JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.9518

FINE-TUNING MODEL BAHASA BESAR OPEN-SOURCE UNTUK LITERASI DAN KREATIVITAS INDONESIA: EVALUASI KOMPARATIF

Yoga Harisfadila[1*], Yulhendri[2]

[1,2]Teknik Informatika/Universitas Esa Unggul; Jl Arjuna Utara No.9, West Jakarta 11510, Indonesia

Keywords:
Large Language Models; Fine-Tuning; Chatbot; Literacy; Creativity

Corespondent Email:
yorisfad@student.esaunggul.ac.id

Abstrak. Adaptasi Model Bahasa Besar (LLM) ke dalam konteks budaya non-Inggris, terutama dalam bidang yang sarat nuansa seperti literasi dan kreativitas, masih menjadi tantangan yang signifikan. Penelitian ini mengatasi kesenjangan tersebut dengan mengembangkan dan mengevaluasi 'Cakrakarsa', sebuah chatbot AI yang dibuat melalui penyempurnaan model open-source Mistral-7B-Instruct. Model ini diadaptasi menggunakan teknik QLoRA pada dataset yang dikurasi khusus berisi 11.330 entri yang mencakup cerita rakyat, sastra, dan puisi Indonesia. Kinerja diukur terhadap model dasar melalui kerangka evaluasi internal yang komprehensif, yang menggabungkan metrik kuantitatif (Perplexity, Evaluation Loss) dan penilaian kualitatif melalui metode LLM-as-a-Judge pada 30 prompt khusus. Model yang telah disesuaikan menunjukkan keunggulan yang signifikan, dengan pengurangan sebesar 67,11% pada Perplexity dan 61,52% pada Evaluation Loss. Secara kualitatif, model inimenang dalam 81,11% perbandingan langsung, menunjukkan peningkatan yangmencolok dalam pemahaman konteks budaya, koherensi, dan kreativitas. Temuan kritis mengungkapkan adanyapertukaran antara spesialisasi domain yang mendalam dan pengetahuan umum yang luas. Penelitian inimenyumbangkan chatbot yang fungsional dan sadar budaya, serta metodologi yang kokoh dan dapat direplikasi untuk menyesuaikan LLM dengan domain budaya spesifik dengan sumber daya komputasi yang terbatas.

Abstract. *The adaptation of Large Language Models (LLMs) to non-English cultural contexts, particularly in nuanced domains like literacy and creativity, remains a significant challenge. This study addresses this gap by developing and evaluating 'Cakrakarsa', an AI chatbot created by fine-tuning the open-source Mistral-7B-Instruct model. The model was adapted using the QLoRA technique on a custom-curated dataset of 11,330 entries comprising Indonesian folklore, literature, and poetry. Performance was measured against the baseline model through a comprehensive internal evaluation framework, combining quantitative metrics (Perplexity, Evaluation Loss) and qualitative assessment via the LLM-as-a-Judge method on 30 dedicated prompts. The fine-tuned model demonstrated significant superiority, achieving a 67.11% reduction in Perplexity and a 61.52% reduction in Evaluation Loss. Qualitatively, it won 81.11% of head-to-head comparisons, showing marked improvements in cultural context understanding, coherence, and creativity. A critical finding reveals a trade-off between deep domain specialization and broad general knowledge. This research contributes a functional, culturally-aware chatbot and a robust, replicable methodology for adapting LLMs to specific cultural domains with limited computational resources.*

1. PENDAHULUAN

Perkembangan pesat Large Language Models (LLM) dalam beberapa tahun terakhir telah membuka jalan bagi pengembangan chatbot berbasis kecerdasan buatan yang mampu mengemulasi interaksi seperti manusia di berbagai domain, termasuk literasi dan kreativitas. Model-model terkemuka seperti ChatGPT, Claude, Gemini, dan Llama telah menunjukkan kemampuan yang signifikan dalam tugas-tugas natural language processing (NLP), sehingga menjadi fondasi yang kuat bagi pengembangan alat bantu pendidikan dan kreativitas berbasis AI [1], [2]. Kemampuan generatif model-model ini telah terbukti dapat mendukung berbagai aplikasi, mulai dari penulisan kreatif hingga pemahaman teks kompleks, menjadikannya kandidat yang menjanjikan untuk diterapkan dalam konteks pendidikan.

