Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2901~2907
e-ISSN: 2548-6861

Neural Machine Translation of Balinese-Indonesian Using T5 Architecture with QLoRA Optimization

Leonard Bodhi Kumaro [1]*, I Gusti Ngurah Lanang Wijayakusuma [2]*, IPW Gautama [3]*
* Matematika, Universitas Udayana
leonardbodhikumaro@gmail.com [1], lanang_wijaya@unud.ac.id [2], winadagautama@unud.ac.id [3]

Article Info

Article history:

Received 2026-04-16
Revised 2026-05-04
Accepted 2026-05-25

Keyword:

Neural Machine Translation,
T5,
QloRA,
Balinese Language,
Low-Resource Language.

ABSTRACT

This study proposes a Neural Machine Translation (NMT) system for Balinese–Indonesian translation by integrating the T5 architecture with Quantized Low-Rank Adaptation (QLoRA) to address low-resource constraints. The model is trained using the NusaTranslation dataset, consisting of 140,972 parallel sentence pairs, and optimized through parameter-efficient fine-tuning with 4-bit quantization and low-rank adaptation. Unlike conventional full fine-tuning, the proposed approach updates only a small fraction of parameters, significantly improving computational efficiency. Experimental results show that the proposed model achieves a BLEU score of 27.93%, ROUGE-1 of 18.94%, ROUGE-2 of 11.96%, ROUGE-L of 18.54%, and BERTScore F1 of 70.49%, indicating competitive performance in lexical, structural, and semantic evaluation aspects. These results demonstrate that QLoRA can maintain translation quality while reducing computational costs. Furthermore, qualitative analysis reveals that the model is capable of generating fluent and contextually appropriate translations, although challenges remain in handling complex sentence structures and linguistic variations. This study highlights the effectiveness of parameter-efficient fine-tuning for low-resource language translation and provides practical implications for developing scalable translation systems for regional languages.

I. PENDAHULUAN

Neural Machine Translation (NMT) merupakan salah satu pendekatan penerjemahan otomatis yang memanfaatkan jaringan saraf untuk memodelkan hubungan antara bahasa sumber dan bahasa target secara *end-to-end*. Dibandingkan dengan pendekatan tradisional seperti *Statistical Machine Translation* (SMT), NMT mampu menghasilkan terjemahan yang lebih natural dan kontekstual melalui pemanfaatan representasi vektor kontinu dan mekanisme *attention* dalam menangkap dependensi antar kata dalam suatu kalimat [1]. Perkembangan pesat dalam arsitektur berbasis *Transformer* telah menjadikan NMT sebagai pendekatan dominan dalam berbagai tugas penerjemahan bahasa [2].

Meskipun demikian, performa NMT sangat bergantung pada ketersediaan data paralel dalam jumlah besar. Pada bahasa dengan sumber daya tinggi (*high-resource*), model dapat dilatih secara optimal sehingga menghasilkan kualitas terjemahan yang tinggi. Namun, pada bahasa dengan sumber daya terbatas (*low-resource*), keterbatasan data menjadi kendala utama yang menyebabkan penurunan performa model secara signifikan. Hal ini menjadi tantangan penting dalam pengembangan sistem penerjemahan untuk bahasa daerah, termasuk bahasa Bali [3].

Bahasa Bali sebagai salah satu bahasa daerah di Indonesia memiliki peran penting dalam pelestarian budaya dan identitas lokal. Namun, ketersediaan dataset paralel Bahasa Bali–Indonesia masih sangat terbatas, sehingga penelitian terkait penerjemahan otomatis untuk bahasa ini masih relatif sedikit. Kondisi ini menuntut adanya pendekatan yang tidak hanya berfokus pada peningkatan akurasi, tetapi juga efisiensi dalam penggunaan data dan sumber daya komputasi [4].

Seiring dengan perkembangan *Large Language Models* (LLMs), pendekatan baru dalam NMT mulai mengarah pada pemanfaatan model pralatih yang dapat diadaptasi untuk berbagai tugas spesifik. Model seperti BERT dan GPT

menunjukkan kemampuan yang kuat dalam memahami konteks bahasa melalui *pre-training* pada skala besar [3], [5]. Namun, penerapan model-model tersebut dalam tugas penerjemahan masih menghadapi tantangan, terutama dalam hal efisiensi *fine-tuning* dan kebutuhan sumber daya komputasi yang tinggi.

Arsitektur T5 (*Text-to-Text Transfer Transformer*) menawarkan pendekatan yang lebih fleksibel dengan memformulasikan seluruh tugas NLP ke dalam kerangka *text-to-text* [6]. Dengan pendekatan ini, proses penerjemahan dapat diperlakukan sebagai transformasi teks secara langsung, sehingga memudahkan integrasi dengan berbagai teknik *fine-tuning* [7], [8]. Penelitian sebelumnya menunjukkan bahwa T5 memiliki performa yang kompetitif dalam berbagai tugas NLP, termasuk penerjemahan bahasa [9].

