, and Komodo-7B in summarization and T5 and PEGASUS in translation, whereas Komodo-7B with LoRA in translation is nearly insensitive; the best learning rate is $1e-4$ for PEGASUS and Komodo-7B and $1e-5$ for T5 in summarization, and the batch size matters mainly at the high learning rate. Third, Komodo-7B needs about 6 to 16 times the training time of T5 and PEGASUS in the best configurations, with GPU memory similar to the smaller models under LoRA, while T5 offers the best balance of quality and cost.

Fourth, PEGASUS is recommended for summarization, Komodo-7B for translation when quality has priority, and T5 when resources are limited. Because the results come from single runs on a largely extractive summarization data set, future work should repeat the runs, include Lead-3 and human evaluation, and compare parameter-efficient tuning for all models.

References

- Adelia, R., Suyanto, S., & Wisesty, U. N. (2019). Indonesian abstractive text summarization using bidirectional gated recurrent unit. *Procedia Computer Science*, 157, 581–588. <https://doi.org/10.1016/j.procs.2019.09.017>
- Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). *IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation*. arXiv. <https://doi.org/10.48550/arXiv.2104.08200>
- Dhanda, N., & Gupta, K. K. (2024). A novel approach to text summarization using machine learning. *Asian Journal of Research in Computer Science*, 17(4), 95–104. <https://doi.org/10.9734/ajrcos/2024/v17i4432>
- Fattahila, A. A. (2023). *Peringkasan artikel berita menggunakan pendekatan abstraktif dengan model Transformers*.
- Gupta, A., Chugh, D., Anjum, & Katarya, R. (2022). Automated news summarization using transformers. In *Sustainable advanced computing* (pp. 249–259). Springer. https://doi.org/10.1007/978-981-16-9012-9_21
- Itsnaini, Q. A., Hayaty, M., Putra, A. D., & Jabari, N. A. M. (2023). Abstractive text summarization using pre-trained language model “Text-to-Text Transfer Transformer (T5).” *ILKOM Jurnal Ilmiah*, 15(1), 124–131. <https://doi.org/10.33096/ilkom.v15i1.1532.124-131>
- Ivanedra, K., & Mustikasari, M. (2019). Implementasi metode recurrent neural network pada text summarization dengan teknik abstraktif. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 6(4), 377–382. <https://doi.org/10.25126/jtiik.201961067>
- Khasanah, A. N., & Hayaty, M. (2023). Abstractive-based automatic text summarization on Indonesian news using GPT-2. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, 10(1), 9–18. <https://doi.org/10.33330/jurteks.v10i1.2492>
- Koto, F., Lau, J. H., & Baldwin, T. (2020). Liputan6: A large-scale Indonesian dataset for text summarization. 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* (pp. 598–608). Association for Computational Linguistics.
- Kurniawan, K., & Louvan, S. (2018). IndoSum: A new benchmark dataset for Indonesian text summarization. In *Proceedings of the 2018 International Conference on Asian Language Processing (IALP)* (pp. 215–220). IEEE. <https://doi.org/10.1109/IALP.2018.8629109>
- Moslem, Y., Haque, R., & Way, A. (2023). *Fine-tuning large language models for adaptive machine translation*. arXiv. <https://doi.org/10.48550/arXiv.2312.12740>
- Muia, C. M., Oireew, A. M., & Ndung'u, R. N. (2024). A comparative study of transformer-based models for text summarization of news articles. *International Journal of Advanced Trends in Computer Science and Engineering*, 13(2), 37–43. <https://doi.org/10.30534/ijatcse/2024/011322024>
- Owen, L., Tripathi, V., Kumar, A., & Ahmed, B. (2024). *Komodo: A linguistic expedition into Indonesia's regional languages*. arXiv. <https://doi.org/10.48550/arXiv.2403.09362>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). *Exploring the limits of transfer learning with a unified text-to-text transformer*. arXiv. <https://doi.org/10.48550/arXiv.1910.10683>
- Saputra, M. A. (2021). *Peringkasan teks otomatis bahasa Indonesia secara abstraktif menggunakan metode long short-term memory*.
- Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9, 13248–13265. <https://doi.org/10.1109/ACCESS.2021.3052783>

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008).
- Wahyu Guntara, T., Aji, A. F., & Prasajo, R. E. (2020). Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora* (pp. 35–43). European Language Resources Association.
- Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2019). *PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization*. arXiv. <https://doi.org/10.48550/arXiv.1912.08777>
- Zhang, M., Zhou, G., Yu, W., Huang, N., & Liu, W. (2022). A comprehensive survey of abstractive text summarization based on deep learning. *Computational Intelligence and Neuroscience*, 2022, Article 7132226. <https://doi.org/10.1155/2022/7132226>
- Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12, 39–57. https://doi.org/10.1162/tacl_a_00632
- Zhang, X., Rajabi, N., Duh, K., & Koehn, P. (2023). Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 468–481). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.wmt-1.43>