Transfer Learning Menggunakan LoRA+ pada Llama 3.2 untuk Percakapan Bahasa Indonesia

Transfer Learning Using LoRA+ on Llama 3.2 for Indonesian Conversation

Faiz Muhammad Kautsar[1], Abdan Hafidz[2], Muhammad Farhan Arya Wicaksono[3], Diana Purwitasari[4], Nanik Suciati[5], Dini Adni Navastara[6], Ilham Gurat Adillion[7]

[1,2,3,4,5,6,7] *Departemen Teknik Informatika, Institut Teknologi Sepuluh Nopember*

E-mail: [1]5054231013@student.its.ac.id, [2]5054231024@student.its.ac.id, [3]5054231011@student.its.ac.id, [4]diana@its.ac.id, [5]nanik@its.ac.id, [6]dini_navastara@its.ac.id, [7]ilhamgurata@its.ac.id

Received 15 March 2025; Revised 28 April 2025; Accepted 29 April 2025

Abstrak

Penelitian ini mengeksplorasi penerapan dari metode *Parameter-Efficient Finetuning* (PEFT) *Low-Rank Adaptation+* (LoRA+) pada *transfer learning* model Llama 3.2 1B, sebuah model bahasa besar. Seiring bertambahnya ukuran model bahasa, *finetuning* yang dilakukan secara konvensional dalam transfer learning semakin tidak fisibel untuk dilakukan tanpa menggunakan komputasi skala besar. Untuk menangani hal tersebut, dapat dilakukan *finetuning* pada beberapa komponen saja, menggunakan komputasi yang relatif minimal berbanding dengan *finetuning* konvensional, metode-metode yang menerapkan prinsip ini disebut juga sebagai PEFT. Penelitian menguji efektifitas metode PEFT, yakni LoRA+, pada *transfer learning* model bahasa besar terhadap domain baru, yakni bahasa Indonesia, menggunakan metrik BLEU, ROUGE, serta Weighted F1. Hasil penelitian menunjukkan bahwa penerapan LoRA+ menghasilkan performa kompetitif dan unggul terhadap *baseline* dalam kemampuan berbahasa Indonesia, dengan peningkatan 112% pada skor BLEU dan 21.7% pada skor ROUGE-L, dengan standar deviasi yang relatif rendah sebesar 3.72 dan 0.00075. Meskipun terjadi penurunan pada skor Weighted F1 sebesar 13% yang disebabkan oleh *domain shift*, model menunjukkan kemampuan transfer lintas-bahasa yang baik.

Kata kunci: *Finetuning*, Model Bahasa Besar, *Parameter-Efficient Finetuning*, *Low-Rank Adaptation*, *Transfer Learning*

Abstract

This research explores the application of the Parameter-Efficient Finetuning (PEFT) method Low-Rank Adaptation+ (LoRA+) on the transfer learning of Llama 3.2 1B, a large language model (LLM). As LLMs continue to grow in size, conventional finetuning becomes increasingly challenging and this approach quickly becomes computationally infeasible. To address this, the application of finetuning could be applied efficiently to certain components only, methods which are also called PEFT. The PEFT method LoRA+ is tested within this research for the transfer learning of Llama 3.2 1B to a new linguistic domain, that is the Indonesian language, using the metrics BLEU, ROUGE, and Weighted F1. The results show that the application of the LoRA+ method yields competitive and superior performance compared to the baseline in Indonesian linguistic capabilities, with improvements of 112% in BLEU scores and 21.7% in ROUGE-L scores, while maintaining relatively low standard deviations of 3.72 and 0.00075, respectively. Though Weighted F1 scores showed an expected decline of 13% due to domain shift, the model demonstrated robust cross-lingual transfer capabilities.

Keywords: *Finetuning, Large Language Model, Parameter-Efficient Finetuning, Low-Rank Adaptation, Transfer Learning*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

1. PENDAHULUAN

Model bahasa, dikenal pula dalam bahasa Inggris sebagai *Language Model*, adalah model yang memodelkan sebuah korpus teks dalam suatu bahasa melalui pemodelan probabilistik. Model bahasa dengan ukuran besar disebut sebagai model bahasa besar, atau *Large Language Model* (LLM). LLM komersial pada umumnya melakukan tahapan *pretraining* pada korpus *dataset* yang predominan tersaturasi oleh teks yang berbahasa Inggris. Kelemahan yang timbul akibat saturasi berlebih tersebut adalah bias terhadap kapabilitas sebuah model yang berorientasi pada kemampuan berbahasa Inggris. Bias terhadap data yang berorientasi pada sebuah bahasa tersebut dapat menyebabkan peningkatan terhadap bias dalam domain-domain lainnya yang tidak berkaitan dengan kapabilitas berbahasa pula, seperti bias kultural [1], bias gender [2], dan sebagainya. [3]

Dataset Cendol Collection adalah dataset *low-resource* yang bertujuan untuk mengaugmentasi kekurangan dari representasi bahasa Indonesia dalam dataset umum, dataset ini merupakan kurasi dari beberapa dataset-dataset yang tersaturasi oleh data bahasa-bahasa Indonesia. Bersama dengan dataset tersebut, dilakukan pula riset mengenai efektivitas dari penerapannya pada beberapa model, seperti Llama 2 dan mT5. Hasil dari penelitian Cendol menunjukkan bahwa penerapan dataset yang memiliki predominansi saturasi pada rumpun bahasa Indonesia dapat meng-augmentasi kekurangan dari pemahaman suatu model dalam berbahasa Indonesia. [4]

Untuk meng-augmentasi kekurangan atau ketidakmampuan sebuah model pada sebuah *domain*, dapat dilakukan proses *transfer learning*, yakni merupakan penggunaan pengetahuan sebuah model terhadap domain yang sudah ada untuk membantu pemahaman pada domain yang baru. [5] Dalam *transfer learning*, terdapat beberapa metodologi mengenai bagian dari model yang digunakan pengetahuan yang sudah ada pada domain baru, salah satu metode yang umum digunakan adalah *fine-tuning*, yakni proses pelatihan sebuah model general terhadap dataset yang berorientasi pada suatu *domain*.