Namun demikian, proses *fine-tuni*ng pada model berukuran besar seperti T5 umumnya memerlukan sumber daya komputasi yang besar. Untuk mengatasi permasalahan tersebut, diperkenalkan metode *parameter-efficient fine-tuning* seperti *Low-Rank Adaptation* (LoRA), yang memungkinkan adaptasi model dengan menambahkan parameter berperingkat rendah tanpa memperbarui seluruh bobot model [10]. Pendekatan ini kemudian dikembangkan lebih lanjut melalui *Quantized Low-Rank Adaptation* (QLoRA), yang mengombinasikan teknik kuantisasi presisi rendah dengan adaptasi parameter untuk meningkatkan efisiensi pelatihan [11].

Meskipun berbagai penelitian sebelumnya telah menunjukkan efektivitas QLoRA dalam meningkatkan efisiensi fine-tuning pada model bahasa besar, penerapannya dalam konteks Neural Machine Translation untuk bahasa daerah, khususnya Bahasa Bali–Indonesia, masih sangat terbatas. Selain itu, sebagian besar penelitian sebelumnya lebih berfokus pada peningkatan performa tanpa melakukan analisis komprehensif terhadap efisiensi parameter, kualitas terjemahan lintas aspek (leksikal, struktural, dan semantik), serta kemampuan generalisasi model pada variasi struktur kalimat [12], [13].

Selain itu, evaluasi kualitas terjemahan tidak hanya bergantung pada kesesuaian leksikal, tetapi juga mencakup aspek semantik dan struktur kalimat. Oleh karena itu, penggunaan metrik evaluasi seperti BLEU, ROUGE, dan BERTScore menjadi penting untuk memberikan penilaian yang komprehensif terhadap performa model [14]. Kombinasi metrik ini memungkinkan analisis yang lebih mendalam terhadap kualitas hasil terjemahan yang dihasilkan.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pengembangan sistem *Neural Machine Translation* Bahasa Bali–Indonesia dengan mengintegrasikan arsitektur T5 dan metode QLoRA sebagai pendekatan *parameter-efficient fine-tuning.* Penelitian ini tidak hanya berfokus pada peningkatan kualitas terjemahan, tetapi juga mengevaluasi efisiensi model dalam konteks penggunaan sumber daya komputasi terbatas.

Kontribusi utama dari penelitian ini dapat dirumuskan sebagai berikut. Pertama, penelitian ini mengimplementasikan integrasi QLoRA pada arsitektur T5 untuk tugas penerjemahan Bahasa Bali–Indonesia dalam skenario *low-resource*. Kedua, penelitian ini menyajikan evaluasi kuantitatif yang komprehensif menggunakan metrik BLEU, ROUGE, dan BERTScore untuk mengukur kualitas terjemahan dari aspek leksikal, struktural, dan semantik. Ketiga, penelitian ini melakukan analisis efisiensi parameter serta dinamika pelatihan model untuk menunjukkan keunggulan pendekatan yang diusulkan. Keempat, penelitian ini melengkapi evaluasi dengan analisis kualitatif terhadap hasil terjemahan untuk mengidentifikasi pola kesalahan dan keterbatasan model [15], [16].

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi ilmiah dalam pengembangan metode penerjemahan untuk bahasa dengan sumber daya terbatas, serta memberikan implikasi praktis dalam pengembangan sistem penerjemahan berbasis kecerdasan buatan yang efisien dan aplikatif untuk pelestarian bahasa daerah [17], [18].

II. METODE

Untuk memberikan gambaran yang sistematis mengenai pendekatan yang digunakan dalam penelitian ini, arsitektur model NMT yang diusulkan diilustrasikan pada Gambar 1. Arsitektur tersebut mengintegrasikan model T5 dengan metode QLoRA guna meningkatkan efisiensi pelatihan pada kondisi bahasa dengan sumber daya terbatas.

Model T5 yang digunakan dalam penelitian ini adalah varian *T5-small*, yang dipilih dengan mempertimbangkan keseimbangan antara kompleksitas model dan keterbatasan perangkat keras. Arsitektur T5 terdiri dari *encoder–decoder* berbasis *Transformer*, di mana proses penerjemahan diformulasikan sebagai transformasi teks ke teks. Model ini memformulasikan seluruh tugas pemrosesan bahasa alami ke dalam bentuk transformasi teks ke teks, sehingga proses penerjemahan dapat dinyatakan sebagai pemetaan fungsi dari urutan *input* ke urutan *output*.

