Available online at http://icsejournal.com/index.php/JCSE
Journal of Computer Science an Engineering (JCSE)
e-ISSN 2721-0251
Vol. 6, No. 2, August 2025, pp. 48-58

Comparative Analysis of Parameter-Efficient-Fine-Tuning and Full Fine-Tuning Approaches for Indonesian Dialogue Summarization using mBART

Ananda Bayu Aji[1], Detty Purnamasari[2]

[1,2] Universitas Gunadarma, Jl Margonda Raya No.100, Depok 16424, Indonesia
[1]anadabayu83@gmail.com*; [2]detty@staff.gunadarma.ac.id
* corresponding author

ARTICLE INFO

Article History:
Received February 11, 2025
Revised July 12, 2025
Accepted July 16, 2025

Keywords:
Dialogue summarization,
Fine-tuning,
mBART,
PEFT,
ROUGE Score

Correspondence:
E-mail: anadabayu83@gmail.com

ABSTRACT

This study addresses the urgent need for efficient Indonesian dialogue summarization systems in remote working contexts by adapting the multilingual mBART-large-50 model. The DialogSum dataset was translated into Indonesian using Opus-MT, and two fine-tuning approaches, full fine-tuning and Parameter-Efficient Fine-Tuning (PEFT) with LoRA, were evaluated. Experiments on 1,500 test samples revealed that full fine-tuning achieved superior performance (ROUGE-1: 0.3726), while PEFT reduced energy consumption by 68.7% with a moderate accuracy trade-off (ROUGE-1: 0.2899). A Gradio-based interface demonstrated practical utility, enabling direct comparison of baseline, fine-tuned, and PEFT models. Critical findings include translation-induced terminology inconsistencies (e.g., "Hebes" vs. "Hebei") and context retention challenges in long dialogues. This work contributes a scalable framework for low-resource language NLP and provides actionable insights for optimizing computational efficiency in real-world applications.

1. Introduction

Conversation plays a crucial role in interpersonal interactions, whether in daily life, workplaces, or online forums, drawing significant attention from both academia and industry [1]. With the rapid advancements in speech recognition systems and the rising trend of remote work, there has been an increasing number of lengthy conversations being recorded and transcribed, such as meeting minutes, interviews, and debates. These long conversations serve as a dense medium of information, posing challenges for users to quickly grasp the key discussions and extract relevant insights within a short timeframe.

This phenomenon aligns with the growing popularity of online meeting platforms such as Zoom, Google Workspace, Microsoft Teams, and Skype, which have become integral to modern-day activities [2]. According to a Populix survey titled *Unveiling the Tech Revolution: How Technology Reshapes the Future of Work* conducted in April 2023, the majority of Indonesians use Zoom (77%), followed by Google Workspace (54%), Microsoft Teams (30%), and Skype (24%) [3]. These platforms play a significant role in workplace productivity and remote education by offering innovative features, such as the ability to record and download chat histories for documentation purposes.

However, downloaded chat logs often contain unstructured text with excessive and irrelevant information from the conversation. This condition makes it difficult for users to efficiently extract essential insights. This challenge highlights the urgent need for a solution capable of summarizing conversations quickly and accurately. In this context, natural language processing (NLP) technology plays a crucial role, particularly in the task of conversation summarization, to optimize the understanding and utilization of conversational data.

Conversation summarization refers to the process of extracting essential information from lengthy conversational texts to generate concise and accurate summaries [4]. This process requires technologies that can capture the multi-layered complexity of conversations, comprehend diverse linguistic styles, and adapt to various use-case scenarios [5]. Additionally, a major challenge in this field is the limited availability of labeled conversational datasets in the Indonesian language, as well as the need for efficient approaches to processing long-sequence data.

Recent breakthroughs in NLP have been significantly driven by pre-trained Transformer-based models, which have demonstrated remarkable performance improvements across multiple NLP tasks. These models have proven particularly effective in training scenarios with limited data availability [6][7]. One such model that has yielded excellent results in various NLP tasks is the multilingual version of the Bidirectional and Auto-Regressive Transformers (BART) model, known as mBART. Built on Transformer architecture, mBART is designed to support text processing in multiple languages, including Indonesian, and enables both bidirectional and autoregressive generation [8][9].

