Perbandingan *LoRA* pada *Qwen2.5* dan *DeepSeek-R1* dengan *Fine-Tuning IndoBERT* untuk Sentimen Indonesia

Comparing LoRA on Qwen2.5 and DeepSeek-R1 with IndoBERT Fine-Tuning for Indonesian Sentiment

[1]Afito Indra Permana*, [2]Muhammad Iqbal Akbar
[1,2]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Malang
[1,2]Jl. Semarang No. 5, Sumbersari, Kec. Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia
*e-mail: indraafito56@gmail.com

(***received*:** 11 October 2026, ***revised*:** 9 November 2026, ***accepted*:** 10 November 2026)

Abstrak

Analisis sentimen berbahasa Indonesia banyak mengandalkan model *encoder* seperti IndoBERT melalui *full fine-tuning*, sementara large language model (LLM) generatif *open-weight* dengan Low-Rank Adaptation (LoRA) menawarkan alternatif yang lebih hemat sumber daya. Penelitian ini membandingkan Qwen2.5-1.5B-Instruct dan DeepSeek-R1-Distill-Qwen-1.5B dengan LoRA (hyperparameter adaptasi disamakan) terhadap IndoBERT-base-p1 dengan *full fine-tuning* pada analisis sentimen tiga kelas *dataset* SmSA dari IndoNLU, dengan pembagian ulang *stratified* 70:15:15. Berdasarkan tiga *seed* acak (42, 43, 44) pada 1.914 data uji, F1-macro rata-rata IndoBERT 0,9242±0,0048, Qwen 0,9195±0,0039, dan DeepSeek 0,8779±0,0048 (Qwen pada *run* awal dengan konfigurasi pelatihan berbeda sebesar 0,9211±0,0023). DeepSeek tertinggal signifikan dari keduanya pada ketiga *seed* (McNemar dengan koreksi Holm dan paired bootstrap), sedangkan selisih Qwen dan IndoBERT kecil (sekitar 0,5 poin) dan interval kepercayaannya mencakup nol, sehingga IndoBERT hanya cenderung sedikit lebih tinggi dan kesebandingan hanya dapat dinyatakan pada toleransi ±0,02 F1-macro. Backbone kedua LLM tidak identik. Kontrol tambahan satu *seed* dengan Qwen2.5-Math-1.5B ber-LoRA (F1-macro 0,8842) tidak berbeda signifikan dari DeepSeek, sehingga ketertinggalan DeepSeek diduga lebih terkait model dasar daripada distilasi *reasoning*. Pra-pemrosesan minimal tidak memberi perbedaan signifikan pada IndoBERT. IndoBERT juga jauh lebih efisien, dengan inferensi 6,12 ms per sampel dibandingkan 154,81 hingga 160,80 ms pada kedua model LoRA (*seed* 42). Temuan ini menunjukkan bahwa LoRA pada LLM generatif tidak otomatis mengungguli *encoder* dengan *full fine-tuning* pada klasifikasi sentimen dalam-distribusi seperti SmSA.
Kata kunci: analisis sentimen, IndoBERT, LoRA, parameter-efficient *fine-tuning*, SmSA

Abstract

Indonesian sentiment analysis has largely relied on encoder-based models, such as IndoBERT, through full fine-tuning. Meanwhile, open-weight generative large language models (LLMs) adapted using Low-Rank Adaptation (LoRA) offer a potentially more resource-efficient alternative. This study compares Qwen2.5-1.5B-Instruct and DeepSeek-R1-Distill-Qwen-1.5B using LoRA, with identical adaptation hyperparameters, against IndoBERT-base-p1 using full fine-tuning for three-class sentiment classification on the SmSA dataset from IndoNLU, with a re-split stratified into 70:15:15 training, validation, and test sets. Across three random seeds (42, 43, and 44), evaluated on 1,914 test samples, the mean macro-F1 scores were 0.9242 ± 0.0048 for IndoBERT, 0.9195 ± 0.0039 for Qwen, and 0.8779 ± 0.0048 for DeepSeek. In an earlier run using a different training configuration, Qwen achieved 0.9211 ± 0.0023. DeepSeek performed significantly worse than both competing models across all three seeds, based on McNemar's test with Holm correction and paired bootstrap analysis. In contrast, the performance gap between Qwen and IndoBERT was small (approximately 0.5 percentage points), and the confidence intervals included zero. Thus, IndoBERT showed only a slight tendency toward better performance, while comparability between the models can only be claimed within a predefined equivalence margin of ±0.02 macro-F1. The backbones of the two LLMs were not identical. An additional single-seed control experiment using LoRA-adapted Qwen2.5-Math-1.5B yielded a macro-F1 score of 0.8842, which was not significantly different from DeepSeek. This

 ISSN:2302-8149
e-ISSN:2540-9719

suggests that DeepSeek's performance gap may be more closely associated with its base model than with reasoning distillation. Minimal preprocessing did not produce a statistically significant difference in IndoBERT's performance. IndoBERT was also substantially more efficient at inference, requiring 6.12 ms per sample, compared with 154.81–160.80 ms for the two LoRA-based models (seed 42). These findings indicate that applying LoRA to generative LLMs does not automatically outperform full fine-tuning of encoder-based models in in-distribution sentiment classification tasks such as SmSA.
Keywords: *sentiment analysis, IndoBERT, LoRA, parameter-efficient fine-tuning, SmSA*

1 Pendahuluan

Analisis sentimen adalah salah satu tugas klasifikasi teks paling banyak diaplikasikan dalam pemrosesan bahasa alami (NLP), dari pemantauan opini publik di media sosial [1] hingga evaluasi kepuasan pelanggan pada platform ulasan daring [2]. Untuk Bahasa Indonesia, tugas ini menghadapi tantangan sebagai bahasa *mid-resource*, dengan korpus berlabel relatif terbatas dibandingkan Bahasa Inggris, dan data sentimen dunia nyata umumnya timpang antarkelas (*class-imbalanced*), dengan kelas netral biasanya paling sedikit datanya. Kelas netral juga termasuk yang paling sulit diklasifikasikan, bahkan oleh anotator manusia [3], seperti yang terlihat pada *dataset* SmSA yang menjadi acuan utama riset sentimen Indonesia [4].

Selama ini, model *encoder* berbasis Transformer yang dilatih khusus untuk Bahasa Indonesia, terutama IndoBERT (diperkenalkan bersama *benchmark* IndoNLU [4]) dan varian media sosialnya, IndoBERTweet [5], menjadi rujukan utama untuk tugas NLU Indonesia termasuk analisis sentimen, umumnya melalui *full fine-tuning* yang layak karena ukurannya relatif kecil (sekitar 124-125 juta parameter). Penelitian terbaru masih memakai IndoBERT sebagai basis, misalnya untuk pengenalan emosi ulasan produk [6], klasifikasi komentar publik [1], dan deteksi spam [7]. Di sisi lain, LLM generatif *open-weight* seperti Qwen2.5 [8] dan turunan distilasi *reasoning* seperti DeepSeek-R1-Distill-Qwen [9] membuka peluang adaptasi melalui *parameter-efficient fine-tuning* (PEFT) [10], khususnya Low-Rank Adaptation (LoRA) [11], yang hanya melatih sebagian kecil parameter tambahan. Pada klasifikasi teks pendek, pendekatan ini memangkas parameter terlatih lebih dari 98% dan memori GPU puncak 23-52% dengan performa mendekati *full fine-tuning* [12].

Ada satu pertanyaan yang belum banyak dijawab secara empiris untuk konteks Bahasa Indonesia, yaitu apakah LLM generatif 1,5 miliar parameter yang di-fine-tuning dengan LoRA dapat menyaingi *encoder* kecil yang di-fine-tuning penuh pada klasifikasi sentimen standar, dan seberapa besar *trade-off* efisiensinya. Untuk Bahasa Inggris, Roumeliotis dkk. [13] menemukan LLaMA-2 7B yang di-fine-tuning hanya unggul tipis atas BERT pada sentimen ulasan maskapai, dengan biaya *fine-tuning* sekitar sembilan belas kali lipat. Pertanyaan ini menjadi lebih menarik ketika dua LLM yang dibandingkan berasal dari keluarga arsitektur dan jumlah parameter yang sama, yaitu Qwen2.5-1.5B-Instruct (*instruction-tuned*) dan DeepSeek-R1-Distill-Qwen-1.5B (distilasi *reasoning* dari DeepSeek-R1). Keduanya tampak berbagi backbone, tetapi verifikasi kartu model Hugging Face menunjukkan backbone-nya tidak identik (Bagian 3.3 dan 4.8).