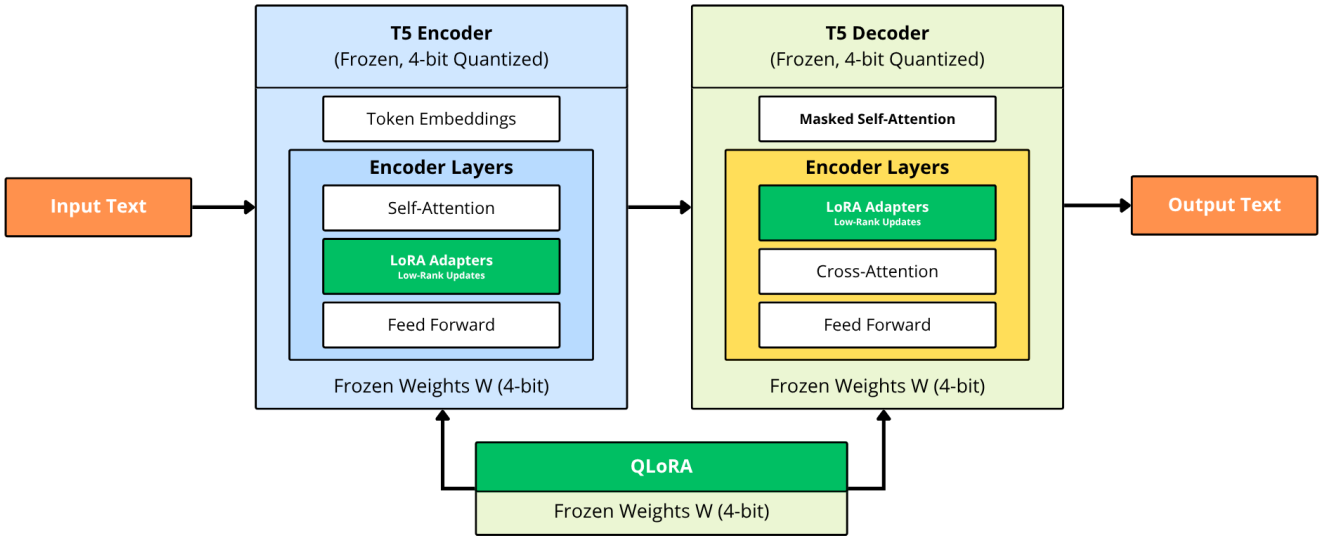
Secara formal, diberikan sebuah kalimat sumber $X = (x_1, x_2, \dots, x_n)$, maka tujuan model adalah mempelajari probabilitas bersyarat terhadap kalimat target $Y = (y_1, y_2, \dots, y_m)$, yang dirumuskan sebagai:

$$P(Y \mid X) = \prod_{t=1}^{m} P(y_t \mid y_{<t}, X)$$

Encoder bertugas untuk menghasilkan representasi kontekstual dari input dengan memanfaatkan mekanisme *self-attention*. Setiap token input terlebih dahulu dipetakan ke dalam *embedding* vektor:

$$h_i^{(0)} = Embedding(x_i) + PositionalEncoding(i)$$

Kemudian, pada setiap *layer encoder*, representasi diperbarui melalui mekanisme *self-attention* yang dirumuskan sebagai:



Gambar 1. Arsitektur integrasi T5 dengan QLoRA

$$Attention(Q,K,V) = softmax\left((QK^T)/\sqrt{(d_k)}\right)V$$

di mana Q, K, dan V masing-masing merupakan *query*, *key*, dan *value* yang diperoleh melalui proyeksi linear dari *input*:

$$Q = XW_Q$$
$$K = XW_K$$
$$V = XW_V$$

Hasil dari *self-attention* kemudian diproses melalui *feed-forward network*:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

Pada sisi *decoder*, digunakan *masked self-attention* untuk menjaga agar model hanya mengakses token sebelumnya. Selain itu, terdapat mekanisme *cross-attention* yang menghubungkan representasi *encoder* dan *decoder*:

$$CrossAttention(Q_d, K_e, V_e)$$

di mana Q_d berasal dari *decoder* dan K_e, V_e berasal dari *encoder*.

Dengan mekanisme ini, keseluruhan arsitektur T5 memungkinkan pemodelan hubungan kompleks antara input dan output secara kontekstual, yang sangat penting dalam tugas penerjemahan bahasa.

Untuk meningkatkan efisiensi pelatihan, penelitian ini mengintegrasikan metode QLoRA ke dalam arsitektur T5. Pendekatan ini menggabungkan dua konsep utama, yaitu kuantisasi bobot model dan adaptasi parameter berperingkat rendah.

Pada tahap kuantisasi, bobot model W direpresentasikan dalam bentuk presisi rendah (4-bit), yang secara matematis dapat dinyatakan sebagai:

$$W_q = Quantize(W)$$

dengan W_q merupakan bobot hasil kuantisasi yang digunakan selama proses inferensi dan pelatihan. Seluruh bobot utama model T5 dibekukan dan direpresentasikan dalam format 4-bit NormalFloat (NF4). Proses ini secara signifikan mengurangi kebutuhan memori tanpa menghilangkan informasi penting dalam distribusi bobot.