Namun demikian, tantangan utama tetap ada dalam mengadaptasi model-model tersebut ke bahasa lokal dan konteks budaya yang spesifik, khususnya di Indonesia. Model-model open-source yang ada saat ini kerap menunjukkan bias budaya yang signifikan, yang merupakan konsekuensi dari dominasi data pelatihan berbahasa Inggris, sehingga menghambat efektivitasnya dalam mendukung literasi dan kreativitas berbahasa Indonesia [3]. Permasalahan ini diperparah oleh indikator-indikator nasional yang memprihatinkan. Berdasarkan Global Innovation Index 2023, Indonesia menempati posisi ke-61 dari 132 negara, menunjukkan kesenjangan yang signifikan dibandingkan negara-negara tetangga seperti Singapura dan Malaysia [4]. Senada dengan itu, hasil Programme for International Student Assessment (PISA) 2022 mengungkapkan bahwa hanya 25% siswa Indonesia yang mencapai kemampuan minimum dalam membaca, jauh di bawah rata-rata OECD sebesar 74% [5]. Kondisi ini menegaskan urgensi pemanfaatan sistem berbasis AI untuk meningkatkan literasi dan kemampuan berpikir kreatif pelajar Indonesia.

Penelitian-penelitian terdahulu menunjukkan bahwa fine-tuning LLM open-source menggunakan dataset yang relevan secara lokal dapat meningkatkan adaptasi kontekstual dan performa model secara signifikan pada domain-domain khusus [6], [7]. Sebagai contoh, penelitian di bidang keuangan telah membuktikan bahwa LLM yang di-fine-tune pada data domain spesifik mampu melampaui performa model umum dalam tugas-tugas terspesialisasi [7]. Di sisi evaluasi, metode seperti LLM-as-a-Judge menyediakan pendekatan terstruktur untuk mengukur aspek-aspek kualitatif seperti kreativitas, yang sering kali tidak tertangkap oleh benchmark tradisional [8]. Meskipun demikian, kerangka kerja komprehensif yang mengintegrasikan fine-tuning dengan evaluasi yang andal dan disesuaikan dengan konteks bahasa dan budaya Indonesia masih sangat terbatas. Khususnya, belum ada penelitian yang secara sistematis mengembangkan dan mengevaluasi LLM open-source untuk domain literasi dan kreativitas Indonesia dengan menggunakan dataset kurasi lokal yang mencakup folklor, sastra, dan puisi secara bersamaan.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk mengembangkan dan mengevaluasi chatbot AI berbahasa Indonesia bernama 'Cakrakarsa' yang dihasilkan melalui proses fine-tuning model Mistral-7B-Instruct menggunakan dataset kurasi lokal. Secara spesifik, penelitian ini menjawab dua pertanyaan utama: (1) Sejauh mana proses fine-tuning dengan dataset lokal dapat meningkatkan performa model secara kuantitatif (Perplexity dan Evaluation Loss) dan kualitatif (penilaian LLM-as-a-Judge) pada domain literasi dan kreativitas Indonesia? dan (2) Apa saja konsekuensi dan trade-off yang muncul dari proses spesialisasi domain tersebut? Penelitian ini memberikan kontribusi berupa chatbot yang fungsional dan sadar budaya, serta metodologi yang dapat direplikasi untuk mengadaptasi LLM pada konteks lokal dengan sumber daya komputasi yang terbatas.

2. TINJAUAN PUSTAKA

2.1 Large Language Models (LLM)

Large Language Models (LLM) merupakan model kecerdasan buatan berbasis arsitektur *Transformer* yang dilatih pada korpus teks berukuran sangat besar untuk memahami dan menghasilkan bahasa alami. Naveed et al. memberikan tinjauan komprehensif tentang perkembangan LLM modern, menjelaskan bahwa model-model ini mampu melakukan berbagai tugas NLP tanpa pelatihan tugas-spesifik (*zero-shot*) maupun dengan sedikit

contoh (*few-shot*) [1]. Brown et al. melalui penelitian GPT-3 membuktikan bahwa LLM berskala besar yang dilatih dengan paradigma *few-shot learning* mampu mencapai performa kompetitif di berbagai *benchmark* tanpa *fine-tuning* eksplisit [2].