Kerangka ini terinspirasi dari Jin dkk. [14] dan diadaptasi ke SmSA dengan model 1,5 miliar parameter yang dapat dilatih pada GPU kelas konsumen (Google Colab, T4 16GB), dengan IndoBERT-base-p1 *full fine-tuning* sebagai *baseline*.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan (1) membandingkan efektivitas (accuracy, precision, recall, F1-macro) kedua LLM ber-LoRA dengan hyperparameter yang disamakan terhadap IndoBERT-base-p1 *full fine-tuning* pada sentimen tiga kelas SmSA, (2) menganalisis *trade-off* efisiensi (waktu latih, waktu inferensi, proporsi parameter terlatih), dan (3) menguji signifikansi selisih performa dengan McNemar, paired bootstrap, dan tiga *seed* acak. Kontribusinya adalah bukti empiris untuk Bahasa Indonesia pada skala model dan sumber daya yang realistis bagi peneliti dengan GPU terbatas, dengan catatan bahwa backbone kedua LLM tidak identik, yang ditelaah melalui tiga eksperimen tambahan (Bagian 4.6).

2 Tinjauan Literatur

Analisis sentimen bertujuan mengklasifikasikan polaritas opini dalam teks, umumnya ke kategori positif, negatif, dan netral. Untuk Bahasa Indonesia, *dataset* SmSA pada *benchmark* IndoNLU [4]

 ISSN:2302-8149
e-ISSN:2540-9719

menjadi salah satu acuan paling banyak digunakan, dengan 12.760 teks dari berbagai platform daring yang dianotasi ahli bahasa Indonesia. Distribusi kelasnya tidak seimbang, dengan positif mendominasi dan netral minoritas, sehingga F1-macro lebih relevan daripada akurasi.

IndoBERT merupakan model bahasa berbasis BERT [15] yang dilatih pada korpus Indo4B (sekitar 4 miliar kata) dan diperkenalkan bersama IndoNLU [4]. Bersama varian IndoBERTweet [5], model ini konsisten unggul dibandingkan model multibahasa seperti mBERT pada tugas NLU Bahasa Indonesia, termasuk analisis sentimen, dan dengan sekitar 124-125 juta parameter full fine-tuning-nya masih layak pada GPU kelas konsumen.

Seiring bertambahnya ukuran LLM, *full fine-tuning* menjadi mahal, sehingga berkembang *parameter-efficient fine-tuning* (PEFT). LoRA [11] membekukan bobot *pretrained* dan menyisipkan matriks dekomposisi rank rendah pada lapisan tertentu sehingga parameter terlatih jauh berkurang dengan performa tetap kompetitif.

Qwen2.5 [8] adalah keluarga model bahasa *open-weight* Alibaba (0,5-72 miliar parameter, dilatih pada 18 triliun *token*) dengan performa kompetitif pada berbagai *benchmark*. DeepSeek-R1-Distill-Qwen-1.5B adalah hasil distilasi *reasoning* dari DeepSeek-R1 [9]; menurut dokumentasi resmi DeepSeek-AI, model ini didistilasi di atas Qwen2.5-Math-1.5B, bukan Qwen2.5-1.5B yang menjadi basis Qwen2.5-1.5B-Instruct. Keduanya berada pada keluarga arsitektur Qwen2 berukuran 1,5 miliar parameter, tetapi *checkpoint* dasar dan beberapa konfigurasi arsitekturnya berbeda sehingga backbone-nya tidak identik (Bagian 3.3 dan 4.8).

Sejumlah penelitian terbaru mengeksplorasi LoRA pada analisis sentimen dan klasifikasi teks Bahasa Indonesia. Suharjanto dan Mustakim [6] menerapkan LoRA pada IndoBERT untuk pengenalan emosi ulasan produk (Tabel 1), menunjukkan PEFT dapat mengurangi kebutuhan komputasi tanpa mengorbankan performa signifikan dibandingkan *full fine-tuning* pada *encoder* yang sama. Nasution dkk. [3] membandingkan 22 LLM open-source dan ChatGPT-4 pada sentimen dan emosi Twitter Indonesia secara *zero-shot* (Tabel 1), dengan kelas netral tetap paling sulit diklasifikasikan bahkan bagi anotator manusia.

Pada tataran penerapan, riset sentimen Bahasa Indonesia bergerak dari pembelajaran mesin klasik ke arsitektur Transformer. Lin dan Nuha [2] mengusulkan model hibrida BERT-DistilBERT dengan LSTM-TCN pada *dataset* sentimen Indonesia campuran sembilan topik dan melaporkan model itu mengungguli skema pembanding berbasis transformer (Tabel 1). Kusoema dan Ibrahim [1] menggabungkan IndoBERT dengan RCNN untuk komentar YouTube sosial-politik (Tabel 1) dengan performa lebih lemah pada kelas netral dan positif akibat tumpang tindih semantik. Al Simabua dan Alfat [7] mengoptimasi *fine-tuning* IndoBERT-Lite pada deteksi spam melalui deduplikasi ekstrem dan *early stopping* untuk menekan kebocoran data. Ketiga studi itu memperlihatkan *encoder* berbahasa Indonesia masih menjadi tulang punggung praktik saat ini, tetapi belum menguji alternatif LLM generatif yang diadaptasi secara efisien.

Di luar Bahasa Indonesia, Roumeliotis dkk. [13] membandingkan LLaMA-2 7B dengan BERT pada sentimen ulasan maskapai (Tabel 1) dan menemukan keunggulan LLaMA-2 hanya tipis dengan biaya jauh lebih mahal. Wang dan Azman [12] menunjukkan LoRA pada *encoder* untuk klasifikasi teks pendek memangkas parameter terlatih drastis dengan performa mendekati *fine-tuning* penuh. Keduanya berbahasa Inggris, tidak membandingkan dua LLM berukuran sama, dan belum menguji signifikansi perbedaan secara berpasangan.

Rujukan kerangka utama penelitian ini adalah Jin dkk. [14], yang membandingkan *encoder* Transformer berbahasa Mandarin dengan Qwen2.5-7B dan DeepSeek-R1-Distill-Qwen-7B berbasis LoRA untuk deteksi teks hasil AI. Studi itu menemukan *encoder* cenderung menghafal data latih namun menurun signifikan pada distribusi berbeda (*out-of-distribution*), sementara Qwen2.5-7B dengan LoRA menggeneralisasi jauh lebih baik. Temuan ini mendasari pertanyaan apakah pola serupa berlaku pada analisis sentimen Indonesia yang relatif homogen dan dalam-distribusi, dibandingkan deteksi teks AI yang inheren rentan pergeseran distribusi antar-generator teks.

Tabel 1 merangkum penelitian terdahulu yang paling relevan. Jin dkk. [14] disertakan karena arsitektur dan rancangan eksperimennya paling dekat dengan penelitian ini.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1. Perbandingan penelitian terdahulu dan posisi penelitian ini.

No	Peneliti	Metode	Domain	Temuan Utama
1	Kusoema dan Ibrahim [1]	IndoBERT + RCNN	Sentimen	Akurasi 84%, F1 83%, performa lebih lemah pada kelas netral dan positif
2	Lin dan Nuha [2]	BERT-DistilBERT + LSTM-TCN (hibrida)	Sentimen	Model hibrida BERT-DistilBERT dan LSTM-TCN unggul atas pembanding transformer
3	Nasution dkk. [3]	22 LLM dan ChatGPT-4 (zero-shot)	Sentimen-emosi	LLM open-source terbaik capai >90% performa ChatGPT-4
4	Suharjanto dan Mustakim [6]	LoRA pada IndoBERT	Emosi	PEFT turunkan komputasi tanpa kurangi performa
5	Saputra dkk. [16]	IndoBERT Large context-conditioned	Sentimen	F1-macro 0,856, unggul 35,6 poin dari baseline
6	Roumeliotis dkk. [13]	LLaMA-2 7B + QLoRA vs BERT-base	Sentimen (Inggris)	Keunggulan LLaMA-2 tipis, biaya jauh lebih mahal
7	Jin dkk. [14]	Encoder vs LLM generatif 7B + LoRA	Deteksi teks AI (Mandarin)	LLM dengan LoRA lebih tahan pergeseran distribusi
8	**Penelitian ini**	**LLM generatif 1,5B + LoRA vs IndoBERT (full FT)**	**Sentimen (SmSA)**	**Ukuran dan hyperparameter LoRA disamakan (backbone tidak identik), diuji signifikansi statistik**

Berdasarkan uraian di atas, ada tiga hal yang sejauh penelusuran kami belum banyak dijawab literatur. Pertama, batas keberlakuan temuan Jin dkk. [14] belum diuji pada klasifikasi sentimen *in-distribution* seperti SmSA. Kedua, efek post-training belum banyak diisolasi untuk Bahasa Indonesia pada LLM berukuran dan konfigurasi LoRA yang disamakan, meskipun backbone pralatihannya tidak selalu identik (Bagian 3.3). Ketiga, *trade-off* efisiensi antara LoRA pada LLM 1,5 miliar parameter dan *full fine-tuning* pada *encoder* 124 juta parameter belum dikuantifikasi secara eksplisit untuk GPU kelas konsumen.