Tahap *pretraining* sebuah model cenderung cukup berat secara komputasional, ini disebabkan oleh komputasi yang dilakukan pada presisi *floating-point* yang tinggi serta ukuran dataset yang cukup besar dan waktu *training* yang lama agar kemampuan emergen muncul. [6, 7] Maka dari itu, selain untuk mengadaptasikan domain sebuah model, untuk mempersingkat dan mengurangi komputasi yang diperlukan untuk melakukan *training*, dapat dilakukan proses *finetuning*.

Finetuning dapat dilakukan dengan melakukan prosedur yang identik dengan *pretraining*, di mana keseluruhan parameter dari model dilatih kembali untuk menyesuaikan dengan sebuah dataset, disebut juga sebagai *full finetuning* atau *continual pretraining*, atau melalui prosedur yang hanya mengubah beberapa parameter saja, yang disebut juga secara general sebagai *parameter-efficient finetuning* (PEFT). [8] Salah satu metode PEFT yang sering digunakan adalah *Low-Rank Adaptation* (LoRA) [9], yang melakukan perubahan aproksimasi perubahan matriks *weight* menggunakan dua matriks *low-rank* untuk perubahan parameter. Kelemahan dari metode tersebut adalah asumsi bahwa kedua matriks *low-rank* harus diubah menggunakan *learning rate* yang sama, LoRA+ [10] memberikan perubahan pada prosedur ini di mana untuk kedua matriks dapat digunakan *learning rate* yang berbeda pada suatu rasio di mana untuk kedua matriks, A dan B, *learning rate* B adalah λ-kalinya *learning rate* A.

Penelitian sebelumnya, seperti Owen et al. [11] dan Cahyawijaya et al. [4], menunjukkan bahwa penerapan *transfer learning* dapat meningkatkan performa sebuah model pada domain berbahasa Indonesia. Akan tetapi, kekurangan dari penerapan-penerapan sebelumnya adalah kurangnya penekanan pada efisiensi komputasi metode pelatihan yang digunakan, menghasilkan keperluan komputasi yang cukup tinggi untuk pelatihan berkala. Oleh karena itu, perlu dilakukan pengkajian lanjutan pada metode-metode yang dapat menerapkan *transfer learning* dengan komputasi yang efisien.

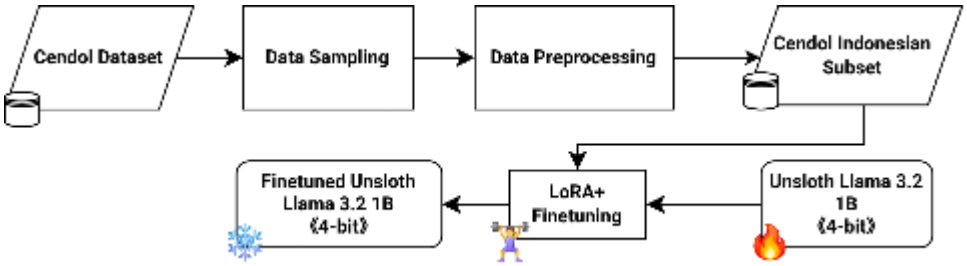
e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

Penerapan LoRA+ pada penelitian ini berdasarkan pada keuntungan fleksibilitas pada penyesuaian *learning rate* yang berbeda untuk kedua matriks *low-rank*. Perubahan ini memungkinkan percepatan konvergensi pada proses pelatihan, di mana dengan lebih cepat tercapainya konvergensi, iterasi pembaharuan pada model dapat dilakukan dengan lebih efisien melalui amortisasi biaya komputasi. Selain itu, dengan membedakan learning rate pada kedua matriks, A dan B, properti konvergensi yang dicapai pada sebuah model akan lebih optimal dan stabil, terutama pada model dengan kelebaran *layer* jaringan yang besar. [10]

Penelitian-penelitian sebelumnya tidak mempertimbangkan stabilitas maupun kualitas dari konvergensi atas adapter matriks *low-rank* yang dihasilkan oleh prosedur pelatihan yang diterapkan. Maka dari itu, penelitian ini bertujuan untuk menerapkan metode LoRA+ dalam *transfer learning* dari *domain* awal Llama 3.2 [12] berbahasa Inggris ke bahasa Indonesia secara efisien dan optimal dalam komputasi, serta mengimplementasikan model tersebut.

2. METODE PENELITIAN

Metodologi penelitian terbagi menjadi empat tahap utama: 1.) Pengumpulan dan *sampling* data; 2.) *Preprocessing* data; 3.) *Training* model; dan 4.) Evaluasi model. Berikut adalah diagram alir yang menggambarkan metodologi penelitian secara garis besar:



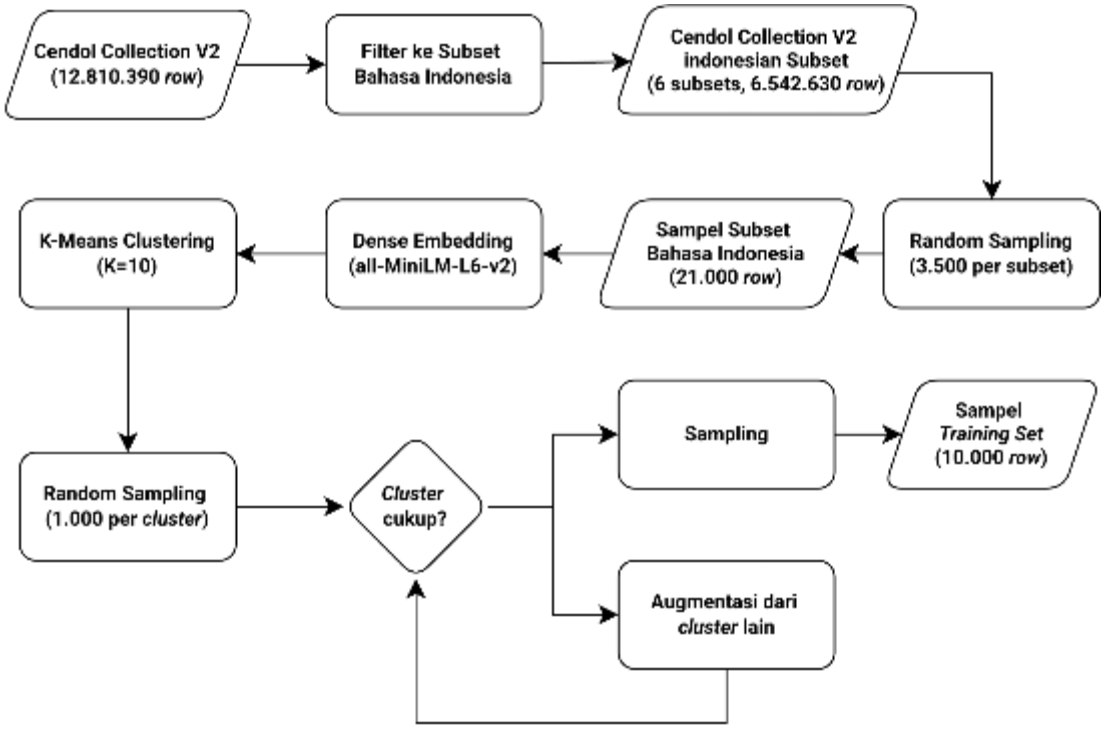
Gambar 1 Diagram alir garis besar metodologi penelitian