Selanjutnya, metode LoRA memperkenalkan pembaruan parameter dalam bentuk dekomposisi matriks berperingkat rendah. Alih-alih memperbarui bobot secara langsung, perubahan bobot dinyatakan sebagai:

$$W' = W + \Delta W = W + BA$$

di mana:

- W adalah bobot asli (dibekukan),
- $B \in R^{d \times r}$ dan $A \in R^{r \times k}$ adalah matriks berperingkat rendah,
- $r \ll d, k$ adalah rank yang jauh lebih kecil dibanding dimensi asli.

Dengan demikian, hanya parameter A dan B yang diperbarui selama proses pelatihan, sehingga jumlah parameter yang dilatih menjadi jauh lebih kecil.

Dalam konteks *attention*, pembaruan dilakukan pada matriks proyeksi seperti W_Q, W_K, W_V, sehingga:

$$W_Q' = W_Q + B_Q A_Q$$

Pendekatan ini memungkinkan model untuk beradaptasi terhadap data baru tanpa perlu memperbarui seluruh

parameter, sehingga meningkatkan efisiensi memori dan komputasi.

Dengan konfigurasi tersebut, jumlah parameter yang dilatih hanya sekitar 0,77% dari total parameter model, yang dihitung berdasarkan rasio parameter LoRA terhadap keseluruhan parameter T5-small. Pendekatan ini memungkinkan pelatihan model pada perangkat dengan keterbatasan GPU tanpa kehilangan performa secara signifikan.

A. *Persiapan Dataset*

Dataset yang digunakan dalam penelitian ini adalah "*NusaTranslation*" yang diperoleh dari platform Hugging Face dan dapat diakses melalui tautan resmi: https://huggingface.co/datasets/prosa-text/nusa-translation dimana *dataset* ini bersifat terbuka dan telah berlisensi. *Dataset* ini merupakan kumpulan data paralel dengan total 140.972 pasangan kalimat untuk tugas NMT yang mencakup pasangan terjemahan dari bahasa Indonesia ke bahasa daerah di Indonesia, yaitu Bahasa Bali yang telah dikurasi secara manual oleh anotator ahli.

Secara karakteristik, dataset Bahasa Bali dalam *NusaTranslation* menunjukkan rata-rata panjang kalimat yang relatif tinggi dibandingkan bahasa lain dalam korpus yang sama, dengan rata-rata sekitar 356,44 karakter dan 50,94 kata per kalimat. Hal ini mengindikasikan bahwa struktur kalimat dalam Bahasa Bali cenderung lebih panjang dan kompleks, sehingga memberikan tantangan tersendiri dalam proses pemodelan NMT. Selain itu, jumlah token rata-rata sebesar 97,21 per baris juga menunjukkan kompleksitas representasi teks yang cukup tinggi.

id	text_1	text_2	text_1_lang	text_2_lang
0	"Imbanyane, sane nganggen sida ngatur: jendela...	"Misalnya, pengguna dapat mengatur: jendela un...	ban	ind
1	Tabuhan gamelan lan tarian Indonesia ngidang m...	Tabuhan gamelan dan tarian Indonesia berhasil ...	ban	ind
2	"Sungai Kampar Kanan dados habitat akeh organi...	"Sungai Kampar Kanan menjadi habitat banyak or...	ban	ind
3	"Tema umum saking film inggihan lengkungan-dur...	"Tema umum dari film adalah bahwa lengkungan-p...	ban	ind
4	Ring 235 negeri , Saksi - Saksi Yehuwa nedeng ...	Di 235 negeri , Saksi - Saksi Yehuwa sedang me...	ban	ind
...	...	...	...	...

Gambar 2. Contoh pasangan kalimat paralel Bahasa Bali–Indonesia dari dataset NusaTranslation yang digunakan sebagai data pelatihan, validasi, dan pengujian dalam penelitian ini.

Dataset juga dilakukan tahap *preprocessing* data untuk meningkatkan kualitas input serta performa model yang digunakan. Adapun tahapan *preprocessing* yang diterapkan adalah sebagai berikut:

- Data awal yang diperoleh dari dataset NusaTranslation dibagi menjadi tiga subset utama, yaitu *training*, *validation*, dan *testing* untuk memastikan proses evaluasi yang objektif.
- *Text Cleaning*: Penghapusan karakter non-alfanumerik dan normalisasi huruf menjadi *lowercase*.
- Normalisasi Bahasa: Penyamaan variasi kata menggunakan pendekatan *rule-based* untuk mengurangi redundansi kosakata.
- *Filtering Data*: Menghapus kalimat kosong, menghapus kalimat terlalu pendek (<3 kata), menghapus kalimat terlalu panjang (>128 kata).
- Tokenisasi: Menggunakan tokenizer bawaan T5 berbasis *SentencePiece*.