Berdasarkan celah tersebut, dirumuskan tiga hipotesis. H1, Qwen2.5-1.5B-Instruct dengan LoRA mencapai performa sebanding atau lebih baik daripada IndoBERT-base-p1 *full fine-tuning*, mengingat kapasitas representasinya yang jauh lebih besar. H2, di antara dua LLM berukuran dan konfigurasi LoRA yang disamakan, model *instruction-tuned* (Qwen2.5-1.5B-Instruct) mengungguli model *reasoning-distilled* (DeepSeek-R1-Distill-Qwen-1.5B), karena tugas sentimen lebih menuntut pemahaman ekspresi sehari-hari daripada rantai penalaran panjang. H3, IndoBERT-base-p1 tetap unggul secara efisiensi komputasi meski seluruh parameternya diperbarui, karena ukurannya jauh lebih kecil. Hipotesis diuji dengan paired bootstrap pada selisih F1-macro, McNemar pada akurasi dengan koreksi Holm-Bonferroni, dan replikasi tiga *seed*.

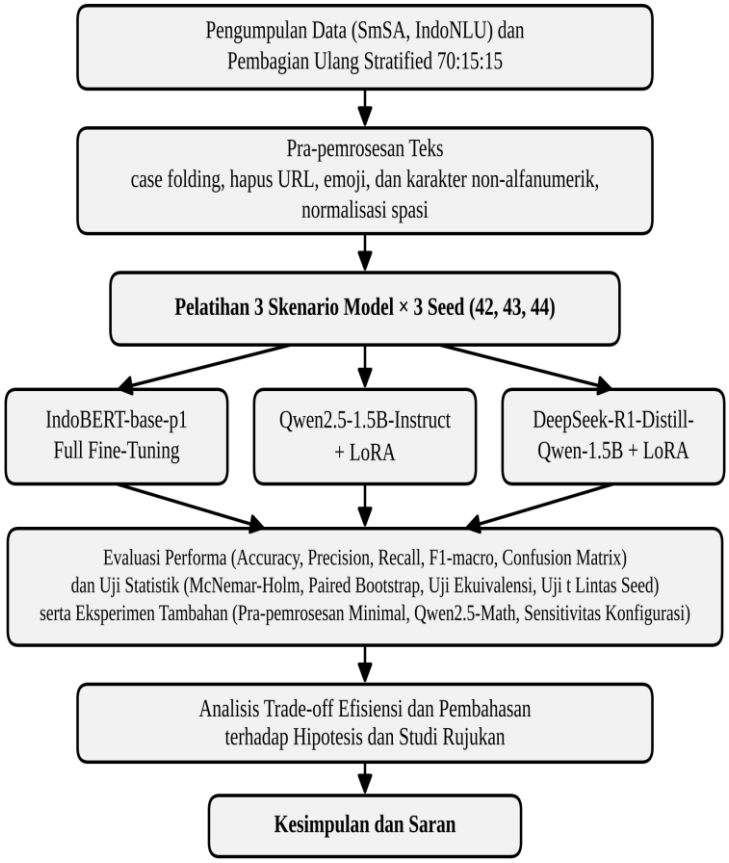
3 Metode Penelitian

3.1 Desain Penelitian

Penelitian ini menggunakan desain eksperimen kuantitatif komparatif, membandingkan tiga skenario *fine-tuning* pada klasifikasi sentimen tiga kelas, yaitu (1) Qwen2.5-1.5B-Instruct dengan LoRA, (2) DeepSeek-R1-Distill-Qwen-1.5B dengan hyperparameter LoRA yang sama, dan (3) IndoBERT-base-p1 *full fine-tuning* sebagai *baseline*. Seluruh eksperimen dijalankan di Google Colab (GPU NVIDIA T4, 16GB VRAM), dengan *run* utama memakai *seed* 42 di seluruh pustaka (Python random, NumPy, PyTorch, Hugging Face Transformers); dua *seed* tambahan (43 dan 44) dijalankan untuk replikasi (Bagian 4.5). Tiga eksperimen tambahan menguji keterbatasan rancangan ini (Bagian 4.6), yaitu (A) IndoBERT dengan pra-pemrosesan minimal (huruf kecil, penghapusan URL, dan normalisasi spasi, tanpa penghapusan tanda baca) pada tiga *seed*, (B) LoRA pada Qwen2.5-Math-1.5B, model dasar DeepSeek, pada *seed* 42, dan (C) pengulangan Qwen *seed* 42 dengan konfigurasi

 ISSN:2302-8149
e-ISSN:2540-9719

hemat memori yang sama dengan lima *run* LoRA lainnya, yang dipakai sebagai *run* utama, sedangkan *run* awal dipertahankan hanya sebagai analisis sensitivitas. Alur penelitian disajikan pada Gambar 1.



Gambar 1. Alur penelitian.

3.2 Dataset dan Pra-pemrosesan

Dataset yang digunakan adalah SmSA dari *benchmark* IndoNLU, dimuat melalui Hugging Face datasets (fallback ke TSV resmi repositori GitHub IndoBenchmark bila *loader* utama gagal). *Split* resmi bawaan berjumlah 11.000 data latih, 1.260 validasi, 500 uji (rasio sekitar 86:10:4) [4]. *Split* ini sendiri bukan pembagian asli sumber data, melainkan dibuat oleh tim IndoNLU khusus untuk *benchmark*, sehingga tidak ada satu pembagian "asli" yang mutlak dipertahankan. Seluruh *split* digabung menjadi satu pool 12.760 data, lalu dibagi ulang secara *stratified* 70:15:15 (random_state=42) menggunakan train_test_split scikit-learn, menghasilkan 8.932 data latih, 1.914 validasi, dan 1.914 uji.

Keputusan membagi ulang didasari dua pertimbangan. Pertama, *split* resmi menyediakan data uji kecil (500 data, sekitar 3,9% dari total) yang kurang memadai untuk pengujian statistik stabil, sedangkan rasio 70:15:15 menjaga data latih tetap besar dengan validasi dan uji yang cukup besar, sejalan dengan Joseph [17] bahwa rasio optimal bergantung pada kompleksitas model dan bukan angka baku (diturunkan untuk regresi linear). Kedua, validation *split* 15% *stratified* dengan *early stopping* (*patience* 2) berdasarkan F1-macro juga diterapkan Saputra dkk. [16] pada riset sentimen Indonesia terkini, dan skema *stratified* dipilih karena distribusi label SmSA tidak seimbang.

Tahap pra-pemrosesan teks mencakup lima langkah, yaitu (1) *case folding*, (2) menghapus URL, (3) menghapus emoji, (4) menghapus karakter non-alfanumerik, dan (5) menormalkan spasi berlebih. Data dengan teks kosong setelah pembersihan dibuang. Langkah dasar ini mengurangi noise teks media sosial atau ulasan daring, dan sebagian juga dipakai pada [2]. Seluruh 12.760 data tetap dipertahankan dengan distribusi label 7.359 positif (57,67%), 1.367 netral (10,71%), dan 4.034 negatif (31,61%).

 ISSN:2302-8149
e-ISSN:2540-9719

Pra-pemrosesan mengubah 87,81% data selain menurunkan huruf kapital, terutama karena pembuangan tanda baca. Tanda tanya muncul pada 8,60% teks negatif dibandingkan 1,36% positif dan 1,61% netral, sedangkan tanda seru muncul pada 9,25% teks negatif, 7,15% positif, dan 1,68% netral. Pembuangan tanda baca ini berpotensi menghilangkan sinyal sentimen yang berkorelasi dengan kelas, sehingga klaim bahwa pra-pemrosesan tidak memengaruhi informasi sentimen tidak dapat dipertahankan. Penghapusan emoji dan URL praktis tidak berdampak karena tidak ada teks yang memuat emoji, karakter non-ASCII, maupun URL. Seluruh model menerima teks hasil pembersihan yang sama agar perbandingan adil. Dampak pra-pemrosesan minimal diuji pada IndoBERT (Bagian 4.6), sedangkan kedua LLM tidak diulang (Bagian 4.8).

Karena data digabung lalu dibagi ulang acak, duplikasi antar-subset diperiksa. Pada teks hasil pembersihan, 15 dari 1.914 baris uji (0,78%) identik dengan baris data latih, dan 20 baris (1,04%) identik dengan data latih atau validasi. Analisis sensitivitas yang mengeluarkan 15 baris bocor (n = 1.899) hanya mengubah F1-macro setiap *run* paling banyak 0,0015 dan tidak mengubah kesimpulan signifikansi mana pun. Hasil utama tetap dilaporkan pada seluruh 1.914 data uji.

Hasil penelitian ini tidak dapat dibandingkan langsung dengan papan peringkat IndoNLU. Data uji resmi berjumlah 500 teks (41,6% positif, 17,6% netral, 40,8% negatif), sedangkan data uji di sini berjumlah 1.914 teks (57,68% positif, 10,71% netral, 31,61% negatif). Perbedaan ukuran dan komposisi kelas ditambah pembagian ulang dan pra-pemrosesan sendiri membuat skor di sini hanya sah dibandingkan antarmodel dalam penelitian ini.