2.1 Pengumpulan Data

Dataset yang digunakan adalah Cendol Collection [4], sebuah dataset instruksi bahasa *low-resource* yang mencakup Bahasa Indonesia serta bahasa-bahasa daerah Indonesia lainnya. Untuk mempercepat konvergensi dan efisiensi komputasi, penelitian hanya mengambil subset data yang berbahasa Indonesia saja. Dilakukan pengambilan subset dengan rasio 0.08% dari dataset asli melalui prosedur sebagai berikut:

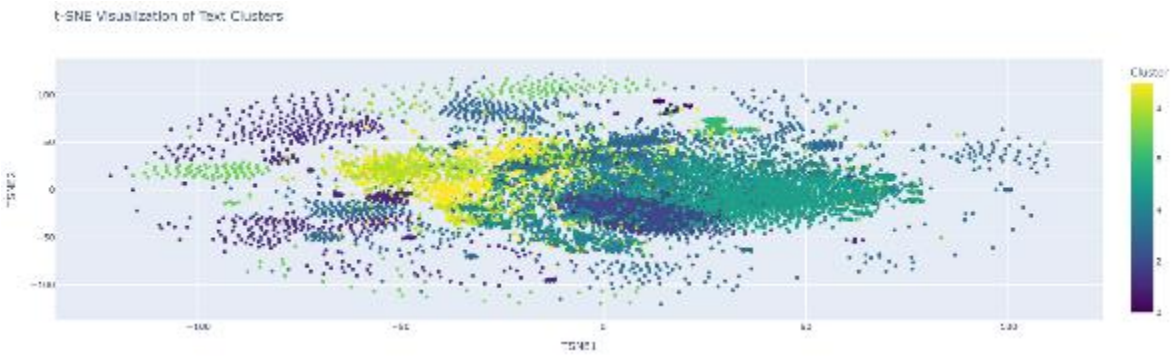
1) Dari dataset Cendol Collection [4] diambil subset-subset berbahasa Indonesia, yakni "indo_puisi", "wikihow", "wikipedia_id", "safety_prompt", "identity_prompt", dan "dolly", yang menghasilkan sampel ±50% populasi awal, menghasilkan sampel berukuran 6.542.630 instansi.
2) Diambil 3.500 instansi dari tiap subset-subset tersebut dengan random sampling, menghasilkan total sampel 21.000 instansi, ±0,3% sampel awal.
3) *Output* tiap set kemudian diambil *dense embedding* menggunakan model "all-MiniLM-L6-v2".
4) Untuk meminimalisir jumlah sampel, dilakukan K-Means clustering dengan K=10. Pengambilan jumlah K berdasarkan pertimbangan efisiensi komputasi serta rasio hasil akhir setelah pengambilan instansi dari tiap *cluster*.
5) Tiap *cluster* diambil 1.000 instansi, dengan augmentasi dari *cluster* lain jika tidak mencukupi, menghasilkan sampel 10.000 instansi.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508



Gambar 2 Diagram alir metode *sampling*
yang digunakan

Hasil dari pengambilan subset menghasilkan 10.000 sampel data yang merepresentasikan variasi dalam dataset asli sesuai dengan data yang ter-*cluster*. Gambar 3 menunjukkan visualisasi clustering yang dilakukan, dengan reduksi dimensi menjadi 2 dimensi menggunakan t-SNE, di mana posisi tiap teks dalam korpus divisualisasi posisinya pada *embedding space* "all-MiniLM-L6-v2" yang tereduksi:

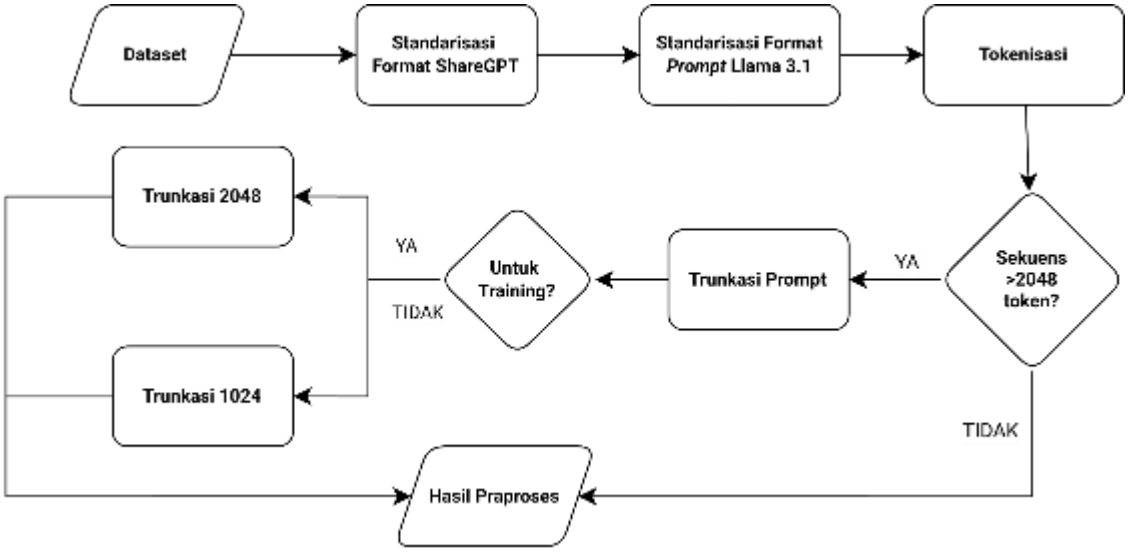


Gambar 3 Visualisasi *clustering sampling*
dengan reduksi t-SNE 2 dimensi

2.2 Preprocessing

Dalam tahap *preprocessing*, dilakukan tokenisasi serta *formatting* untuk menyesuaikan dengan format *instruction-tuning* yang digunakan oleh *base model* agar adapter LoRA dapat direspon dengan baik oleh model.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508



Gambar 4 Diagram alir proses *preprocessing*

Tokenisasi yang dilakukan menggunakan *tokenizer* dari *base model*, dengan format *instruction-tuning* digunakan merupakan format Llama 3.1 [12], mengikuti spesifikasi dari *base model*. *Sequence length* maksimum 2048 *token*, dengan *preprocessing* melakukan trunkasi pada *prompt* yang terlalu panjang pada 2048 *token* untuk *batch* pelatihan dan 1024 *token* untuk *batch* evaluasi.

Tabel 1 Langkah-langkah proses preprocessing beserta sampel hasil tiap langkah

	Langkah	Hasil
1	*Input Data*	{'dataset_name': 'nusa_t2t_v2', 'subset_name': 'wikipedia_id', 'prompt_id': 'cd34b736-1ce0-48c1-b9ed-eab9342a419f', 'template_name': 'wikipedia_subsection_4', 'dataset_key': 'train', 'input': 'Gw mau tau Komposisi dari HD 40307 g', 'output': 'Penemu kode Hugh Jones, dari University of Hertfordshire di Inggris, …"'}
2	SHAREGPT *Formatting*	{'content': 'Gw mau tau Komposisi dari HD 40307 g', 'role':'user'}, {'content': "Penemu kode Hugh Jones, dari University of Hertfordshire di Inggris … , 'role': 'assistant'}]
3	Llama 3.1 *Formatting*	<|begin_of_text|><|start_header_id|>system<|end_header_id|>\n\nCutting Knowledge Date: December 2023\nToday Date: 26 July 2024\n\n<|eot_id|><|start_header_id|>user<|end_header_id|>\n\nGw mau tau Komposisi dari HD 40307 g<|eot_id|><|start_header_id|>assistant<|end_header_id|>\n\n"Penemu kode Hugh Jones, dari University of Hertfordshire di

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

	Langkah	Hasil
		Inggris …".<|eot_id|
4	Tokenisasi	tensor([128000, 128000, 128006, 9125, 128007, 271, 38766, 1303, 33025, ...], device='cuda:0')
5	Trunkasi	tensor([128000, 128000, 128006, 9125, 128007, 271, 38766, 1303, 33025, ...], device='cuda:0')
6	*Attention Masking*	tensor([0, 0, 0, 0, 0, 0, ..., 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, ...], device='cuda:0')
7	Hasil Akhir	\n\n"Penemu kode Hugh Jones, dari University of Hertfordshire di Inggris … <|eot_id|>

Attention masking diterapkan pada prosedur pelatihan untuk mempersempit objektif optimasi dari model, di mana model hanya dapat melakukan komputasi *loss attention* pada posisi-posisi yang berada di *mask* dengan valuasi 1. Hal ini dilakukan dengan rasionalisasi bahwa prosedur pelatihan pada model hanya perlu mengubah bagaimana model merespon sebuah input *prompt*, sehingga bagian dari input yang berupa *prompt* tidak dikomputasikan *loss*-nya.

2.3 Model Training

Base model digunakan adalah Llama 3.2 [12] varian ukuran 1 miliar parameter kuantisasi 4-bit yang dilakukan oleh unsloth. Proses pelatihan dilakukan menggunakan LoRA+ *rank* 32 dengan LoRA α 16, dengan stabilisasi *rank* pada *learning rate* awal 2×10^{-4} menggunakan *cosine annealing scheduler*. Proses pelatihan dilakukan hingga 5 *epoch*, menghasilkan adapter berukuran 22.544.384 parameter.

2.4 Model Testing dan Evaluation

Dalam tahap evaluasi, penelitian menerapkan beberapa metrik pengukuran standar yang umum digunakan dalam evaluasi model *Natural Language Processing* (NLP) yaitu *Bilingual Evaluation Understudy* (BLEU), *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE), dan Weighted F1 Score. Ketiga metrik ini dipilih untuk mendapatkan perbandingan komprehensif terkait performa model sebelum dan sesudah proses training dilakukan.

BLEU score digunakan untuk mengukur kualitas teks yang dihasilkan model dengan membandingkannya terhadap teks referensi manusia. Sementara ROUGE berfokus pada pengukuran *recall* yang mengevaluasi seberapa banyak kata-kata dari referensi yang berhasil dimunculkan dalam output model. Weighted F1 Score memberikan nilai keseimbangan antara precision dan recall dengan mempertimbangkan bobot dari masing-masing kelas.

Parameter pengujian adalah sebagai berikut, dengan semua pengujian dilakukan 5 kali agar didapatkan standar deviasi untuk menangani stokastisitas dalam *sampling* yang dilakukan saat *generation*, di mana untuk parameter *sampling* digunakan *default* dari *library* transformers:

1) BLEU dilakukan dengan ukuran korpus testing 9800 sampel, dengan referensi 1 ukuran 48 token;
2) Pengujian ROUGE menggunakan varian ROUGE-L, ROUGE-2, ROUGE-1, dengan ukuran per sampel 48 token; dan
3) Weighted F1 Score dikalkulasi menggunakan kemiripan semantik *embedding* menggunakan model "all-MiniLM-L6-v2" pada metrik kemiripan kosinus dengan ukuran 32 token dengan *threshold* kemiripan 0.5 untuk menentukan "kebenaran" sebuah *output*.