Meskipun demikian, LLM yang ada saat ini umumnya dilatih dengan dominasi data berbahasa Inggris, sehingga menghasilkan bias budaya yang signifikan ketika diterapkan pada bahasa dan konteks non-Inggris. Tao et al. menunjukkan bahwa bias budaya dalam LLM bukan hanya masalah linguistik, tetapi juga menyangkut representasi nilai, norma, dan perspektif yang tertanam dalam data pelatihan [3]. Kondisi ini menjadi tantangan utama dalam pemanfaatan LLM untuk konteks bahasa Indonesia.

2.2 Fine-Tuning

Fine-tuning merupakan proses adaptasi model bahasa yang telah dilatih secara umum (*pre-trained*) terhadap tugas atau domain tertentu menggunakan dataset yang lebih spesifik. Wei et al. membuktikan bahwa model bahasa yang di-*fine-tune* pada kumpulan instruksi yang beragam (*instruction tuning*) mampu menggeneralisasi kemampuannya ke tugas-tugas baru yang belum pernah dilihat sebelumnya (*zero-shot learners*) [6]. Pendekatan ini terbukti lebih efisien dibandingkan melatih model dari awal (*training from scratch*), terutama ketika sumber daya komputasi terbatas.

Dalam konteks domain spesifik, Jeong mendemonstrasikan bahwa *fine-tuning* model Mistral-7B pada data keuangan menghasilkan peningkatan performa yang signifikan dibandingkan model umum untuk tugas-tugas analisis finansial [7]. Temuan ini mengkonfirmasi bahwa spesialisasi domain melalui *fine-tuning* adalah strategi yang valid dan efektif, sekaligus menjadi dasar pendekatan yang digunakan dalam penelitian ini untuk domain literasi dan kreativitas Indonesia.

2.3 LoRA dan QLoRa

Tantangan utama dalam *fine-tuning* LLM berskala besar adalah kebutuhan sumber daya komputasi yang sangat tinggi. Untuk mengatasi hal ini, Hu et al. memperkenalkan *Low-Rank Adaptation* (LoRA), sebuah teknik *Parameter-Efficient Fine-Tuning* (PEFT) yang bekerja dengan menyisipkan matriks berdimensi rendah (*low-rank matrices*) yang dapat dilatih ke dalam lapisan-lapisan arsitektur *Transformer*, sementara bobot model asli dibekukan [9]. Dengan cara ini, jumlah parameter yang perlu dilatih berkurang secara drastis tanpa mengorbankan performa model secara signifikan.

Dettmers et al. selanjutnya mengembangkan LoRA menjadi QLoRA (*Quantized LoRA*), yang menggabungkan teknik kuantisasi 4-bit pada model dasar dengan pelatihan adapter LoRA [10]. Kombinasi ini memungkinkan *fine-tuning* model berparameter miliaran pada GPU tunggal dengan VRAM terbatas, yang sebelumnya tidak memungkinkan. QLoRA menjadi terobosan penting dalam demokratisasi riset LLM, karena memungkinkan peneliti dengan sumber daya terbatas untuk tetap melakukan eksperimen *fine-tuning* yang bermakna.

2.4 Evaluasi Model

Evaluasi LLM secara komprehensif memerlukan pendekatan yang melampaui metrik kuantitatif standar. Zhao et al. menyoroti bahwa aspek-aspek seperti kreativitas dan orisinalitas dalam keluaran LLM sulit diukur secara objektif menggunakan *benchmark* tradisional, sehingga diperlukan pendekatan evaluasi yang lebih holistik [8]. Dalam konteks ini, metode *LLM-as-a-Judge* yang di mana LLM lain digunakan sebagai penilai (*judge*) untuk membandingkan kualitas respons telah berkembang sebagai alternatif yang menjanjikan.