3.3 Model dan Skema Fine-Tuning

Tabel 2 merangkum ketiga model yang dibandingkan beserta skema fine-tuning-nya. Qwen mewakili LLM instruction-tuned, DeepSeek mewakili LLM hasil distilasi reasoning, dan IndoBERT mewakili encoder konvensional.

Tabel 2. Model dan skema fine-tuning yang dibandingkan.

Model	Identifier Hugging Face	Metode	Keterangan
Qwen2.5-1.5B-Instruct	Qwen/Qwen2.5-1.5B-Instruct	LoRA	LLM instruction-tuned, ~1,56 miliar parameter
DeepSeek-R1-Distill-Qwen-1.5B	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	LoRA	Distilasi reasoning DeepSeek-R1 di atas Qwen2.5-Math-1.5B, ~1,56 miliar parameter (jumlah parameter sama dengan Qwen)
IndoBERT-base-p1	indobenchmark/indobert-base-p1	Full fine-tuning	Encoder BERT khusus Bahasa Indonesia, ~124-125 juta parameter (baseline pembanding)

Pada pemilihan model, IndoBERT-base-p1 menjadi *baseline encoder* karena paling banyak dirujuk pada IndoNLU [4], Qwen2.5-1.5B-Instruct mewakili LLM *instruction-tuned open-weight* [8], dan DeepSeek-R1-Distill-Qwen-1.5B adalah distilasi *reasoning* DeepSeek-R1 dengan ukuran sama [9]. Keduanya berkapasitas serupa tetapi backbone-nya tidak identik, sebagaimana diuraikan berikut.

Hyperparameter adaptasi LoRA disamakan untuk kedua model agar selisih performa tidak berasal dari konfigurasi adaptasi, dengan rank (r)=16, alpha=32, dropout=0,05, bias="none", target modules seluruh lapisan Linear pada blok attention dan feed-forward (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj). Kombinasi rank=16 dan alpha=32 (rasio 2) lazim dipakai dalam praktik LoRA; Hu dkk. [11] tidak menetapkan satu nilai rank baku dan menunjukkan rank kecil memadai pada banyak tugas, sehingga pemilihan ini praktis dan tidak diuji dengan ablasi.

Kartu model Hugging Face Qwen2.5-1.5B-Instruct mencantumkan base_model Qwen/Qwen2.5-1.5B, sedangkan kartu DeepSeek-R1-Distill-Qwen-1.5B mencantumkan Qwen2.5-Math-1.5B sebagai model dasar. Berkas konfigurasi keduanya juga berbeda pada eos_token_id (151645 dan 151643), max_position_embeddings (32.768 dan 131.072), rope_theta (1.000.000 dan 10.000), serta

 ISSN:2302-8149
e-ISSN:2540-9719

tie_word_embeddings (True dan False). Karena itu selisih performa kedua LLM tidak dapat diatribusikan murni pada objektif post-training, dan penelitian ini hanya menyatakan bahwa ukuran model serta hyperparameter LoRA keduanya disamakan, bukan bahwa backbone-nya identik. Qwen2.5-Math-1.5B ditambahkan sebagai kontrol model dasar (Bagian 4.6).

Seluruh model diadaptasi sebagai pengklasifikasi melalui AutoModelForSequenceClassification, bukan pembangkitan teks. Pada Qwen dan DeepSeek, lapisan linear baru bernama score dipasang di atas hidden state terakhir dan dilatih bersama *adapter* LoRA (modules_to_save). Logit kelas diambil dari hidden state pada *token* non-pad paling kanan (last-token pooling). Tokenizer Qwen memiliki pad *token* bawaan <|endoftext|> (id 151643) yang berbeda dari eos *token* <|im_end|> (id 151645), sedangkan DeepSeek memakai eos *token* yang sama sebagai pad *token* (id 151643). Padding di sisi kanan dengan panjang maksimum 128 *token*, dan 0,418% teks uji terpotong pada kedua LLM. IndoBERT-base-p1 memakai head klasifikasi standar BERT di atas *token* [CLS]. Karena prediksi berasal dari satu vektor logit, kemampuan penalaran berantai DeepSeek-R1-Distill tidak digunakan; penelitian ini hanya menilai kualitas representasi tersembunyi untuk klasifikasi, bukan *reasoning* atau penjelasan prediksi.

Dengan konfigurasi ini, parameter terlatih Qwen dan DeepSeek sekitar 18 juta dari total sekitar 1,56 miliar (1,18%), sedangkan IndoBERT-base-p1 di-fine-tuning penuh (sekitar 124-125 juta parameter, 100% *trainable*), pembanding konvensional yang tetap standar dan terjangkau untuk *encoder* kecil [15] dan juga dipakai Suharjanto dan Mustakim [6].

3.4 Hyperparameter Pelatihan

Hyperparameter pelatihan kedua model LoRA dibuat sama. Learning rate 2×10^{-4} (lazim lebih tinggi daripada *full fine-tuning* pada LoRA [11], [10]) dengan penjadwalan cosine dan warmup ratio 0,06, *epoch* maksimum 3 dengan *early stopping* (*patience*=2) berdasarkan F1-macro validasi, optimizer AdamW, dan *batch* efektif 32 (840 langkah optimizer untuk tiga *epoch*). Konfigurasi komputasi per langkah seragam pada keenam *run* LoRA utama (Qwen dan DeepSeek, *seed* 42, 43, 44), yaitu per-device *batch* 8 dengan gradient checkpointing untuk menghindari kendala memori GPU T4 16GB. Run Qwen *seed* 42 yang pertama memakai per-device *batch* 16 tanpa gradient checkpointing; *run* itu digantikan pengulangan pada konfigurasi hemat dan hanya dipertahankan sebagai analisis sensitivitas (Bagian 4.6), karena perbedaan konfigurasi memengaruhi interpretasi efisiensi (Bagian 4.3). Pada akhir pelatihan dimuat *checkpoint* dengan F1-macro validasi terbaik, dan *early stopping* tidak terpicu pada *run* mana pun sehingga seluruh *run* menyelesaikan tiga *epoch*.

Untuk IndoBERT-base-p1, digunakan *learning rate* 2×10^{-5} (lazim untuk *full fine-tuning* BERT [15]), jauh lebih kecil dari LoRA karena seluruh bobot *pretrained* diperbarui langsung, dengan *epoch* 3 dan *batch* 16. Max_length seluruh model diset 128 *token* karena mayoritas teks SmSA berupa kalimat atau paragraf pendek.

3.5 Metrik Evaluasi dan Pengujian Statistik

Efektivitas model dievaluasi dengan accuracy, precision, recall, dan F1-score macro-averaging, mengingat distribusi label SmSA yang tidak seimbang. Pemilihan F1-macro mengacu pada Opitz [18], yang menekankan bahwa metrik macro dipilih sadar ketika setiap kelas, termasuk minoritas seperti netral, dianggap sama penting, sejalan dengan Henning dkk. [19] yang merekomendasikan pelaporan metrik per kelas dan macro-averaged untuk data timpang. Efisiensi diukur melalui waktu latih (waktu dinding pelatihan, termasuk evaluasi validasi per *epoch*), memori GPU puncak (torch.cuda.max_memory_allocated selama pelatihan), waktu inferensi per sampel (waktu dinding prediksi 1.914 data uji dibagi jumlah data, *batch* evaluasi 16 pada LoRA hemat dan 32 pada IndoBERT, sehingga bukan latensi permintaan tunggal), dan proporsi parameter terlatih. Jumlah *epoch* (3) dan langkah optimizer aktual dicatat pada Tabel 4.

Pengujian signifikansi menggunakan tiga pendekatan yang menjawab pertanyaan berbeda. Pertama, uji McNemar dengan koreksi kontinuitas [20] membandingkan akurasi dua model pada data uji yang sama berdasarkan tabel kontingensi 2×2 dari ketepatan prediksi berpasangan (Tabel 5; α=0,05), dan tidak ditafsirkan sebagai uji terhadap F1-macro. Kedua, selisih F1-macro diuji dengan paired bootstrap, yaitu 5.000 resampling dengan pengembalian pada indeks data uji yang sama untuk kedua model [21], dengan CI 95% persentil untuk selisih berpasangan itu, bukan CI masing-masing

 ISSN:2302-8149
e-ISSN:2540-9719

model [22]. Untuk rata-rata tiga *seed*, selisih dihitung pada resampel indeks yang sama di ketiga *seed* lalu dirata-ratakan. Ketiga, konsistensi lintas *seed* diuji dengan paired t-test dan Wilcoxon signed-rank pada selisih F1-macro per *seed* (n=3). Nilai-p McNemar per *seed* dan nilai-p uji t lintas *seed* dikoreksi dengan Holm-Bonferroni atas tiga pasangan model, uji McNemar per kelas aktual atas seluruh keluarga 27 uji, dan eksperimen tambahan atas 6 uji. Hasil yang tidak signifikan tidak diperlakukan sebagai bukti kesetaraan, yang diperiksa dengan uji ekuivalensi (two one-sided tests), yaitu apakah CI 90% selisih berpasangan berada seluruhnya di dalam margin ±0,01 atau ±0,02 F1-macro. Kedua margin ini ditetapkan penulis sebagai toleransi praktis dan tidak berasal dari ketentuan baku di literatur. Prosedur yang sama dipakai pada eksperimen tambahan (Bagian 4.6).