2.5 Model Deployment

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

Langkah akhir dalam penerapan sistem machine learning adalah memastikan model yang telah bersertifikasi dapat diimplementasikan secara efektif. Pada tahap ini, insinyur operasional ML bertanggung jawab untuk menjawab pertanyaan kritis terkait infrastruktur, seperti bagaimana data baru akan diintegrasikan ke dalam model, apakah prediksi dilakukan secara batch atau satu per satu, toleransi latensi yang dapat diterima, serta mekanisme interaksi pengguna dengan sistem.

Dalam penelitian ini, dikembangkan sebuah aplikasi berbasis web menggunakan framework Django yang mengimplementasikan *design pattern Model-View-Controller* (MVC) dengan integrasi layanan prediksi menggunakan *WebSocket*. Alur sistem dijelaskan sebagai berikut:

1. **Pengaksesan Website**: User mengakses antarmuka web yang dibangun dengan Django. Antarmuka ini memungkinkan user untuk mengirimkan *prompt* melalui sebuah antarmuka *chat box.*
2. **Pengiriman Prompt ke Backend**: Setelah user mengirimkan *prompt*, data diterima oleh *controller*. Controller bertanggung jawab memvalidasi dan meneruskan *prompt* ke layanan prediksi.
3. **Worker Mengeksekusi Prediksi**: Layanan prediksi dijalankan oleh subproses *worker* yang telah meng-*load* model sebelumnya. *Worker* menjalankan tugas prediksi secara terpisah dari proses utama Django dengan memproses *prompt* input melalui model dan mengembalikan *stream*.
4. **Penggunaan Redis untuk Komunikasi Antarproses**: Redis digunakan sebagai *cache database* untuk menyimpan *prompt* dan status hasil prediksi. Worker dan Django dapat berkomunikasi dengan bertukar data melalui Redis
5. **Streaming Hasil Prediksi melalui WebSocket**: Setelah worker mengirimkan hasilnya ke Redis, WebSocket digunakan untuk mengirimkan hasil prediksi secara *real-time* ke klien secara kontinu.

3. HASIL DAN PEMBAHASAN

Penelitian melakukan beberapa iterasi percobaan alterasi pada metodologi sebagai studi ablasi. Prosedur akhir yang digunakan merupakan prosedur yang diterapkan pada iterasi ke-6 (5 pada tabel, konvensi penulisan semua figur bermulai dari indeks 0). Perubahan pada tiap iterasi dapat dilihat pada tabel 2.

Tabel 2 Perbandingan Metodologi Tiap Iterasi Beserta *Hyperparameter* dan *Dataset Sampling Method*

Model	r	α	Rank Stab.	Sampling Method	Scheduler	Initial LR	Epochs	Train Loss	Val Loss
0	16	16	FALSE	Random (50k)	cosine	1.00E-04	1	1.551	NA
1	16	16	FALSE	6 subsets, 1k each	linear	2.00E-04	5	1.955	NA
2	16	16	FALSE	6 subsets, 1k each	cosine with restarts	2.00E-04	11.6	1.585	1.721
3	16	16	FALSE	6 subsets, 2k each	cosine with restarts	2.00E-04	5	1.717	NA
4	32	16	TRUE	Stratified 1%	cosine	1.00E-04	1	2.13	2.012
5	32	16	TRUE	6 subsets, 3.5k each, K=10, 1k per cluster	cosine	2.00E-04	5	1.526	1.495

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

Model	r	α	Rank Stab.	Sampling Method	Scheduler	Initial LR	Epochs	Train Loss	Val Loss
6	32	16	TRUE	Stratified 1%	cosine	1.00E-04	1	2.13	2.012
7	32	16	FALSE	Stratified 1%, K=50, 1k per cluster	cosine	1.00E-04	1	2.254	1.969
8	32	32	TRUE	Stratified 16.7%, K=100, 1k per cluster	cosine	1.00E-04	1	2.072	1.783
9	32	4	TRUE	Stratified 16.7%, K=100, 1k per cluster	cosine	1.00E-04	1	2.158	1.883

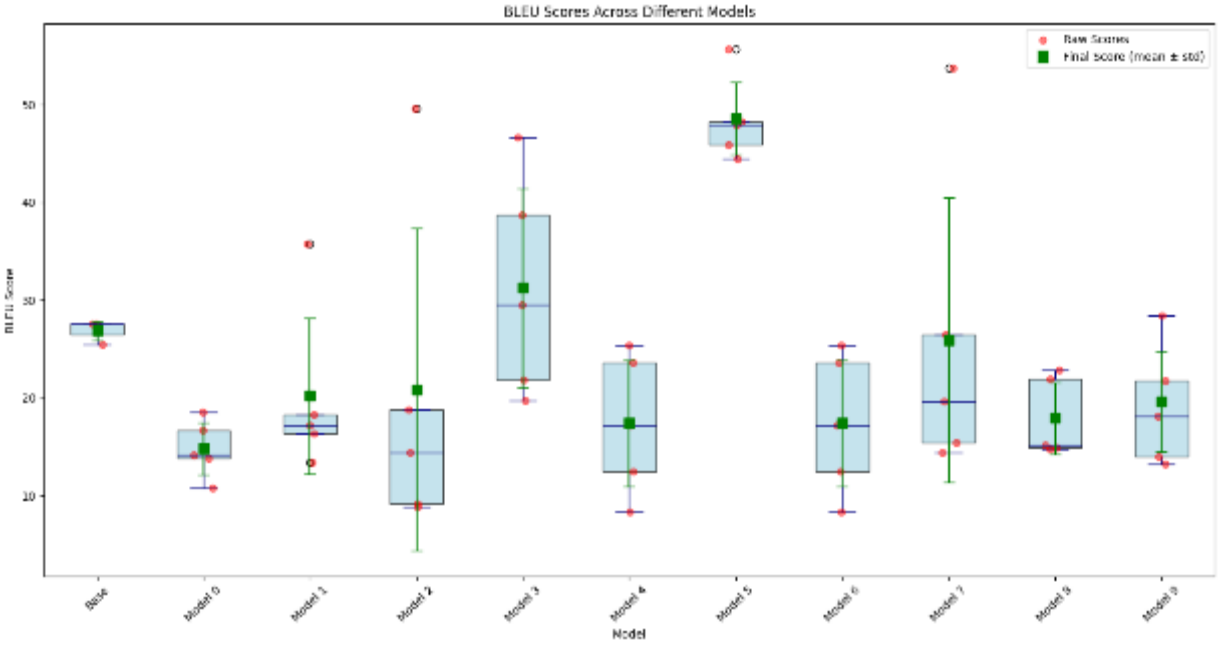
Hasil evaluasi pada Tabel 3 menunjukkan bahwa iterasi 6 merupakan *overall best performer* berdasarkan pertimbangan stabilitas performa dan peningkatan performa, di mana untuk mayoritas metrik selain *weighted F1*, standar deviasi merupakan terendah berbanding terhadap peningkatan performa.