Namun, penerapan metode ini tidak lepas dari tantangan. Liu et al. menemukan bahwa kualitas evaluasi *LLM-as-a-Judge* sangat sensitif terhadap formulasi *prompt* evaluasi, yang dapat menimbulkan inkonsistensi hasil [11]. Lebih lanjut, Chen et al. mengidentifikasi bahwa LLM yang digunakan sebagai *judge* memiliki bias inheren, termasuk kecenderungan memilih respons yang lebih panjang atau yang berasal dari model tertentu [12]. Temuan-temuan ini menjadi dasar pertimbangan dalam merancang protokol evaluasi yang ketat pada penelitian ini, termasuk penggunaan panel hakim (*panel of judges*) dan protocol *blind test* untuk meminimalkan bias.

2.5 Chatbot

Chatbot merupakan program komputer yang dirancang untuk menirukan percakapan layaknya interaksi antarmanusia dengan memanfaatkan kecerdasan buatan serta pemrosesan bahasa alami, sehingga mampu memahami dan merespons pertanyaan pengguna secara otomatis [13]. Cara kerja *chatbot* dimulai dari masukan pengguna dalam bentuk bahasa alami, kemudian sistem mengolah *input* tersebut dan menghasilkan respons yang relevan dan logis sesuai konteks percakapan. Pemanfaatan *chatbot* telah meluas ke berbagai sektor industri karena kemampuannya beroperasi tanpa batasan waktu operasional manusia sekaligus mengurangi beban kerja staf secara signifikan.

3. METODE PENELITIAN

Penelitian ini dirancang sebagai studi eksperimental komparatif yang bertujuan mengadaptasi LLM open-source untuk domain literasi dan kreativitas Indonesia melalui proses fine-tuning, kemudian mengevaluasi hasilnya secara komprehensif. Secara keseluruhan, alur metodologi penelitian ini terdiri dari empat tahap utama: (1) pemilihan model baseline, (2) kurasi dan komposisi dataset, (3) fine-tuning model, dan (4) evaluasi performa model. Seluruh eksperimen dilakukan dalam lingkungan GPU tunggal untuk membuktikan kelayakan penelitian dengan sumber daya komputasi terbatas.

3.1 Pemilihan Model Baseline

Pemilihan model *baseline* merupakan keputusan strategis yang secara langsung mempengaruhi kelayakan dan dampak proses *fine-tuning*. Setelah mempertimbangkan berbagai alternatif model *open-source* dan *closed-source* yang tersedia, model **Mistral-7B-Instruct** dipilih sebagai *baseline* dalam penelitian ini. Perbandingan performa model dalam kategori *Creative Writing* berdasarkan skor ELO dari *Lmsys Chatbot Arena* [14].

Tabel 1. Perbandingan Skor ELO Penulisan Kreatif

Model	Tipe	ELO
Claude 3.5 Sonnet	Closed-Source	1367
Gemini 1.5 Pro	Closed-Source	1359
GPT-4o	Closed-Source	1339
Llama 3.1 405B	Open-Source	1309
Mistral-7B	Open-Source	1104

Pemilihan Mistral-7B didasarkan pada tiga pertimbangan utama:

1. **Efisiensi dan Kelayakan:** Dengan 7,3 miliar parameter, Mistral-7B dirancang untuk dapat di-*fine-tune* pada GPU tunggal menggunakan teknik kuantisasi seperti QLoRA, sehingga memungkinkan eksperimen dengan sumber daya terbatas.
2. **Aksesibilitas *Open-Source*:** Model ini dirilis di bawah lisensi Apache 2.0 yang permisif, menjamin reproduksibilitas dan transparansi penuh untuk keperluan penelitian.
3. **Potensi Kompetitif melalui *Fine-Tuning*:** Penelitian sebelumnya telah menunjukkan bahwa model 7B seperti Mistral dapat melampaui performa GPT-4 pada tugas-tugas bahasa yang terkontrol melalui kombinasi *fine-tuning* dan *alignment* yang tepat [15].