By fine-tuning mBART on specific datasets, its capabilities can be optimized for tasks such as conversation summarization. Several studies have explored the application of mBART in different languages, including Russian [10][11], Vietnamese [12][13][14], and many others. The evaluation results from these studies indicate promising outcomes. For instance, a study on Vietnamese [12][14][15]) achieved ROUGE-1, ROUGE-2, and ROUGE-L scores of 55.21, 25.69, and 37.33, respectively, using the WikiLingua dataset, and 59.81, 28.28, and 38.71, respectively, using the Vietnews dataset.

One of the primary challenges in conversation summarization for the Indonesian language is the lack of high-quality, labeled datasets. The DialogSum dataset was developed to support research in conversation summarization by providing dialogue data enriched with real-world conversational scenarios. It integrates data from three public dialogue corpora Dailydialog [16], DREAM [17], and MuTual [18] as well as from English language learning websites. Dialogues in this dataset cover various daily life topics such as education, work, health, shopping, entertainment, and travel. Most of these conversations occur between friends, colleagues, or between service providers and customers. One of the key advantages of this dataset is its structured communication patterns and clear intents, making it an ideal resource for conversation summarization tasks [19].

However, a major limitation of DialogSum is that it is available only in English. Consequently, a preprocessing step is necessary to translate this dataset into Indonesian to ensure relevance to this study. To facilitate this process, Opus-MT, a Transformer-based machine translation model developed under the Open Parallel Corpus (OPUS) project, will be utilized [20]. Opus-MT is known for its ability to generate high-quality translations by leveraging multilingual parallel corpora, making it an ideal tool for adapting DialogSum into Indonesian [21].

Given the background and challenges outlined above, this research aims to develop an automatic conversation summarization system in the Indonesian language using the mBART model. By adapting the DialogSum dataset through Opus-MT for translation into Indonesian, this study will explore and compare two different fine-tuning approaches: full fine-tuning and parameter-efficient fine-tuning (PEFT). These methods will be comprehensively evaluated using the ROUGE metric to assess the quality of generated summaries.

2. Method

This study adopts a systematic framework to develop Indonesian dialogue summarization systems through multilingual model adaptation, addressing both data scarcity and computational efficiency

challenges. The methodology comprises three main stages: (1) dataset translation and preprocessing to localize the English DialogSum corpus into Indonesian, (2) model adaptation via full fine-tuning and parameter-efficient fine-tuning (PEFT) of the mBART-large-50 model, and (3) performance evaluation using both quantitative metrics (ROUGE scores) and qualitative analysis (visual comparison of model outputs), alongside energy consumption monitoring.

2.1 Data Preparation

The foundation of this study lies in adapting the English-language DialogSum dataset to Indonesian, addressing the critical gap in labeled conversational data for Bahasa Indonesia. The original DialogSum corpus comprises 14,460 dialogues spanning real-world scenarios such as healthcare consultations, workplace discussions, and customer service interactions, annotated with human-written summaries and topic labels. To preserve the richness of conversational contexts while localizing the data, a multi-stage translation and preprocessing pipeline was implemented using computational resources from an NVIDIA DGX Station A-100, ensuring efficient handling of large-scale text processing.