Tabel 3 Perbandingan metrik pada tiap iterasi metodologi.
Model 5 terpilih sebagai *overall best performer*, di-**bold**.

Model Type	BLEU (Mean ± Std)	ROUGE-1	ROUGE-2	ROUGE-L (Mean ± Std)	Weighted F1 (Mean ± Std)
Base	24.76 ± 1.20	0.055	0.012	0.046 ± 0.0008	0.503 ± 0.003
Model 0	17.45 ± 2.65	0.048	0.011	0.042 ± 0.00049	0.385 ± 0.004
Model 1	28.15 ± 7.96	0.05	0.013	0.043 ± 0.00049	0.379 ± 0.005
Model 2	37.39 ± 16.54	0.041	0.01	0.035 ± 0.00049	0.36 ± 0.004
Model 3	41.45 ± 10.18	0.067	0.018	0.057 ± 0.00075	0.482 ± 0.006
Model 4	23.88 ± 6.47	0.061	0.02	0.054 ± 0.0008	0.398 ± 0.005
Model 5	**52.35 ± 3.72**	**0.062**	**0.023**	**0.056 ± 0.00075**	**0.438 ± 0.002**
Model 6	23.88 ± 6.47	0.061	0.02	0.054 ± 0.0008	0.398 ± 0.005
Model 7	40.46 ± 14.53	0.053	0.016	0.047 ± 0.00049	0.38 ± 0.004
Model 8	21.58 ± 3.67	0.057	0.019	0.051 ± 0.00075	0.384 ± 0.002
Model 9	24.72 ± 5.12	0.053	0.018	0.048 ± 0.00063	0.356 ± 0.008

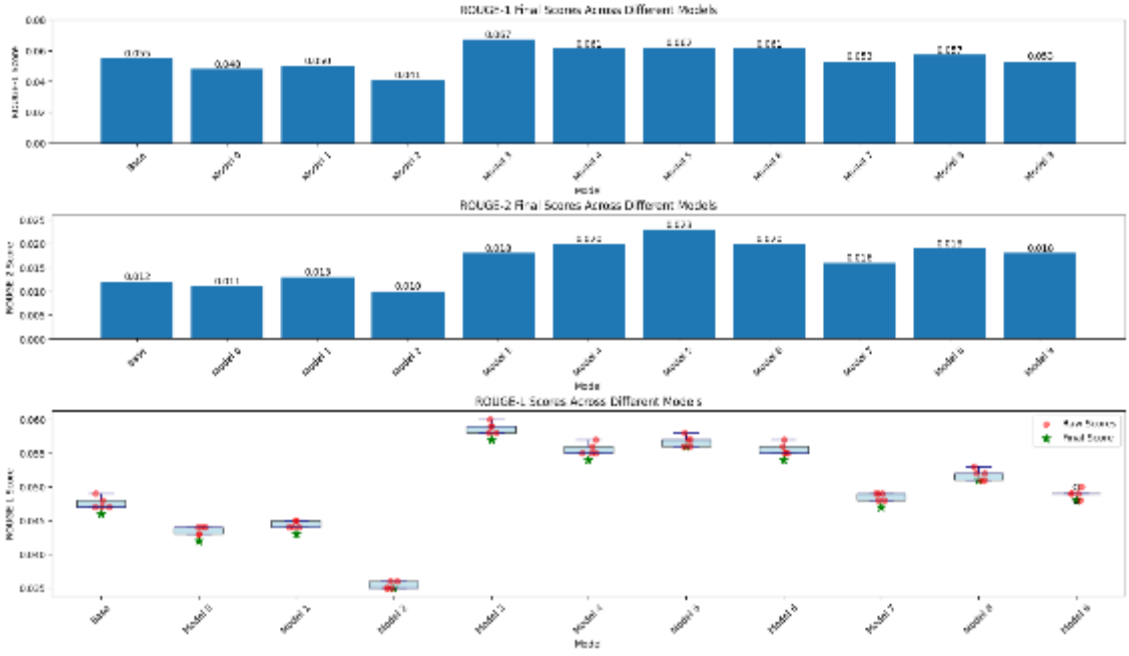
Pada Gambar 5, evaluasi BLEU iterasi 6 memiliki peningkatan performa sebesar $\pm$112% dengan standar deviasi relatif kecil, 3.72, dibanding iterasi lain seperti iterasi 4, 10.18, dan 8, 14.53. Ini menunjukkan bahwa augmentasi pada skala n-gram pada model berhasil meningkatkan korelasi terhadap domain yang baru, yakni bahasa Indonesia.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508



Gambar 5 Box plot perbandingan skor BLEU tiap iterasi dan baseline

Hasil uji ROUGE pada Gambar 5 menunjukkan bahwa iterasi 6 memiliki performa yang relatif kompetitif dibandingkan iterasi lainnya, dengan peningkatan ±21.7% ROUGE-L dengan standar deviasi rendah berbanding dengan iterasi lain seperti 4 dan 7, menunjukkan bahwa model dapat dengan lebih baik melakukan inferensi relevan terhadap konteks berbahasa Indonesia.