3.2 Kurasi dan Komposisi Dataset

Efektivitas *fine-tuning* sangat bergantung pada kualitas dan relevansi data pelatihan. Untuk penelitian ini, dikembangkan dataset kurasi yang terdiri dari **11.330 entri** yang diselaraskan dengan tujuan peningkatan literasi dan kreativitas dalam konteks Indonesia. Data distrukturkan dalam format pesan JSONL (*JavaScript Object Notation Lines*), yang banyak diadopsi untuk tugas *instruction-tuning*. Setiap entri terdiri dari percakapan yang direpresentasikan dalam skema berbasis *chat*, memfasilitasi kompatibilitas dengan sifat *instruction-following* model dasar.

Dataset mengintegrasikan lima sumber data utama, masing-masing dipilih untuk

memberikan atribut budaya dan linguistik yang berbeda, sebagaimana diuraikan pada list dibawah ini:

1. Balinese Story: Narasi budaya Bali; kedalaman budaya dan nilai tradisional
2. Folklore Indonesia: Kumpulan cerita rakyat dari berbagai daerah; representasi geografis dan budaya
3. Indocorpus-Sastra: Subset proyek IndoCorpus; karya sastra; kekayaan gaya bahasa Indonesia
4. IndoLEM: Dataset linguistik terstruktur; penalaran logis dan koherensi
5. Koleksi Puisi: Koleksi puisi; fleksibilitas puitis dan kreativitas generasi Bahasa

Dengan mengintegrasikan kelima sumber tersebut, dataset mencapai keseimbangan antara keaslian budaya, variasi linguistik, dan ekspresi sastra. Kombinasi ini memungkinkan model yang di-*fine-tune* untuk beradaptasi tidak hanya pada tugas literasi fungsional, tetapi juga pada ekspresi kreatif, sehingga menjawab dimensi pragmatis sekaligus artistik penggunaan bahasa Indonesia.

3.3 Metodologi Fine-Tuning

Untuk mengadaptasi model *baseline* secara efisien dalam keterbatasan komputasi, penelitian ini menggunakan Teknik *Parameter-Efficient Fine-Tuning* (PEFT), yaitu *Quantized Low-Rank Adaptation* (QLoRA) [10]. Pendekatan QLoRA bekerja dengan mengkuantisasi model dasar ke 4-bit untuk mengurangi jejak memori, kemudian melatih hanya sekumpulan kecil matriks berdimensi rendah (*LoRA adapters*) yang disisipkan ke dalam arsitektur *Transformer* [9]. Dengan cara ini, *fine-tuning* model berskala besar pada GPU tunggal menjadi layak secara komputasi.

Untuk mengoptimalkan proses pelatihan lebih lanjut, digunakan pustaka akselerasi Unsloth [16]. Pustaka ini meningkatkan efisiensi komputasi secara signifikan melalui pemanfaatan *custom Triton kernels*, yang menghasilkan waktu pelatihan lebih cepat dan penggunaan VRAM lebih rendah dibandingkan implementasi standar. Konfigurasi *hyperparameter* kunci yang digunakan dalam proses *fine-tuning* dirinci pada Tabel 2.

Tabel 2. Hyperparameter Kunci Proses Fine-Tuning

Hyperparameter	Nilai
Model Dasar	unsloth/mistral-7b-instruct-v0.3-bnb-4bit
LoRA Rank (r)	16
LoRA Alpha	16
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Effective Batch Size	8 (2 per perangkat × 4 langkah akumulasi)
Learning Rate	2e-4
Epoch	1
Optimizer	adamw_8bit

Konfigurasi ini dipilih secara strategis untuk memastikan adaptasi model dasar yang efektif dan dapat direproduksi pada domain target, sekaligus beroperasi dalam anggaran komputasi yang tersedia.

3.4 Kerangka Evaluasi

Untuk mengukur dampak proses *fine-tuning* secara komprehensif, ditetapkan kerangka evaluasi multi-aspek yang menggabungkan metrik kuantitatif dan penilaian kualitatif.