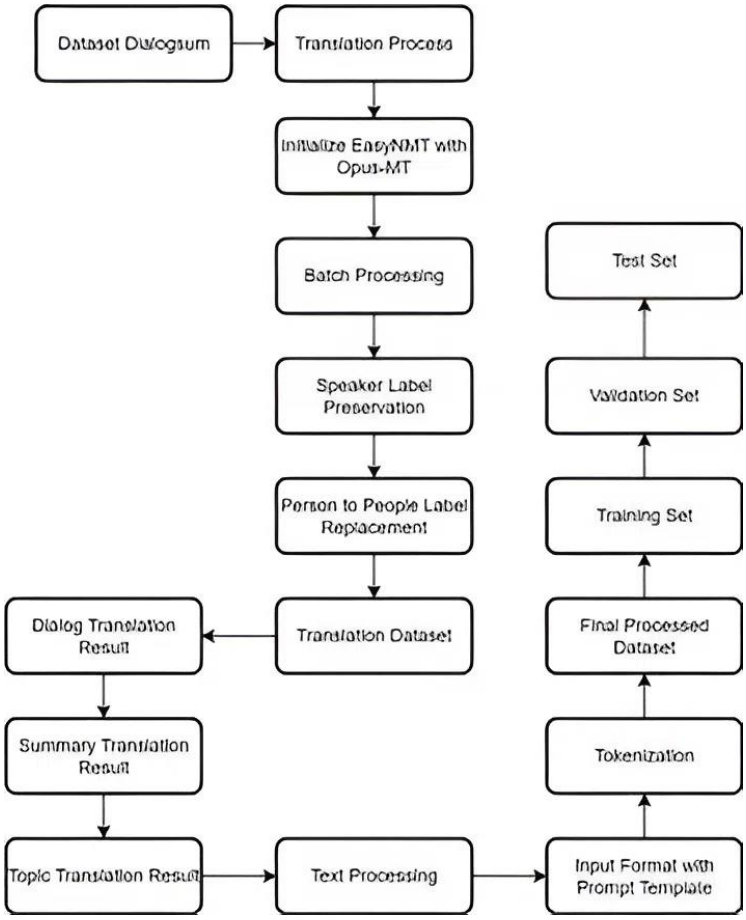

Figure 1. Data Preparation

The translation process began with the initialization of the Opus-MT model via the EasyNMT library, selected for its robust performance on low-resource language pairs. As illustrated in **Figure 1**, the pipeline follows a systematic sequence, starting with batch processing of 50 dialogues to balance memory utilization and processing speed. A key challenge in this phase involved preserving speaker identifiers (e.g., #Person1#) while maintaining translation accuracy. Regular expressions were employed to isolate speaker labels from dialogue content, thus preserving structural integrity. For example, the utterance #Person1#: Let's reschedule the meeting was translated into #Orang1#: Mari jadwal ulang rapat, retaining the speaker tag while adapting the semantic content to fluent Indonesian.

Subsequently, speaker labels were standardized into the form #OrangX# to align with Indonesian morphological conventions. Additionally, the system incorporated a label replacement mechanism to handle common mistranslations and enhance naturalness. This included, for instance, correcting literal translations such as “stomachache” to the idiomatic “sakit perut,” avoiding outputs like “sakit perut-ache” that emerged from incorrect token alignment. Following translation, the resulting dataset underwent further preprocessing, including text normalization, splitting into training, validation, and test subsets, tokenization, and formatting into input prompt templates. The overall workflow, as depicted in **Figure 1**, ensured the generation of a linguistically consistent and model-compatible Indonesian version of the DialogSum corpus.

Table 1. Translation Result

id	dialogue	summary	topic	dialogue_translated	summary_translated	topic_translated
0	#Person1#: Hi, Mr. Smith. I’m Doctor Hawkins. ...	Mr. Smith’s getting a check-up, and Doctor Haw..	get a checkup	#Orang1#: Hai, Tn. Smith. Aku Dokter Hawkins. ...	Mr Smith mendapatkan check-up, dan Dokter Hawk...	mendapatkan checkup
1	#Person1#: Hello Mrs. Parker, how have you bee...	Mrs Parker takes Ricky for his vaccines. Dr. P...	vaccines	#Orang1#: Halo Mrs Parker, bagaimana kabarmu?	Mrs Parker mengambil Ricky untuk vaksinnya. Dr...	vaksin

The construction of the translated corpus, named **IndoSum**, serves as the primary outcome of the preprocessing pipeline. This dataset comprises 12,460 instances allocated for training, 500 for validation, and 1,500 for testing. Each instance retains the essential structure of the original DialogSum dataset, consisting of multi-turn dialogues, abstractive summaries, and associated topic labels. All components were localized into Bahasa Indonesia through a controlled translation process that preserved semantic and structural consistency. **Table 1** presents representative examples of this alignment, demonstrating the parallel structure between English and Indonesian texts. Notably, speaker labels were preserved and adapted appropriately (e.g., #Person1# to #Orang1#), while conversational utterances and summaries were translated in a context-sensitive manner to maintain naturalness and fluency in the target language.