4 Hasil dan Pembahasan

4.1 Performa Klasifikasi Keseluruhan

Tabel 3 merangkum accuracy, precision-macro, recall-macro, dan F1-macro ketiga model pada 1.914 data uji untuk *run* utama (*seed* 42). Hasil utama penelitian ini adalah rata-rata dan simpangan baku tiga *seed* pada Tabel 7 (Bagian 4.5), sedangkan Tabel 3 merinci *run* utama.

Tabel 3. Perbandingan metrik efektivitas ketiga model pada data uji (n = 1.914, run utama seed 42).

Model [Metode]	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
Qwen2.5-1.5B-Instruct [LoRA]	0,9373	0,9248	0,9083	0,9159
DeepSeek-R1-Distill-Qwen-1.5B [LoRA]	0,9122	0,8805	0,8723	0,8762
IndoBERT-base-p1 [Full fine-tuning]	0,9399	0,9312	0,9081	0,9189

IndoBERT-base-p1 dengan *full fine-tuning* mencapai F1-macro tertinggi pada *run* utama di Tabel 3 (0,9189), sedikit di atas Qwen2.5-1.5B-Instruct dengan LoRA (0,9159, selisih sekitar 0,3 poin persentase), sedangkan DeepSeek-R1-Distill-Qwen-1.5B yang berukuran dan ber-hyperparameter LoRA sama dengan Qwen tertinggal cukup jauh (0,8762, sekitar 4,0 poin di bawah Qwen dan 4,3 poin di bawah IndoBERT).

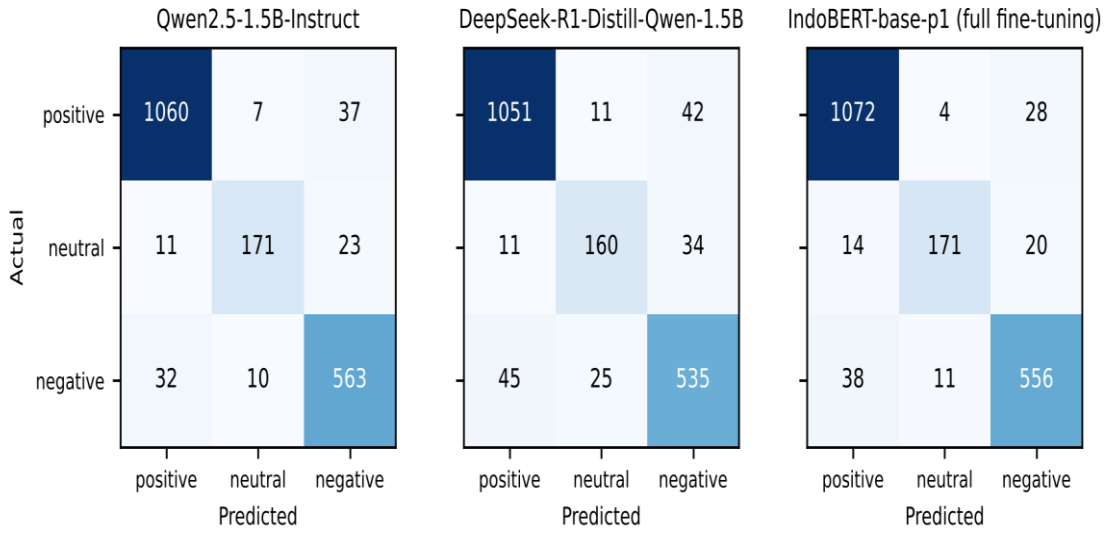
Kedua LLM memiliki jumlah parameter dan hyperparameter LoRA yang sama, tetapi backbone dasarnya tidak identik (Bagian 3.3), sehingga selisih itu dapat bersumber dari tujuan post-training, model dasar (Qwen2.5-Math-1.5B dibandingkan Qwen2.5-1.5B), atau keduanya; rancangan utama tidak dapat memisahkannya.

4.2 Analisis Kesalahan Klasifikasi per Kelas

Gambar 2 menyajikan *confusion matrix* ketiga model. Kelas netral, dengan data uji paling sedikit (205 dari 1.914), paling sering salah diklasifikasikan, dengan recall 0,8341 pada Qwen dan IndoBERT serta 0,7805 pada DeepSeek, konsisten dengan Nasution dkk. [3] bahwa netral paling sulit diklasifikasikan pada sentimen Indonesia bahkan bagi anotator manusia.

DeepSeek juga menunjukkan kesalahan lebih tinggi pada kelas negatif dibandingkan Qwen dan IndoBERT (45 dari 605 data negatif salah diklasifikasikan positif, berbanding 32 dan 38) pada *run* utama. Pola ini belum dapat disebut sistematis. Pada uji McNemar per kelas aktual (Qwen terhadap DeepSeek), DeepSeek lebih sering salah pada *seed* 42 (44 berbanding 16 kasus diskordan), 43 (41 berbanding 18), dan 44 (40 berbanding 13). Setelah koreksi Holm atas 27 uji, *seed* 42 (p=0,009) dan 44 (p=0,006) signifikan, sedangkan *seed* 43 (p=0,076) tidak, sehingga arahnya konsisten tetapi dukungan statistiknya belum.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 2. Confusion matrix ketiga model pada data uji (run utama seed 42).

4.3 Trade-off Efisiensi Komputasi

Tabel 4 merangkum konfigurasi komputasi, waktu latih, memori GPU puncak, dan waktu inferensi per sampel pada seluruh *run*. Parameter terlatih Qwen dan DeepSeek masing-masing 18.469.376 dari 1.562.188.288 (1,18%), sedangkan IndoBERT-base-p1 memperbarui seluruh 124.443.651 parameternya (100%).

Tabel 4. Konfigurasi komputasi, waktu latih, memori puncak, dan waktu inferensi pada seluruh run (termasuk run awal Qwen dan run tambahan B).

Model	Seed	Konfigurasi	Langkah	Waktu Latih (detik)	Memori Puncak (MB)	Infer. (ms)
Qwen	42	b8, GC	840	13.131,7	3.572,4	160,80
Qwen	43	b8, GC	840	13.948,7	3.574,4	168,69
Qwen	44	b8, GC	840	4.263,2	3.574,4	152,79
Qwen (awal)	42	b16, tanpa GC	840	9.176,3	11.434,7	142,76
DeepSeek	42	b8, GC	840	4.332,4	3.574,4	154,81
DeepSeek	43	b8, GC	840	8.276,3	3.574,4	144,03
DeepSeek	44	b8, GC	840	8.178,7	3.574,4	143,07
Math (B)	42	b8, GC	840	7.909,4	3.572,4	139,11
IndoBERT	42	full FT, b16	1677	582,3	2.599,0	6,12
IndoBERT	43	full FT, b16	1677	567,3	2.599,0	5,94
IndoBERT	44	full FT, b16	1677	573,1	2.599,0	5,83

Pola paling jelas pada Tabel 4 adalah keunggulan IndoBERT-base-p1 pada seluruh aspek efisiensi meski di-fine-tuning penuh. Pada *run* utama, waktu latihnya (582,3 detik) sekitar 7,4 kali lebih singkat dari DeepSeek dan 22,6 kali dari Qwen, dan inferensinya (6,12 ms per sampel) sekitar 25-26 kali lebih cepat dari kedua model LoRA (160,80 dan 154,81 ms). LoRA hanya mengurangi parameter yang diperbarui, sehingga *checkpoint* adapter-nya sekitar 304 MB berbanding sekitar 1.900 MB pada

 ISSN:2302-8149
e-ISSN:2540-9719

IndoBERT, tetapi komputasi forward-backward tetap melewati seluruh 1,56 miliar parameter dasar, sehingga efisiensinya bersifat memori dan penyimpanan, bukan latensi. Besarnya memori puncak sendiri sangat dipengaruhi pengaturan pelatihan [23].

Pada *run* utama, kedua model LoRA memakai konfigurasi komputasi yang sama (*batch* 8 dengan gradient checkpointing, GC; Bagian 3.4), sehingga memori puncaknya hampir identik (3.572,4 MB pada Qwen dan 3.574,4 MB pada DeepSeek). Run Qwen *seed* 42 yang pertama (Qwen (awal) di Tabel 4; *batch* 16 tanpa GC) membutuhkan 11.434,7 MB (sekitar 3,2 kali lipat) pada jumlah parameter, *epoch*, dan langkah optimizer (840) yang sama, sehingga memori puncak ditentukan oleh konfigurasi pelatihan, bukan model.