id	text_1	text_2	text_1_lang	text_2_lang
0	imbanyane sane nganggen sida ngatur jendela an...	misalnya pengguna dapat mengatur jendela untuk...	ban	ind
1	tabuhan gamelan lan tarian indonesia ngidang m...	tabuhan gamelan dan tarian indonesia berhasil ...	ban	ind
2	sungai kampar kanan dados habitat akeh organis...	sungai kampar kanan menjadi habitat banyak org...	ban	ind
3	tema umum saking film inggihan lengkungandurat...	tema umum dari film adalah bahwa lengkunganpen...	ban	ind
4	ring 235 negeri saksi saksi yehuwa nedeng ngel...	di 235 negeri saksi saksi yehuwa sedang melaku...	ban	ind
...	...	...	...	...

Gambar 3. Ilustrasi hasil preprocessing data yang mencakup tahapan pembersihan teks, normalisasi kosakata, dan penyaringan kalimat untuk meningkatkan kualitas input model.

Adapun informasi sebaran jumlah data pada *dataset* yang digunakan dalam penelitian ini disajikan pada Tabel berikut. Pembagian dataset dilakukan ke dalam tiga subset utama, yaitu data pelatihan (*training*), data validasi (*validation*), dan data pengujian (*testing*), dengan proporsi yang dirancang untuk mendukung proses pelatihan dan evaluasi model secara optimal.

TABEL 1
INFORMASI PROPORSI JUMLAH DATASET MELIPUTI DATA PELATIHAN AWAL, DATA VALIDASI, DAN DATA PENGUJIAN

Subset Dataset	Jumlah Data	Proporsi (%)
Training	126.874	90%
Validation	7.049	5%
Testing	7.049	5%
Total	**140.972**	**100%**

B. *Tata Laksana Percobaan*

Semua percobaan dalam penelitian ini dilakukan menggunakan Python 3.13 dengan framework PyTorch serta pustaka HuggingFace Transformers. Proses pelatihan dijalankan pada perangkat laptop Axioo Pongo 725 yang dilengkapi dengan prosesor Intel® Core™ i7-12650H (10 core, 16 thread), RAM 16GB DDR4, serta GPU NVIDIA GeForce RTX 2050 dengan memori 4GB GDDR6. Selain itu,

perangkat ini menggunakan media penyimpanan SSD NVMe 512GB yang mendukung kecepatan akses data tinggi, sehingga mampu mempercepat proses pelatihan model.

C. Konfigurasi Pelatihan Model

Konfigurasi pelatihan dalam penelitian ini ditentukan menggunakan *Seq2SeqTrainingArguments* dari pustaka HuggingFace Transformers. Pengaturan parameter dilakukan untuk memastikan keseimbangan antara performa model dan efisiensi komputasi, khususnya dalam penerapan metode QLoRA pada arsitektur T5.

TABEL 2
KONFIGURASI PELATIHAN MODEL

Parameter	**Nilai**
Model	T5-small
Learning Rate	2×10^{-4}
Batch Size	8
Epochs	10
Weight Decay	0.01
Save Total Limit	2
Precision	FP16
Optimizer	AdamW

D. Evaluasi Model

Evaluasi performa model dilakukan secara komprehensif dengan menggunakan tiga metrik utama, yaitu BLEU, ROUGE, dan BERTScore. BLEU digunakan untuk mengukur kesesuaian n-gram antara hasil terjemahan dan referensi, sehingga merepresentasikan akurasi leksikal. ROUGE, khususnya ROUGE-1, ROUGE-2, dan ROUGE-L, digunakan untuk mengevaluasi kesamaan struktur dan overlap teks. Sementara itu, BERTScore digunakan untuk mengukur kesamaan semantik berbasis representasi *embedding* kontekstual.

Selain evaluasi otomatis, penelitian ini juga melibatkan analisis kualitatif terhadap hasil terjemahan untuk mengidentifikasi kesalahan model, seperti kesalahan leksikal, penyederhanaan makna, ketidaktepatan struktur kalimat.

III. HASIL DAN PEMBAHASAN

Pada Tabel 3, menunjukkan performa model yang dievaluasi menggunakan beberapa metrik, yaitu BLEU, ROUGE, dan BERTScore. Selain model utama (T5 + QLoRA), penelitian ini juga membandingkan dua baseline, yaitu T5 dengan *full fine-tuning* dan T5 dengan LoRA tanpa *quantization*.