Each translated dialogue contains an average of 94 tokens, whereas summaries are condensed to approximately 20 tokens. This level of abstraction is indicative of practical summarization requirements in real-world conversational systems. The distribution of thematic categories reflects that of the original English dataset, with dominant domains including Healthcare (23%), Workplace Coordination (19%), and Travel Planning (15%). These three categories collectively account for 57% of the corpus, indicating a high degree of alignment in content coverage and ensuring relevance to downstream applications in dialog-based summarization.

The tokenization process was conducted using a predefined tokenizer with a maximum sequence length of 128 tokens, applying padding to max length and truncation to handle longer sequences efficiently. Each dialogue input was converted into token IDs, with padding added to standardize input lengths and an attention mask generated to differentiate actual content from padding. For the target summaries (labels), padding tokens were masked with -100 to prevent them from contributing

to the loss computation during training. The final tokenized dataset consists of 12,460 training samples, 500 validation samples, and 1,500 test samples, each containing input_ids, attention_mask, and labels. This preprocessing step ensures compatibility with Transformer architectures while optimizing sequence length for computational efficiency. A batch size of 100 was selected to balance GPU memory utilization and prevent out-of-memory (OOM) errors, ensuring the dataset is well-structured for fine-tuning the pre-trained language model (PLM) on the text summarization task.

2.2 Modelling

This study addresses two fundamental challenges in Indonesian conversational summarization: the scarcity of high-quality labeled datasets and the need for computationally efficient model training. To overcome these limitations, the multilingual mBART-large-50 model was selected as the foundational architecture due to its demonstrated capacity for handling cross-lingual tasks and its suitability for morphologically rich languages such as Bahasa Indonesia..

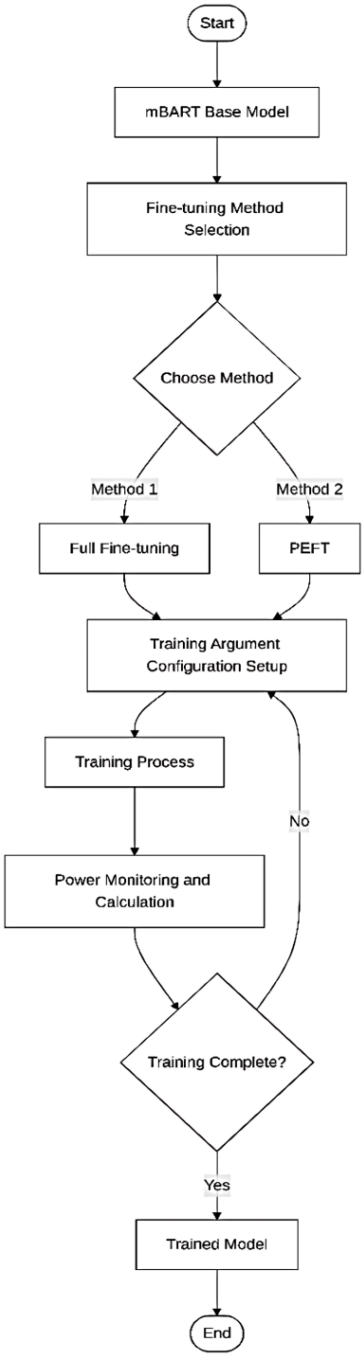

Figure 2. Modelling Process

Figure 2 illustrates the model adaptation workflow designed to explore the trade-off between training accuracy and resource efficiency. The process begins with the initialization of the mBART base model, followed by the selection of a fine-tuning strategy. Two distinct approaches were implemented: (1) full fine-tuning, which updates all model parameters for optimal task-specific performance, and (2) Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA), which modifies only a subset of parameters to reduce computational and memory overhead.

After selecting the fine-tuning method, training arguments—such as learning rate, batch size, and sequence length—are configured before executing the training process using the translated IndoSum dataset. Power consumption was continuously monitored using hardware-level sensors on the NVIDIA DGX Station A100 to evaluate energy efficiency. Upon training completion, the resulting model was stored for downstream evaluation and inference.