Dengan model dan konfigurasi hemat yang sama, waktu latih Qwen berkisar 4.263,2 hingga 13.948,7 detik pada tiga *seed* dan DeepSeek 4.332,4 hingga 8.276,3 detik, selisih hingga sekitar 3,3 kali. Waktu latih DeepSeek *seed* 42 (4.332,4 detik) hanya sekitar 33,0% dari Qwen *seed* 42 (13.131,7 detik), tetapi *run* awal Qwen tanpa GC (9.176,3 detik) berada di rentang yang sama, sehingga selisih waktu tidak dijelaskan oleh konfigurasi dan kemungkinan terkait kondisi sesi Colab (GPU T4 bersama) yang tidak dikontrol. Waktu latih Qwen-Math (7.909,4 detik) hanya mencakup 560 langkah terakhir (Bagian 4.6). Waktu dinding tidak dipakai untuk menyimpulkan perbedaan efisiensi Qwen dan DeepSeek. Waktu inferensi per sampel pada kedelapan *run* LoRA berkisar 139,1 hingga 168,7 ms tanpa urutan konsisten dan jauh di atas IndoBERT (5,8 hingga 6,1 ms).

4.4 Signifikansi Statistik

Signifikansi diuji pada tiga pasangan model dengan pendekatan Bagian 3.5. Tabel 5 memuat tabel kontingensi 2×2 ketepatan prediksi berpasangan pada *seed* 42 beserta McNemar pada akurasi (bukan F1-macro). DeepSeek tertinggal signifikan terhadap Qwen (benar sendiri pada 45 sampel berbanding 93) maupun IndoBERT (36 berbanding 89), sedangkan Qwen dan IndoBERT tidak berbeda signifikan (41 berbanding 46). Pola sama muncul pada *seed* 43 dan 44 (DeepSeek tertinggal dengan p Holm di bawah 0,001; Qwen dan IndoBERT p Holm 0,538 hingga 0,837).

Tabel 5. Tabel kontingensi 2×2 ketepatan prediksi berpasangan dan uji McNemar pada akurasi (seed 42, n = 1.914).

Pasangan (A vs B)	Keduanya benar	Hanya A benar	Hanya B benar	Keduanya salah	χ^2	p Holm
Qwen vs DeepSeek	1.701	93	45	75	16,01	$1{,}26\times10^{-4}$
Qwen vs IndoBERT	1.753	41	46	74	0,18	0,668
DeepSeek vs IndoBERT	1.710	36	89	79	21,63	$9{,}91\times10^{-6}$

Tabel 6 melaporkan selisih F1-macro berpasangan beserta CI-nya. Selisih Qwen terhadap DeepSeek berkisar +0,036 hingga +0,050 per *seed* dengan CI 95% selalu di atas nol, dan DeepSeek terhadap IndoBERT berkisar −0,042 hingga −0,054 dengan CI 95% selalu di bawah nol. Selisih Qwen terhadap IndoBERT negatif pada ketiga *seed* (−0,0030; −0,0064; −0,0048), tetapi CI 95% per *seed* dan CI rata-ratanya mencakup nol, yang hanya berarti perbedaan tidak terdeteksi pada ukuran data uji ini, bukan bukti bahwa kedua model setara. CI 90% selisih rata-rata tiga *seed* (−0,0136 hingga +0,0043) berada di dalam margin ±0,02 tetapi tidak ±0,01, sehingga kesetaraan Qwen dan IndoBERT hanya dapat dinyatakan pada toleransi ±0,02 yang ditetapkan penulis. Pasangan yang melibatkan DeepSeek tidak ekuivalen pada kedua margin.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 6. Selisih F1-macro berpasangan (paired bootstrap, 5.000 resampel) dan uji ekuivalensi pada rata-rata tiga seed.

Pasangan (A vs B)	Seed	ΔF1 (A−B)	CI 95%	CI 90%	Ekuiv. ±0,01	Ekuiv. ±0,02
Qwen vs DeepSeek	3 seed	0,0416	[0,0291; 0,0553]	[0,0310; 0,0533]	Tidak	Tidak
Qwen vs IndoBERT	3 seed	−0,0047	[−0,0152; 0,0056]	[−0,0136; 0,0043]	Tidak	Ya
DeepSeek vs IndoBERT	3 seed	−0,0463	[−0,0603; −0,0325]	[−0,0580; −0,0347]	Tidak	Tidak

4.5 Validasi Multi-Seed (Mean ± Std dan Uji Lintas-Seed)

Untuk menjawab keterbatasan single-run, eksperimen diulang dengan tiga *seed* (42, 43, 44) mengingat varians *fine-tuning* akibat *seed* acak [24], yang terdiri atas satu *run* utama (Bagian 4.1 hingga 4.4) dan dua *run* tambahan (*seed* 43 dan 44). Rata-rata ± simpangan baku pada Tabel 7 menjadi hasil utama. Seluruh *run* LoRA pada tiga baris pertama memakai konfigurasi komputasi yang sama (Bagian 3.4); *run* awal berkonfigurasi berbeda (baris Qwen awal) disajikan sebagai analisis sensitivitas (Bagian 4.6).

Tabel 7. Agregasi hasil multi-seed (mean ± std; tiga baris terakhir dari eksperimen tambahan, Bagian 4.6).

Model	N Seed	Accuracy (mean±std)	F1-macro (mean±std)	Precision (mean±std)	Recall (mean±std)
Qwen2.5-1.5B-Instruct [LoRA]	3	0,9392±0,0018	0,9195±0,0039	0,9262±0,0031	0,9137±0,0051
DeepSeek-R1-Distill-Qwen-1.5B [LoRA]	3	0,9121±0,0024	0,8779±0,0048	0,8868±0,0055	0,8703±0,0077
IndoBERT-base-p1 [Full fine-tuning]	3	0,9418±0,0025	0,9242±0,0048	0,9318±0,0033	0,9177±0,0084
Qwen awal [LoRA]	3	0,9410±0,0016	0,9211±0,0023	0,9281±0,0035	0,9149±0,0033
IndoBERT minimal [Full FT]	3	0,9441±0,0014	0,9297±0,0030	0,9382±0,0048	0,9230±0,0099
Qwen-Math [LoRA]	1	0,9154	0,8842	0,8846	0,8840

Setelah dirata-rata tiga *seed* (Tabel 7), IndoBERT-base-p1 mencatat F1-macro rata-rata tertinggi (0,9242±0,0048), sedikit di atas Qwen (0,9195±0,0039), urutan yang sama dengan *run* utama (Tabel 3). Selisih Qwen terhadap IndoBERT negatif pada ketiga *seed*, tetapi CI-nya mencakup nol (Tabel 6), sehingga tidak ada bukti perbedaan pada ukuran data uji ini; rentang mean±std yang tumpang tindih tidak dijadikan dasar kesimpulan. Dengan *run* awal Qwen, rata-rata Qwen 0,9211±0,0023 (Bagian

 ISSN:2302-8149
e-ISSN:2540-9719

4.6). DeepSeek tertinggal pada ketiga *seed* (0,8779±0,0048), dengan CI selisih terhadap kedua model selalu di bawah nol.

Uji t berpasangan pada selisih F1-macro lintas tiga *seed* sejalan dengan McNemar dan paired bootstrap untuk DeepSeek, yang signifikan lebih buruk dari Qwen (t=10,18; p=0,0095; p Holm=0,0217) maupun IndoBERT (t=−11,69; p=0,0072; p Holm=0,0217). Untuk Qwen dan IndoBERT, uji t memberi t=−4,86 dan p=0,0398, tetapi McNemar per *seed* dan CI paired bootstrap tidak signifikan dan uji t pada tiga pengamatan lemah, sehingga hasil ini dibaca sebagai petunjuk bahwa IndoBERT cenderung sedikit lebih tinggi, bukan bukti kuat. Wilcoxon signed-rank tidak dijadikan bukti karena dengan n=3 nilai-p minimumnya 0,25.

4.6 Eksperimen Tambahan pada Pra-pemrosesan, Backbone, dan Konfigurasi Pelatihan

Tiga eksperimen tambahan menguji keterbatasan rancangan utama (Bagian 3.1) dengan data uji, pembagian data, dan hyperparameter yang sama. Hasilnya dilaporkan pada Tabel 7 dan Tabel 8 (koreksi Holm atas 6 uji McNemar; pada baris tiga *seed*, p Holm adalah nilai terbesar antar-seed).

Eksperimen A (pra-pemrosesan minimal pada IndoBERT). Mempertahankan tanda baca menaikkan F1-macro rata-rata IndoBERT dari 0,9242±0,0048 menjadi 0,9297±0,0030 (selisih +0,0055; CI 95% −0,0014 hingga +0,0124). Selisih per *seed* berganti tanda (+0,0104; +0,0075; −0,0015) dan McNemar tidak signifikan, sehingga tidak ditemukan bukti bahwa pembuangan tanda baca merugikan IndoBERT. CI 90% (−0,0003 hingga +0,0113) berada di dalam margin ±0,02 tetapi tidak ±0,01. Kedua LLM tidak diulang dengan pra-pemrosesan ini, sehingga skor tersebut tidak dibandingkan dengan LLM.