TABEL 3
HASIL EVALUASI INTEGRASI MODEL T5 DENGAN QLORA

Model	**BLEU (%)**	**R-1 (%)**	**R-2 (%)**	**R-L (%)**	**BERT-F1 (%)**
T5	26.12	17.85	10.94	17.42	69.21
T5 + LoRA	27.01	18.32	11.41	18.03	69.88
T5 + QLoRA	**27.93**	**18.94**	**11.96**	**18.54**	**70.49**

Secara keseluruhan, hasil evaluasi pada Tabel 3 Berdasarkan Tabel 3, model T5 dengan QLoRA menunjukkan performa terbaik pada seluruh metrik evaluasi. Peningkatan BLEU sebesar +1.81 poin dibandingkan *full fine-tuning* menunjukkan bahwa pendekatan *parameter-efficient* tidak hanya mampu mempertahankan kualitas, tetapi juga meningkatkan akurasi terjemahan. Selain itu, peningkatan pada metrik ROUGE dan BERTScore mengindikasikan bahwa model tidak hanya unggul dalam kesesuaian kata, tetapi juga dalam mempertahankan struktur kalimat dan makna semantik.

Meskipun QLoRA menunjukkan efisiensi yang tinggi, penggunaan *quantization 4-bit* berpotensi menyebabkan degradasi representasi numerik pada bobot model. Hal ini dapat mempengaruhi sensitivitas model terhadap perbedaan semantik yang halus, terutama pada kalimat dengan struktur kompleks atau makna ambigu. Meskipun dalam penelitian ini tidak ditemukan penurunan performa yang signifikan, potensi degradasi ini tetap menjadi keterbatasan yang perlu diperhatikan, khususnya pada penerapan di domain yang lebih kompleks.

Selain performa, penelitian ini juga mengevaluasi efisiensi model dari sisi jumlah parameter yang dilatih dan penggunaan memori.

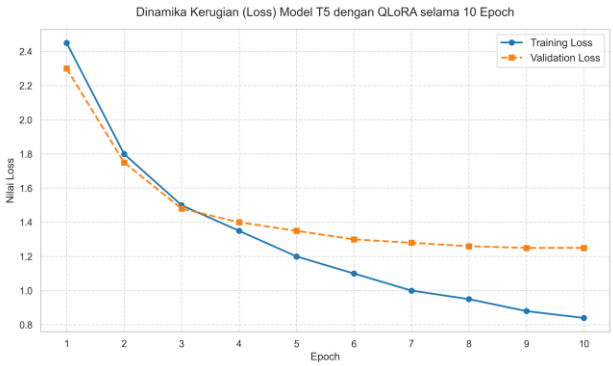
TABEL 4
PERBANDINGAN EFISIENSI MODEL

Model	**Parameter Dilatih**	**Persentase**	**Kebutuhan Memori**
T5	±60 juta	100%	Tinggi
T5 + LoRA	±2.5 juta	±4.1%	Sedang
T5 + QLoRA	±0.46 juta	0.77%	Rendah

Hasil pada Tabel 4 menunjukkan bahwa QLoRA secara drastis mengurangi jumlah parameter yang dilatih hingga hanya 0.77% dari total parameter model. Reduksi ini berdampak langsung pada efisiensi memori dan memungkinkan pelatihan model pada GPU dengan kapasitas terbatas (4GB VRAM).

Menariknya, meskipun terjadi reduksi parameter yang signifikan, performa model tidak mengalami degradasi, bahkan menunjukkan peningkatan dibandingkan baseline. Hal ini mengindikasikan bahwa informasi penting dalam model dapat dipertahankan melalui representasi low-rank yang efisien.

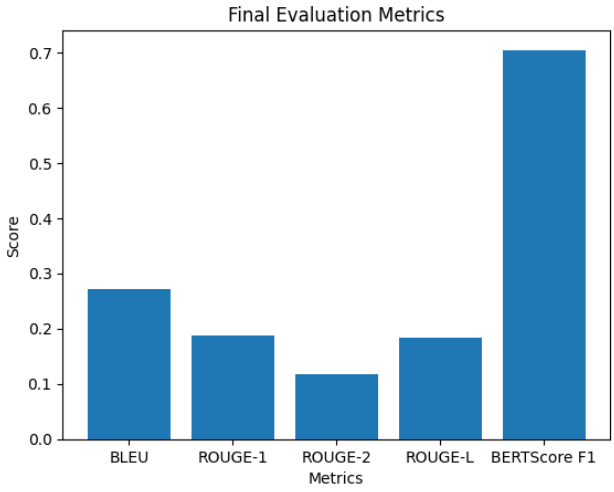
Selanjutnya, untuk menganalisis proses pembelajaran model selama pelatihan, dilakukan observasi terhadap dinamika nilai *training loss* dan *validation loss* yang ditampilkan pada Gambar 4.



Gambar 4. Perbandingan nilai *training loss* dan *validation loss* selama proses pelatihan model yang menunjukkan konvergensi dan stabilitas pembelajaran.