The pipeline in **Figure 2** supports a structured and reproducible comparison between full fine-tuning and PEFT under consistent experimental conditions. Further architectural and configuration details for the mBART implementation are summarized in **Table 2**.

Table 2. Model Architecture

Architecture Parameters	Specifications
Model Type	mBART-large-50
Encoder-Decoder Layers	12 layers
Attention Heads	16
Total Parameters	610,879,488
Trainable Parameters (Full Fine-tuning)	610,879,488 (100%)
Trainable Parameters (PEFT-LoRA)	4,718,592 (0.77%)

In the full fine-tuning approach, all 610,879,488 parameters of the model were optimized to learn Indonesian conversation-specific patterns. Conversely, the PEFT implementation with LoRA modified only 0.77% of the parameters (4.7 million) by inserting LoRA matrices into the attention mechanism's query (q_proj) and value (v_proj) layers, significantly reducing computational overhead while maintaining adaptive capabilities. The training configuration parameters are outlined in **Table 3**.

Table 3. Training Configuration Parameters

Training Parameters	**Value**
Learning Rate	1e-5
Batch Size (Training)	8
Batch Size (Evaluation)	1
Number of Epochs	24
Optimizer	AdamW (fused)
Learning Rate Scheduler	Cosine with Restarts
Precision	Mixed Precision (FP16)
Gradient Accumulation Steps	16

The training process employed an AdamW optimizer with fused implementation and a learning rate of 1e-5 with cosine annealing with restarts to prevent stagnation. Mixed Precision Training (FP16) was implemented to reduce memory consumption by up to 40% without sacrificing precision. All implementations were conducted on an NVIDIA DGX Station A100 infrastructure with real-time power consumption monitoring via nvidia-smi utilities. The mBART-large-50 architecture was chosen for its bidirectional text generation capabilities, which proved crucial for comprehending complex conversational contexts in the Indonesian language domain.

3. Results and Discussion

The results of the training analysis, obtained under identical configuration settings for both fine-tuning approaches, are depicted in **Figure 3**:

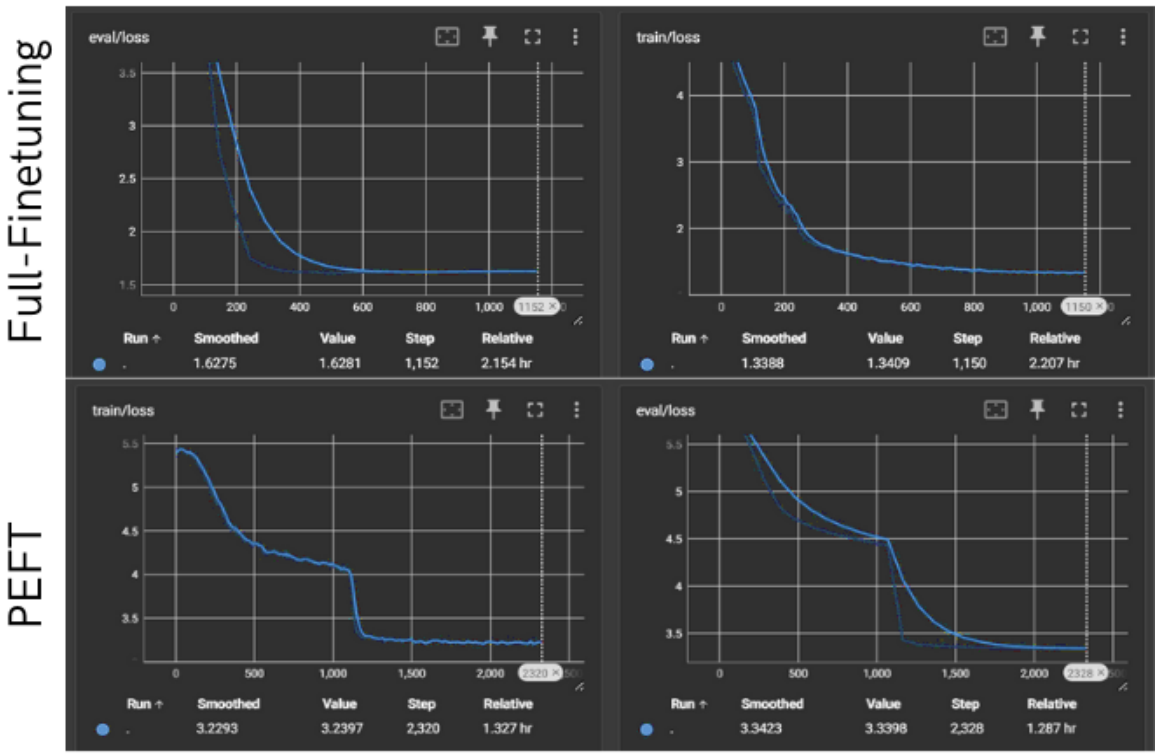

Figure 3. Training Loss

Full Fine-Tuning exhibited rapid convergence in its training process, where all model parameters (610 million, 100%) were adjusted. The loss curves demonstrate:

- Training Loss (top right): Shows quick initial descent, stabilizing after approximately 1,000 steps and reaching 1.34 after 1,150 steps
- Evaluation Loss (top left): Displays sharp initial decline before gradually stabilizing at 1.62 after 1,152 steps

In contrast, the PEFT approach with LoRA, which modified only 0.77% of parameters (4.7 million), showed different training dynamics:

- Training Loss (bottom left): Demonstrates slower descent compared to Full Fine-Tuning, with significant changes after 1,500 steps, eventually stabilizing around 3.23 after 2,320 steps
- Evaluation Loss (bottom right): Started high at approximately 5.0, gradually decreasing to stabilize around 3.34 after 2,328 steps

While PEFT demonstrated higher final loss values compared to Full Fine-Tuning, it achieved significant computational efficiency, reducing training time by approximately 41%. These results suggest potential for further optimization through hyperparameter tuning, particularly in learning rate, LoRA rank, or training steps.

The qualitative evaluation shown in **Table 4** below reveals that the baseline model fails to generate meaningful summaries, as it merely replicates the instruction format without extracting key information. The full fine-tuned model effectively captures the main topic of the conversation—

conflict between travel plans and the sandstorm threat—while maintaining coherence, but it omits specific details such as wind intensity and health impacts. In contrast, the PEFT model retains numerical details like "six-degree strong winds" but suffers from translation artifacts, leading to terminology inconsistencies (e.g., "Hebes" instead of "Hebei"). These findings highlight a trade-off between completeness and precision, where full fine-tuning prioritizes coherence through generalization, while PEFT preserves granular details at the cost of potential translation errors.

Table 4. Result Data

ORIGINAL DIALOGUE:
#Orang1#: Di mana Anda akan untuk perjalanan Anda?
#Orang2#: Saya pikir Hebei adalah tempat yang baik.
#Orang1#: Tapi kudengar utara Cina mengalami badai pasir yang parah!
#Orang2#: Benarkah?
#Orang1#: Ya, dikatakan bahwa Hebes mengalami enam derajat angin kencang.
#Orang2#: Bagaimana badai ini mempengaruhi orang - orang yang tinggal di daerah ini? #Orang1#: Laporan itu mengatakan jumlah penderita infeksi saluran pernapasan cenderung meningkat setelah badai pasir. Pasir masuk ke hidung dan tenggorokan orang dan menciptakan iritasi.
#Orang2#: Kedengarannya badai pasir adalah masalah bagi semua orang!
#Orang1#: Kau benar.

BASELINE HUMAN SUMMARY:
#Orang2# #Berrencana untuk melakukan perjalanan di Hebei tapi
#Orang1# #berkata ada badai pasir di sana.