Eksperimen C (sensitivitas terhadap konfigurasi pelatihan Qwen). Run Qwen *seed* 42 yang pertama (*batch* 16 tanpa GC) diulang pada konfigurasi hemat (*batch* 8 dengan GC), dan pengulangan ini dipakai sebagai *run* utama agar ketiga *seed* Qwen dan DeepSeek berkonfigurasi sama. F1-macro turun dari 0,9209 menjadi 0,9159 (akurasi 0,9425 menjadi 0,9373; selisih −0,0049; CI 95% −0,0190 hingga +0,0092; p Holm=1,000), tidak signifikan, sedangkan memori puncak turun dari 11.434,7 menjadi 3.572,4 MB. Dengan *run* awal, F1-macro rata-rata Qwen 0,9211±0,0023 (Tabel 7), selisihnya terhadap IndoBERT −0,0031, dan uji t lintas *seed* memberi p=0,3512 (berbanding p=0,0398 pada konfigurasi seragam). Kesimpulan Qwen terhadap IndoBERT karenanya peka terhadap satu *run* pada uji t tiga pengamatan; IndoBERT hanya dapat dikatakan cenderung sedikit lebih tinggi (sekitar 0,5 poin), tanpa bukti keunggulan Qwen. Selisih Qwen terhadap DeepSeek tetap besar pada kedua konfigurasi.

Eksperimen B (backbone). Qwen2.5-Math-1.5B, model dasar DeepSeek-R1-Distill-Qwen-1.5B, dilatih dengan LoRA pada konfigurasi hemat yang sama (*seed* 42) dan mencapai F1-macro 0,8842 (akurasi 0,9154). Hasil ini tidak berbeda signifikan dari DeepSeek (selisih +0,0080; CI 95% −0,0092 hingga +0,0249) dan lebih rendah daripada Qwen2.5-1.5B-Instruct (selisih −0,0317; p Holm=0,0024). Ini konsisten dengan dugaan bahwa ketertinggalan DeepSeek lebih berkaitan dengan model dasar daripada distilasi *reasoning*, tetapi belum dapat disimpulkan karena hanya satu *seed*, hasil tidak signifikan bukan bukti kesetaraan, perbandingan dengan Qwen2.5-1.5B-Instruct juga mengubah tahap post-training, dan Qwen2.5-Math-1.5B tidak identik dengan DeepSeek (max_position_embeddings 4.096 dan tie_word_embeddings True, berbanding 131.072 dan False).

Tabel 8. Selisih F1-macro berpasangan pada eksperimen tambahan (paired bootstrap 5.000 resampel; p McNemar dengan Holm atas 6 uji).

Perbandingan (A − B)	Data	ΔF1	CI 95%	p Holm
IndoBERT praproses minimal − standar	3 seed	0,0055	[−0,0014; 0,0124]	1,000
Qwen − Qwen awal	Seed 42	−0,0049	[−0,0190; 0,0092]	1,000
Qwen-Math − DeepSeek	Seed 42	0,0080	[−0,0092; 0,0249]	1,000
Qwen-Math − Qwen	Seed 42	−0,0317	[−0,0506; −0,0134]	0,0024

Run Qwen *seed* 42 (utama), Qwen-Math, dan IndoBERT praproses minimal dijalankan dengan PyTorch 2.14.1, Transformers 5.22.2, dan PEFT 0.21.2, sedangkan versi pustaka *run* lainnya tidak dicatat, dan *run* Qwen-Math dilanjutkan dari *checkpoint* langkah 280 setelah sesi Colab terputus, sehingga waktu latihnya di Tabel 4 hanya mencakup 560 langkah terakhir dan tidak sebanding dengan *run* lain. Pengaruh perbedaan lingkungan tidak dapat dikesampingkan.

4.7 Pembahasan dan Kaitan dengan Studi Rujukan

Status hipotesis (Bagian 2) adalah sebagai berikut. H1 (Qwen sebanding atau lebih baik daripada IndoBERT) hanya didukung sebagian karena Qwen tidak lebih unggul, rata-ratanya sedikit di bawah IndoBERT (−0,0047; −0,0031 pada *run* awal), kesebandingan hanya dapat dinyatakan pada toleransi ±0,02 F1-macro, dan uji t lintas *seed* cenderung menguntungkan IndoBERT (p=0,040; p=0,351 pada *run* awal) walaupun McNemar per *seed* dan CI bootstrap tidak signifikan. H2 (Qwen mengungguli DeepSeek) didukung konsisten pada ketiga *seed*, tetapi karena backbone tidak identik tidak dapat diatribusikan pada objektif post-training semata. H3 (IndoBERT lebih efisien) didukung secara deskriptif pada parameter terlatih, waktu inferensi (5,8 hingga 6,1 ms berbanding 139,1 hingga 168,7 ms), dan memori puncak, sedangkan waktu latih dan memori Qwen versus DeepSeek tidak menunjukkan perbedaan yang dapat dikaitkan dengan model; efisiensi tidak diuji secara formal.

Temuan ini memberi nuansa terhadap Jin dkk. [14] yang melaporkan LLM *decoder* berbasis LoRA mengungguli *encoder full fine-tuning* pada deteksi teks AI Mandarin, terutama saat terjadi pergeseran distribusi. Pada SmSA keunggulan itu tidak terlihat. Perbedaan tingkat pergeseran distribusi mungkin berperan, tetapi itu hanya dugaan karena pengujian di luar distribusi (OOD) tidak dilakukan, dan perbedaan domain, bahasa, serta ukuran model juga dapat berperan.

Ketertinggalan DeepSeek meski berukuran dan ber-hyperparameter LoRA sama dengan Qwen menunjukkan bahwa pemilihan model dasar untuk LoRA perlu mempertimbangkan jenis post-training dan model dasarnya, bukan semata reputasi reasoning-nya; kontrol Qwen2.5-Math-1.5B menguatkan pentingnya model dasar, meski baru pada satu *seed* (Bagian 4.6). Pada skenario sumber daya terbatas dan inferensi cepat, IndoBERT-base-p1 tetap kompetitif karena selisih F1-macro terhadap Qwen kecil (sekitar 0,5 poin pada rata-rata tiga *seed*) dan tidak menguntungkan Qwen.

4.8 Keterbatasan Penelitian

Keterbatasan paling mendasar adalah bahwa kedua LLM tidak berbagi backbone identik (Bagian 3.3), yaitu Qwen2.5-1.5B-Instruct berbasis Qwen2.5-1.5B, sedangkan DeepSeek-R1-Distill-Qwen-1.5B didistilasi di atas Qwen2.5-Math-1.5B dengan konfigurasi arsitektur berbeda. Kontrol Qwen2.5-Math-1.5B (Bagian 4.6) hanya satu *seed* dan tidak identik dengan DeepSeek, sehingga belum memisahkan efek model dasar dari efek distilasi. Run Qwen *seed* 42 hasil pengulangan berkonfigurasi hemat memakai versi pustaka berbeda dari *run* Qwen lainnya. Waktu latih diukur pada sesi Colab tidak terkontrol (variasi hingga sekitar 3,3 kali), sehingga efisiensi pelatihan hanya dapat dibaca deskriptif. Model dipasang sebagai pengklasifikasi tanpa pembangkitan teks, sehingga *reasoning* dan penjelasan prediksi tidak diuji.

Evaluasi hanya mencakup satu *dataset* (SmSA) tanpa pengujian di luar distribusi, sehingga generalisasi ke domain lain dan penjelasan melalui pergeseran distribusi belum dapat dipastikan. Tiga *seed* mengurangi keterbatasan single-run [24] tetapi masih moderat sehingga uji Wilcoxon tidak memiliki power memadai. Konfigurasi LoRA (r=16, *epoch*=3) dipilih berdasarkan praktik umum tanpa ablasi, dan ablasi pra-pemrosesan minimal belum diulang pada kedua LLM. Kebocoran antar-subset kecil (0,78% data uji) dan tidak mengubah kesimpulan pada analisis sensitivitas, tetapi tidak dibersihkan dari hasil utama. Hasil tidak dapat dibandingkan langsung dengan papan peringkat IndoNLU (Bagian 3.2), dan margin ekuivalensi ditetapkan penulis.