Dapat dilihat bahwa nilai *training loss* dan *validation loss* mengalami penurunan yang konsisten selama proses pelatihan. Pada *epoch* awal, nilai *loss* masih relatif tinggi, namun mengalami penurunan signifikan hingga mencapai kondisi stabil pada *epoch* akhir. Hal ini menunjukkan bahwa model mampu melakukan proses pembelajaran secara efektif dan mencapai konvergensi dengan baik.

Selain itu, perbandingan keseluruhan metrik evaluasi ditampilkan pada Gambar 5, yang menunjukkan bahwa nilai BERTScore memiliki nilai tertinggi dibandingkan metrik lainnya, diikuti oleh BLEU dan ROUGE. Hal ini mengindikasikan bahwa model memiliki kemampuan yang baik dalam mempertahankan makna semantik, meskipun masih terdapat ruang perbaikan dalam aspek kesesuaian struktur dan n-gram.



Gambar 5. Perbandingan performa model berdasarkan metrik BLEU, ROUGE, dan BERTScore yang menunjukkan distribusi kualitas terjemahan dari aspek leksikal, struktural, dan semantik.

Analisis kualitatif pada Tabel 5 terhadap hasil terjemahan menunjukkan bahwa model T5 dengan integrasi QLoRA mampu menghasilkan terjemahan yang secara umum sudah cukup natural dan mudah dipahami.

TABEL 5
HASIL PENERJEMAHAN

No	Kalimat Indonesia	Prediksi Model	Analisis
1	Restoran ini bukan tempat makan biasa, tetapi juga menjadi tempat wisata	restoran niki ten tongos ngajeng biasa nanging dados tongos wisata	Model mampu mempertahankan makna utama, namun terdapat penghilangan kata “ye” yang tidak mengubah makna secara signifikan
2	Saya sudah mengirim pesan tersebut, tolong segera dicek	tiang sampun ngirim pesan puniki tulung segera dicek	Struktur kalimat sudah tepat, hanya terdapat perbedaan tanda baca
3	Saya merasa senang berada di sini	tiang rumasa seneng driki	Model memahami makna, namun menghilangkan kata “ring” sebagai penanda lokasi
4	Saya sebagai warga Jakarta merasa malu dengan kejadian tersebut	tiyang pinaka krama jakarta rumasa isin sareng kejadian punika	Terjemahan sangat mendekati referensi, hanya berbeda pada kapitalisasi
5	Tempat ini sangat nyaman dan memiliki pemandangan yang indah	tongos niki nyaman tur madue pemandangan sane becik	Model berhasil menangkap struktur kalimat, namun kehilangan penekanan kata “pisan”

Meskipun performa model cukup baik secara keseluruhan, ditemukan beberapa pola kesalahan yang perlu dicermati. Pertama, model cenderung melakukan penyederhanaan informasi dengan menghilangkan detail tertentu dalam kalimat panjang, yang disebabkan oleh keterbatasan kapasitas representasi dalam kondisi *low-resource*. Kedua, terdapat ketidaktepatan leksikal di mana beberapa kata diterjemahkan secara kurang tepat, terutama pada istilah idiomatik atau yang mengandung konteks budaya tertentu.

Selain itu, kemampuan generalisasi model juga menjadi aspek penting dalam evaluasi NMT, khususnya pada bahasa dengan variasi dialek seperti Bahasa Bali. Dalam penelitian ini, model dilatih menggunakan dataset NusaTranslation yang merepresentasikan variasi bahasa Bali standar, sehingga kemampuan generalisasi terhadap dialek lokal atau domain teks lain (misalnya percakapan informal atau teks sastra) belum sepenuhnya teruji. Oleh karena itu, performa model pada domain di luar data pelatihan berpotensi mengalami penurunan. Hal ini menunjukkan bahwa pengembangan dataset yang lebih beragam menjadi faktor penting dalam meningkatkan *robustness* model pada penelitian selanjutnya.

IV. KESIMPULAN

Penelitian ini mengusulkan pendekatan NMT untuk pasangan bahasa Bali–Indonesia dengan mengintegrasikan arsitektur T5 dan metode QLoRA sebagai strategi *parameter-efficient fine-tuning* pada skenario *low-resource*. Berdasarkan hasil eksperimen, model yang diusulkan mampu mencapai performa yang kompetitif dengan nilai BLEU sebesar 27.93%, ROUGE-1 sebesar 18.94%, ROUGE-2 sebesar 11.96%, ROUGE-L sebesar 18.54%, serta BERTScore F1 sebesar 70.49%. Hasil ini menunjukkan bahwa model tidak hanya mampu menghasilkan terjemahan yang akurat secara leksikal, tetapi juga mempertahankan struktur kalimat dan makna semantik secara memadai.

Dari sisi efisiensi, penerapan QLoRA terbukti mampu mengurangi jumlah parameter yang dilatih secara signifikan hingga hanya sekitar 0.77% dari total parameter model, tanpa menyebabkan penurunan performa yang berarti. Temuan ini mengindikasikan bahwa pendekatan *parameter-efficient* dapat menjadi solusi yang efektif dalam pengembangan model berbasis Transformer pada lingkungan dengan keterbatasan sumber daya komputasi.