BASELINE MODEL SUMMARY:
[TASK] Summarize the following conversation. [CONTEXT] #Orang1#: Di mana Anda akan untuk perjalanan Anda? #Orang2#: Saya pikir Hebei adalah tempat yang baik. #Orang1#: Tapi kudengar utara Cina mengalami badai pasir yang parah! #Orang2#: Benarkah? #Orang1#: Ya, dikatakan bahwa Hebes mengalami enam derajat angin kencang. #Orang2#: Bagaimana badai ini mempengaruhi orang - orang yang tinggal di daerah ini? #Orang1#: Laporan itu mengatakan jumlah penderita infeksi saluran pernapasan cenderung meningkat setelah badai pasir. Pasir masuk ke hidung dan tenggorokan orang dan menciptakan iritasi. #Orang2#: Kedengarannya badai pasir adalah masalah bagi semua orang! #Orang1#: Kau benar. [OUTPUT]

FINE-TUNED MODEL SUMMARY:
#Orang1# #Orang2# #Tentang badai pasir yang parah di utara Cina. #Orang2# #berpikir itu adalah masalah bagi semua orang.

PEFT MODEL SUMMARY:
#Orang1# dan #Orang2# berbicara tentang utara Cina mengalami badai pasir yang parah. #Orang1# mengatakan bahwa Hebes mengalami enam derajat angin kencang.

Building upon the training dynamics analysis, the quantitative evaluation in **Table 5** of both approaches provides deeper insights into their respective performance and computational trade-offs. By systematically assessing summarization quality and resource efficiency, this evaluation highlights the strengths and limitations of Full Fine-Tuning and PEFT. The results not only quantify their effectiveness in generating high-quality summaries but also reveal crucial considerations regarding training time and energy consumption.

Table 5. ROUGE Score (test set, n = 1,500)

Model	ROUGE-1 (±σ)	ROUGE-2 (±σ)	ROUGE-L (±σ)
Baseline	0.1632 (±0.0628)	0.0417 (±0.0416)	0.1257 (±0.0506)

Full Fine-Tuning	0.3726 (±0.1413)	0.1286 (±0.1274)	0.3060 (±0.1320)
PEFT	0.2899 (±0.1590)	0.0909 (±0.1071)	0.2356 (±0.1357)

The ROUGE score evaluation on 1,500 test samples highlights the relative effectiveness of each approach. Full Fine-Tuning consistently outperformed across all metrics, surpassing PEFT by 22% in ROUGE-1, largely due to its comprehensive parameter updates. Meanwhile, PEFT, though yielding slightly lower scores, still demonstrated respectable performance. However, its higher standard deviation (±0.159) indicates greater variability, suggesting potential instability in handling domain-specific terms.

Table 6. Power Monitoring

Metric	Full Fine-Tuning	PEFT	Reduction
Time (minutes)	263.67	159.67	39.4%
Energy (kWh)	0.9591	0.3001	68.7%

The computational efficiency analysis in **Table 6** reveals PEFT's substantial advantages in resource utilization. Compared to Full Fine-Tuning, PEFT significantly reduces training time and energy consumption, making it a more efficient alternative. However, this efficiency comes at the cost of performance, as reflected in a 22% decrease in ROUGE-1 scores. Despite this trade-off, PEFT remains a viable choice for scenarios where computational efficiency is prioritized over peak accuracy. Its 68.7% reduction in energy consumption makes it particularly suitable for resource-constrained environments, offering a balanced approach between performance and efficiency.

4. Conclusion

This study successfully addresses the critical need for efficient Indonesian dialogue summarization systems by adapting the multilingual mBART model through two distinct fine-tuning approaches. Leveraging the translated IndoSum dataset, our experiments demonstrate that full fine-tuning achieves superior performance, yielding ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.3726, 0.1286, and 0.3060, respectively. However, the Parameter-Efficient Fine-Tuning (PEFT) approach with LoRA emerges as a compelling alternative, reducing energy consumption by 68.7% and training time by 39.4%, albeit with moderate trade-offs in accuracy (ROUGE-1: 0.2899). Critical challenges identified include translation-induced terminology inconsistencies (e.g., "Hebes" vs. "Hebei") and diminished context retention in lengthy dialogues, which warrant targeted improvements in future work. Despite these promising results, several limitations should be acknowledged. The reliance on machine translation introduces potential semantic drift and cultural nuances that may not be fully captured in Indonesian contexts. The fixed sequence length of 128 tokens may truncate important contextual information, while the evaluation relies primarily on ROUGE metrics without human assessment. Additionally, the PEFT approach showed higher variance in performance, indicating potential instability in handling domain-specific terminology. Future research should explore hybrid fine-tuning strategies that combine PEFT efficiency with full fine-tuning performance, develop Indonesian-specific pre-trained models, and implement comprehensive evaluation frameworks including human assessment. Priority should be given to creating native Indonesian dialogue datasets, investigating adaptive sequence length mechanisms, and developing model compression techniques for resource-constrained deployment. Cross-domain extensions and multi-modal integration represent additional promising directions. This research contributes a scalable framework for adapting multilingual models to low-resource languages and provides actionable insights for balancing computational efficiency with performance in NLP applications. These advancements will