5 Kesimpulan

Penelitian ini membandingkan Qwen2.5-1.5B-Instruct dan DeepSeek-R1-Distill-Qwen-1.5B dengan LoRA (hyperparameter disamakan) terhadap IndoBERT-base-p1 *full fine-tuning* untuk

analisis sentimen Bahasa Indonesia pada SmSA. Berdasarkan rata-rata tiga *seed*, F1-macro IndoBERT 0,9242±0,0048, Qwen 0,9195±0,0039 (0,9211±0,0023 pada *run* awal berkonfigurasi berbeda), dan DeepSeek 0,8779±0,0048. Selisih Qwen dan IndoBERT kecil (sekitar 0,5 poin) dan tidak signifikan menurut McNemar per *seed* maupun CI 95% paired bootstrap, tetapi paired t-test lintas *seed* memberi p=0,040 (p=0,3512 pada *run* awal); IndoBERT karenanya cenderung sedikit lebih tinggi tanpa bukti keunggulan Qwen, dan kesebandingan hanya dapat dinyatakan pada toleransi ±0,02 F1-macro. DeepSeek tertinggal signifikan dari keduanya pada ketiga *seed* (sekitar 4,2 poin terhadap Qwen dan 4,6 poin terhadap IndoBERT). Karena backbone kedua LLM tidak identik, ketertinggalan ini tidak dapat diatribusikan pada objektif post-training semata, dan kontrol Qwen2.5-Math-1.5B pada satu *seed* (0,8842) mengarah pada peran model dasar. Seluruh model dipasang sebagai pengklasifikasi, sehingga kemampuan *reasoning* tidak diuji.

Dari sisi efisiensi, IndoBERT-base-p1 jauh lebih cepat pada inferensi (5,8 hingga 6,1 ms per sampel dibandingkan 139,1 hingga 168,7 ms pada model LoRA) dan lebih hemat memori, karena ukurannya sekitar 12,5 kali lebih kecil, sehingga penghematan LoRA berlaku relatif terhadap *full fine-tuning* model 1,5 miliar parameter, bukan terhadap *encoder* kecil. Perbedaan waktu latih dan memori Qwen versus DeepSeek tidak dapat disimpulkan sebagai perbedaan efisiensi model. Keunggulan LLM berbasis LoRA atas *encoder* yang dilaporkan Jin dkk. [14] tidak terlihat pada SmSA.

Untuk penelitian selanjutnya, kontrol backbone Qwen2.5-Math-1.5B dan ablasi pra-pemrosesan minimal perlu diulang pada lebih banyak *seed* dan pada kedua LLM, evaluasi diperluas ke *dataset* dan domain sentimen Indonesia lain, dan ablasi rank LoRA, jumlah *epoch*, serta metode PEFT lain seperti QLoRA [25] atau DoRA [26] dapat dilakukan.

Referensi

[1] W. J. Kusoema and I. Ibrahim, "*Sentiment Analysis on the PT Pertamina Corruption Case using IndoBERT and RCNN Methods*," *Sistemasi: Jurnal Sistem Informasi*, Vol. 14, No. 5, pp. 2246–2257, 2025, DOI: 10.32520/stmsi.v14i5.5392.

[2] C.-H. Lin and U. Nuha, "*Sentiment Analysis of Indonesian Datasets based on a Hybrid Deep-Learning Strategy*," *Journal of Big Data*, Vol. 10, No. 1, p. 88, 2023, DOI: 10.1186/s40537-023-00782-9.

[3] A. H. Nasution, A. Onan, Y. Murakami, W. Monika, and A. Hanafiah, "*Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets*," *IEEE Access*, Vol. 13, pp. 94009–94025, 2025, DOI: 10.1109/ACCESS.2025.3574629.

[4] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "*IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*," in *Proc. 1st Conf. Asia-Pacific Chapter ACL and 10th Int. Joint Conf. Natural Language Processing (AACL-IJCNLP)*, 2020, pp. 843–857, DOI: 10.18653/v1/2020.aacl-main.85.

[5] F. Koto, J. H. Lau, and T. Baldwin, "*IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization*," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 10660–10668, DOI: 10.18653/v1/2021.emnlp-main.833.

[6] D. Suharjanto and M. Mustakim, "*Optimizing Emotion Recognition in Indonesian Product Reviews using LoRA on Transformer*," *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, Vol. 5, No. 2, pp. 118–127, 2026, DOI: 10.20885/snati.v5.i2.48248.

[7] F. M. Al Simabua and L. Alfat, "*Optimization of IndoBERT-Lite Fine-Tuning for Spam Detection in Digital Customer Services*," *Sistemasi: Jurnal Sistem Informasi*, Vol. 15, No. 5, pp. 1886–1899, 2026, DOI: 10.32520/stmsi.v15i5.6398.

[8] Qwen Team, "*Qwen2.5 Technical Report*," arXiv preprint arXiv:2412.15115, 2024, DOI: 10.48550/arXiv.2412.15115.

[9] D. Guo et al., "*DeepSeek-R1 Incentivizes Reasoning in LLMs Through Reinforcement Learning*," Nature, Vol. 645, pp. 633–638, 2025, DOI: 10.1038/s41586-025-09422-z.

 ISSN:2302-8149
e-ISSN:2540-9719

[10] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "*Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey*," arXiv preprint arXiv:2403.14608, 2024, DOI: 10.48550/arXiv.2403.14608.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "*LoRA: Low-Rank Adaptation of Large Language Models*," in Proc. 10th Int. Conf. Learning Representations (ICLR), 2022, URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[12] X. Wang and A. Azman, "*LoRA-based Efficient Fine-Tuning of Transformer Models for Short Text Classification*," in *Advances in Visual Informatics (IVIC 2025), Lecture Notes in Computer Science,* Vol. *16215*. Singapore: Springer, 2026, pp. 168–179, DOI: 10.1007/978-981-95-3808-9_13.

[13] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "*Optimizing Airline Review Sentiment Analysis: A Comparative Analysis of LLaMA and BERT Models through Fine-Tuning and Few-Shot Learning*," *Computers, Materials & Continua*, Vol. 82, No. 2, pp. 2769–2792, 2025, DOI: 10.32604/cmc.2025.059567.

[14] H. Jin, N. Ashrafi, A. Abdollahi, W. Liu, J. Wang, G. Gui, M. Pishgar, and H. Feng, "*LLM Encoder vs. Decoder: Robust Detection of Chinese AI-Generated Text with LoRA*," arXiv preprint arXiv:2509.00731, 2025, DOI: 10.48550/arXiv.2509.00731.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "*BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*," in *Proc. 2019 Conf. North American Chapter ACL: Human Language Technologies,* Vol. 1, 2019, pp. 4171–4186, DOI: 10.18653/v1/N19-1423.

[16] M. A. A. Saputra, A. Alamsyah, D. P. Ramadhani, T. S. Siadari, and H. Fakhrurroja, "*IndoBERT-Sentiment: Context-Conditioned Sentiment Classification for Indonesian Text*," arXiv preprint arXiv:2604.07057, 2026, DOI: 10.48550/arXiv.2604.07057.

[17] V. R. Joseph, "*Optimal Ratio for Data Splitting*," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Vol. 15, No. 4, pp. 531–538, 2022, DOI: 10.1002/sam.11583.

[18] J. Opitz, "*A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice*," *Transactions of the Association for Computational Linguistics*, Vol. 12, pp. 820–836, 2024, DOI: 10.1162/tacl_a_00675.

[19] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, "*A Survey of Methods for Addressing Class Imbalance in Deep-Learning based Natural Language Processing*," in *Proc. 17th Conf. European Chapter ACL (EACL)*, 2023, pp. 523–540, DOI: 10.18653/v1/2023.eacl-main.38.

[20] Q. McNemar, "*Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages*," *Psychometrika*, Vol. 12, No. 2, pp. 153–157, 1947, DOI: 10.1007/BF02295996.

[21] B. Efron, "*Bootstrap Methods: Another Look at the Jackknife*," *The Annals of Statistics*, Vol. 7, No. 1, pp. 1–26, 1979, DOI: 10.1214/aos/1176344552.

[22] Y. Bestgen, "*Please, Don't Forget the Difference and the Confidence Interval when Seeking for the State-of-the-Art Status*," in *Proc. 13th Language Resources and Evaluation Conf. (LREC)*, 2022, pp. 5956–5962, URL: https://aclanthology.org/2022.lrec-1.640.

[23] Y. Kim, M. Lee, and S. Lee, "*A Survey on Memory-Efficient Fine-Tuning for Large Language Models*," *Transactions of the Association for Computational Linguistics*, Vol. 14, pp. 960–982, 2026, DOI: 10.1162/tacl.a.692.

[24] N. T. Bui, G. Savova, and L. Wang, "*Assessing the Macro and Micro Effects of Random Seeds on Fine-Tuning Large Language Models*," arXiv preprint arXiv:2503.07329, 2025, DOI: 10.48550/arXiv.2503.07329.

[25] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "*QLoRA: Efficient Finetuning of Quantized LLMs*," in Advances in Neural Information Processing Systems 36 (NeurIPS), 2023, DOI: 10.48550/arXiv.2305.14314.

[26] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "*DoRA: Weight-Decomposed Low-Rank Adaptation*," in *Proc. 41st Int. Conf. Machine Learning (ICML)*, 2024, pp. 32100–32121, DOI: 10.48550/arXiv.2402.09353.