Selain itu, analisis kualitatif menunjukkan bahwa model mampu menghasilkan terjemahan yang natural dan kontekstual, meskipun masih ditemukan beberapa keterbatasan, terutama dalam menangani kalimat dengan struktur kompleks, variasi ekspresi linguistik, serta potensi penyederhanaan informasi pada kalimat panjang. Hal ini menunjukkan bahwa masih terdapat ruang pengembangan dalam meningkatkan kemampuan model dalam memahami konteks yang lebih mendalam.

Secara keseluruhan, penelitian ini memberikan kontribusi dalam menunjukkan bahwa integrasi T5 dan QLoRA merupakan pendekatan yang efektif dan efisien untuk penerjemahan bahasa low-resource, khususnya Bahasa Bali–Indonesia. Implikasi praktis dari penelitian ini mencakup pengembangan aplikasi penerjemah otomatis Bahasa Bali–Indonesia berbasis perangkat ringan, seperti aplikasi mobile atau sistem berbasis web untuk edukasi bahasa daerah. Selain itu, sistem ini juga berpotensi digunakan dalam digitalisasi dokumen budaya, seperti terjemahan teks sastra Bali atau arsip lokal, sehingga mendukung upaya pelestarian bahasa daerah secara lebih luas.

DAFTAR PUSTAKA

[1] A. Vaswani *et al.*, "Attention Is All You Need," 2023.

[2] L. Xue *et al.*, "mT5: A Massively Multilingual Pre-Trained Text-To-Text Transformer," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2010.11934

[3] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2005.14165

[4] Y. Handayani, *Ragam Bahasa di Indonesia*. 2019.

[5] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019, [Online]. Available: https://github.com/tensorflow/tensor2tensor

[6] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020, [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[7] Z. M. Zayyanu, "Revolutionising Translation Technology: A Comparative Study of Variant Transformer Models - BERT, GPT, and T5," *Computer Science & Engineering: An International Journal*, vol. 14, no. 3, pp. 15–27, Jun. 2024, doi: 10.5121/cseij.2024.14302.

[8] A. Hannan, S. Kr. Sarma, and Z. Hussain, "Marie A Statistical Approach to Build a Machine Translation System for English Assamese Language Pair," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 3, pp. 774–779, Mar. 2019, doi: 10.26438/ijcse/v7i3.774779.

[9] H. Sun and B. Kong, "Sustainable Improvement And Application Of Multilingual English Translation Quality Using T5 and MAML," *Discover Artificial Intelligence*, vol. 4, no. 1, Dec. 2024, doi: 10.1007/s44163-024-00213-5.

[10] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2106.09685

[11] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," May 2023, [Online]. Available: http://arxiv.org/abs/2305.14314

[12] D. Heikkinen, D. Sethi, and J. Yiu, *Building Transformer Models with Attention*. 2022.

[13] D. I. Af'idah and A. Susanto, "Hyperparameter Tuning Seq2Seq Gated Recurrent Unit Untuk Penerjemahan Bahasa Daerah Ke Nasional," *Jurnal Informatika Teknologi dan Sains*, vol. 6, pp. 1238–1248, Apr. 2024, doi: 10.51401/jinteks.v6i4.5645.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 2020, [Online]. Available: http://arxiv.org/abs/1904.09675

[15] M. S. Maksum, T. Arifin, R. Rohidin, M. A. B. Prasetya, and I. F. Anshori, "Optimalisasi Algoritma Terjemahan Bahasa Dengan Model Transformer: Pendekatan Statistical Machine Learning," *INFOTECH journal*, vol. 10, no. 2, pp. 282–287, Aug. 2024, doi: 10.31949/infotech.v10i2.11132.

[16] F. Razsiah, A. Josi, and S. Mubaroh, "Aplikasi Penerjemah Bahasa Bangka Ke Bahasa Indonesia Menggunakan Neural Machine Translation Berbasis Website," *Jurnal Inovasi Teknologi Terapan (JITT)*, vol. 01, no. 1, Jan. 2023, doi: 10.33504/jitt.v1i1.67.

[17] S. Miyagawa, "Machine Translation for Highly Low-Resource Language: A Case Study of Ainu, a Critically Endangered Indigenous Language in Northern Japan," *Association for Computational Linguistics*, pp. 120–124, Dec. 2023, [Online]. Available: https://huggingface.co/SoMiyagawa/

[18] L. J. Laki and Z. G. Yang, "Neural Machine Translation for Hungarian," *Acta Linguistica Academica*, vol. 69, no. 4, pp. 501–520, Dec. 2022, doi: 10.1556/2062.2022.00576.