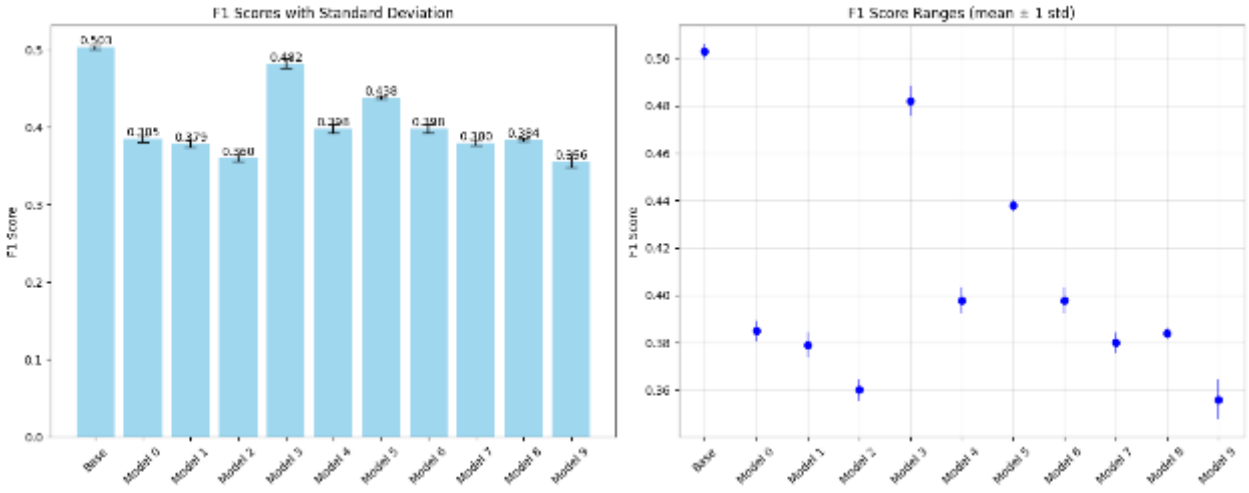


Gambar 6 Bar plot dan box plot perbandingan skor ROUGE-1, ROUGE-2, dan ROUGE-L tiap iterasi dan baseline

Weighted F1 pada Gambar 7 menunjukkan bahwa kemiripan semantik terhadap input cenderung menurun setelah *finetuning* secara keseluruhan. Interpretasi dalam penelitian adalah bahwa ini menunjukkan terjadinya *domain shift* yang menyebabkan *forgetting* pada beberapa

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

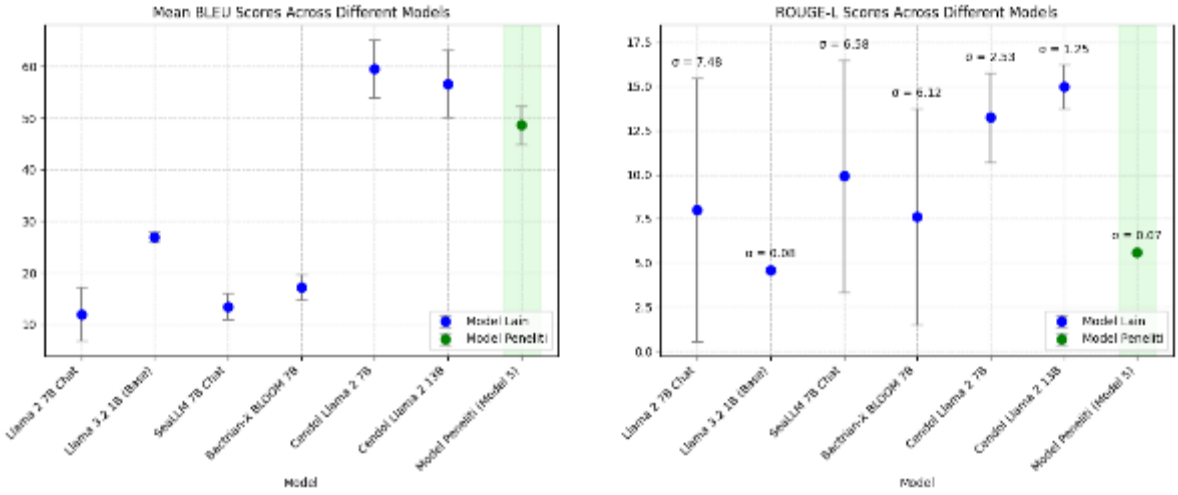
domain original. Pertimbangan pengambilan *best performer* adalah pengambilan iterasi dengan penurunan yang cukup rendah dan stabilitas performa pada metrik lain, yakni iterasi 6.



Gambar 7 Bar plot dan scatter plot perbandingan skor Weighted F1 tiap iterasi dan baseline

Pada gambar 8, iterasi 6 menunjukkan peningkatan skor BLEU terhadap *baseline*, serta peningkatan yang relatif substansial ketika dibandingkan dengan model dengan ukuran yang jauh lebih besar dalam penelitian lain, seperti ±260% pada SeaLLM 7B [14] dan ±182% pada Bactrian-X BLOOM 7B [15], sedangkan pada ROUGE-L, terlihat bahwa model lain memiliki instabilitas pada performa yang terlihat pada standar deviasi yang relatif tinggi dibanding dengan model, di mana *lower-bound* skor pada model meningkat sebesar ±197% terhadap SeaLLM 7B [14] dan ±561.3% terhadap Bactrian-X BLOOM 7B [15].

Hal ini menunjukkan bahwa prosedur pelatihan berhasil meningkatkan performa secara substansial, bahkan pada model pada ukuran yang relatif kecil seperti 1 miliar parameter, yang merupakan 14.2% dan 7.6% dari jumlah parameter yang digunakan pada model-model pembanding.