further empower organizations to harness conversational data effectively in resource-constrained environments, aligning with the growing demands of remote work ecosystems.

References

[1] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). Dialogptâ¯: Large-scale generative pretraining for conversational response generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics.

[2] Angelia, D. (2022). Platform Online Meeting Terpopuler Selama Pandemi Covid-19. [Accessed: 07 January 2025].

[3] Lavinda (2023). Survei populix: Mayoritas orang indonesia gunakan zoom dan google. [Accessed: 07 January 2025].

[4] Lundberg, C., Sánchez Viñuela, L., and Biales, S. (2022). Dialogue summarization using BART. In Shaikh, S., Ferreira, T., and Stent, A., editors, Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges, pages 121–125, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

[5] Chen, Y., Liu, Y., and Zhang, Y. (2021b). DialogSum challenge: Summarizing real-life scenario dialogues. In Belz, A., Fan, A., Reiter, E., and Sripada, Y., editors, Proceedings of the 14th International Conference on Natural Language Generation, pages 308–313, Aberdeen, Scotland, UK. Association for Computational Linguistics.

[6] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. Advances in neural information processing systems, 33:17283–17297.

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

[8] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). 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. Association for Computational Linguistics.

[9] Uthus, D., Ontañón, S., Ainslie, J., and Guo, M. (2023). mlongt5: A multilingual and efficient text-to-text transformer for longer sequences. arXiv preprint arXiv:2305.11129

[10] Gusev, I. (2020). Dataset for Automatic Summarization of Russian News, pages 122–134. Springer International Publishing.

[11] Taunk, D. and Varma, V. (2023). Summarizing indian languages using multilingual transformers based models. arXiv preprint arXiv:2303.16657.

[12] Nguyen, H., Phan, L., Anibal, J., Peltekian, A., and Tran, H. (2021). Viesum: how robust are transformer-based models on vietnamese summarization? arXiv preprint arXiv:2110.04257.

[13] Nguyen, T.-H. and Do, T.-N. (2022). Text summarization on large-scale vietnamese datasets. Journal of information and communication convergence engineering, 20(4):309–316.

[14] Zheng, K. and Zheng, W. (2022). Deep neural networks algorithm for vietnamese word segmentation. Scientific Programming, 2022:1–11

[15] Wang, Y., Du, J., Kuang, J., Chen, C., Li, M., and Wang, J. (2023). Twoscaled identification of landscape character types and areas: A case study of the yunnanâvietnam railway (yunnan section), china. Sustainability, 15(7):6173.

[16] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957

[17] Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., and Cardie, C. (2019). Dream: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics, 7:217–231.

[18] Cui, L., Wu, Y., Liu, S., Zhang, Y., and Zhou, M. (2020). Mutual: A dataset for multi-turn dialogue reasoning. arXiv preprint arXiv:2004.04494.

[19] Chen, Y., Liu, Y., Chen, L., and Zhang, Y. (2021a). Dialogsum: A real-life scenario dialogue summarization dataset. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics.

[20] Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L., and Forcada, M. L., editors, Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

[21] Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.- A., Nieminen, T., Raganato, A., Scherrer, Y., Vázquez, R., and Virpioja, S. (2023). Democratizing neural machine translation with opus-mt. Language Resources and Evaluation, 58(2):713–755.