Gambar 8 Scatter plot perbandingan skor model peneliti terhadap model lain pada metrik BLEU dan ROUGE-L

Hasil jangka waktu metodologi pelatihan akhir mengikuti prinsip Chinchilla-optimal, di mana dengan waktu pelatihan yang lebih lama dengan data yang lebih banyak, sebuah model dapat berkonvergensi dengan lebih baik, bahkan tanpa tambahan parameter yang berlebih. [13]

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

Penemuan peningkatan performa yang berkorelasi dengan kurasi data sampel yang baik juga sesuai dengan riset yang menunjukkan bahwa dengan dilakukannya suplementasi data yang cukup beragam, model dapat menggeneralisasi dengan lebih cepat dan mencapai efisiensi pareto yang optimal. [16, 17]

4. KESIMPULAN DAN SARAN

PEFT memiliki potensi yang tinggi dalam augmentasi kekurangan domain model LLM pada pengetahuan berbahasa *low-resource* secara efisien pada komputasi *resource-constrained*. Akan tetapi, diperlukan preparasi *hyperparameter* serta preparasi dataset yang baik agar PEFT dapat meng-augmentasi kekurangan tersebut, dan bukan justru memperburuk kemampuan model untuk berbahasa. Di sini peneliti menemukan bahwa penekanan performa terdapat pada dua faktor utama, yakni jangka waktu pelatihan serta variabilitas sampel yang digunakan dalam pelatihan.

Penerapan metode PEFT LoRA+ pada model menunjukkan sebuah peningkatan yang substansial pada metrik BLEU dan ROUGE, bahkan meningkatkan performa model berbanding terhadap model-model dengan ukuran yang lebih besar, walaupun *domain shift* menyebabkan penurunan kemampuan *general knowledge* dari model. Dari hasil tersebut, dapat disimpulkan bahwa metode LoRA+ memiliki potensi signifikan dalam meningkatkan kemampuan berbahasa sebuah model pada sebuah bahasa baru dengan efisiensi komputasi yang cukup baik.

Peneliti menyarankan penelitian lanjutan untuk melakukan eksplorasi lebih luas dari domain *hyperparameter* yang dapat digunakan, serta eksplorasi penggunaan *sampling* yang lebih baik dalam mendapatkan kurasi sampel yang digunakan dalam pelatihan, sebab kedua faktor tersebut terlihat sebagai tren *emerging* yang muncul dari hasil penelitian. Selain itu, dapat dilakukan ablasi pada ukuran model yang digunakan serta proporsi adapter terhadap ukuran model.

DAFTAR PUSTAKA

[1] Khandelwal, K., Tonneau, M., Bean, A. M., Kirk, H. R., Hale, S. A., Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models, *Proceedings of the 2024 International Conference on Information Technology for Social Good*, Bremen, Germany, 2024, hal 231-239.

[2] Sharma, S., dalam Chen, Z., Zhang, J. M., Hort, M., Harman, M., Sarro, F., 2024, Fairness testing: A comprehensive survey and analysis of trends, *ACM Transactions on Software Engineering and Methodology*, vol 33, No. 5, hal 1-59.

[3] Sheng, E., Chang, K. W., Natarajan, P., Peng, N., Societal Biases in Language Generation: Progress and Challenges, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, 2021, hal 4275-4293.

[4] Cahyawijaya, S., Lovenia, H., Koto, F., Putri, R., Cenggoro, W., Lee, J., Akbar, S., Dave, E., Nuurshadieq, N., Mahendra, M., Putri, R., Wilie, B., Winata, G., Aji, A., Purwarianti, A., Fung, P., 2024, Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, hal 14899-14914, Association for Computational Linguistics, Bangkok, Thailand.

[5] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., A Comprehensive Survey on Transfer Learning, *Proceedings of the IEEE*, vol. 109, no. 1, hal 43-76, 2021.

[6] Rae, J. W. et al., dalam Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov,

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508

A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N., 2023, PaLM: Scaling Language Modeling with Pathways, *Journal of Machine Learning Research*, No. 240, Vol. 24, hal 1-113.

[7] Kaplan, J. et al., dalam Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., Dollar, P., Girshick, R., 2023, Segment Anything, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oktober, hal 4015-4026.

[8] Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S. Q., 2024, Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey, *Transactions on Machine Learning Research (TMLR)*, Journal of Machine Learning Research.

[9] Hu, E. J. et al., dalam Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., Sun, M., 2023, Parameter-efficient fine-tuning of large-scale pre-trained language models, *Nature Machine Intelligence*, vol 5, No. 3, hal 220-235.

[10] Hayou, S. et al., dalam Shao, M., Basit, A., Karri, R., Shafique, M., 2024, Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges, *IEEE Access*, vol 12, hal 188664-188706.

[11] Owen, L. et al., dalam Koto, F., Mahendra, R., Aisyah, N., Baldwin, T., 2024, IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces, *Transactions of the Association for Computational Linguistics*, Vol. 12, hal 1703-1719.

[12] Dubey, A. et al., dalam Dominguez-Olmedo, R., Hardt, M., Mendler-Dünner, C., 2024, Questioning the Survey Responses of Large Language Models, *Advances in Neural Information Processing Systems*, Vol. 37, Curran Associates, Inc., hal 45850-45878.

[13] Hoffmann, J. et al., dalam Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., Wei, J., 2024, Scaling Instruction-Finetuned Language Models, *Journal of Machine Learning Research*, No. 70, Vol. 25, hal 1-53.

[14] Nguyen, X. P., Zhang, W., Li, X., Aljunied, M., Hu, Z., Shen, C., Chia, Y. K., Li, X., Wang, J., Tan, Q., Cheng, L., Chen, G., Deng, Y., Yang, S., Liu, C., Zhang, H., Bing, L., SeaLLMs - Large Language Models for Southeast Asia, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024, hal 294-304.

[15] Koto, F. et al., dalam Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., Yu, P. S., A Survey of Multilingual Large Language Models, *Patterns*, Vol. 6(1), 2025, Elsevier.

[16] Miranda, B., Lee, A., Sundar, S., Casasola, A., Koyejo, S., Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data, *Proceedings of the Data-centric Machine Learning Research Workshop at ICLR 2024*, Vienna, Austria, 2024.

[17] Sorscher, B. et al., dalam Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., Raffel, C. A., 2023, Scaling Data-Constrained Language Models, *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Vol. 36, hal 50358-50376.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI :10.62411/tc.v24i2.12508