Evaluasi Kuantitatif dilakukan menggunakan dua metrik pemodelan bahasa standar:

a. ***Perplexity* (PPL):** Dihitung pada *proxy test set* yang dibuat secara manual dan tidak terlihat selama pelatihan (TestLK.jsonl) untuk mengukur kemampuan generalisasi model pada data baru. Skor PPL yang lebih rendah mengindikasikan pemahaman pola bahasa yang lebih baik.
b. ***Evaluation Loss*:** Dihitung pada seluruh dataset pelatihan untuk memvalidasi sejauh mana model berhasil menginternalisasi materi pelatihan. Skor yang lebih rendah mengindikasikan proses pembelajaran yang lebih efektif.

Evaluasi Kualitatif dilakukan menggunakan paradigma LLM-as-a-Judge, mengingat metrik kuantitatif tidak dapat sepenuhnya menangkap aspek-aspek bernuansa seperti kreativitas dan pemahaman konteks budaya [11]. Untuk memitigasi risiko inkonsistensi dan bias inheren yang teridentifikasi dalam metode ini [12], protokol yang ketat diterapkan melalui tiga strategi:

1. ***Panel* Hakim AI:** Tiga model *frontier* (OpenAI o3, Claude 4 Opus, dan Gemini

2.5 Pro) digunakan secara paralel untuk mengurangi bias model tunggal.
2. **Protokol *Blind Test*:** Respons dari kedua model disajikan dalam urutan acak untuk mencegah bias posisi.
3. ***Rubrik* Objektif:** Seluruh penilaian didasarkan pada *rubrik* terperinci yang telah ditetapkan sebelumnya, bukan pada preferensi subjektif.

Inti dari penilaian ini Adalah *rubrik* objektif yang dirancang untuk mengukur dimensi-dimensi kunci kualitas respons, sebagaimana dirangkum pada list dibawah ini:
1. Relevance & Instruction Following: Ketepatan dalam memahami dan memenuhi seluruh bagian prompt
2. Coherence & Fluency: Kualitas penulisan, alur logis, dan penggunaan bahasa Indonesia yang natural
3. Cultural Context Understanding: Kemampuan memproses dan menerapkan nuansa budaya Indonesia yang spesifik
4. Creativity & Originality: Imajinasi, kebaruan, dan kualitas artistik dalam tugas kreatif
5. Accuracy & Depth of Analysis: Ketepatan faktual dan kedalaman penalaran dalam tugas analitis

4. HASIL DAN PEMBAHASAN

4.1 Hasil Kuantitatif

Evaluasi kuantitatif memberikan bukti numerik yang objektif mengenai peningkatan performa yang dicapai oleh model hasil *fine-tuning* (selanjutnya disebut 'Cakrakarsa') dibandingkan dengan model *baseline*-nya (Mistral-7B-Instruct). Performa diukur menggunakan dua metrik utama: *Perplexity* (PPL) untuk menilai kemampuan generalisasi pada data yang belum pernah dilihat, dan *Evaluation Loss* untuk memvalidasi internalisasi materi pelatihan oleh model.

Temuan 1: Penurunan *Perplexity* sebesar 67,11%

Perplexity digunakan untuk mengukur tingkat 'kebingungan' model ketika memprediksi teks dari *proxy test set* yang tidak terlihat selama pelatihan. Skor PPL yang lebih rendah mengindikasikan pemahaman yang lebih baik terhadap pola bahasa, gaya narasi, dan kosakata domain target. Hasil perbandingan disajikan pada Tabel 3.

Tabel 3. Perbandingan Skor Perplexity pada Test Set

Model	Skor *Perplexity* (PPL)
Baseline (Model B)	293,77
Cakrakarsa (Model A)	96,62

Model Cakrakarsa mencapai skor PPL sebesar 96,62, merepresentasikan penurunan substansial sebesar 67,11% dibandingkan skor *baseline* sebesar 293,77. Temuan ini memberikan bukti statistik yang kuat bahwa proses *fine-tuning* secara signifikan meningkatkan kemampuan model dalam memahami dan memprediksi teks dalam domain literasi dan kreativitas Indonesia.

Temuan 2: Penurunan Evaluation Loss sebesar 61,52%

Evaluation Loss dihitung pada seluruh dataset pelatihan untuk mengkonfirmasi seberapa efektif setiap model mempelajari materi yang diberikan. Metrik ini mengkuantifikasi rata-rata kesalahan prediksi, di mana skor yang lebih rendah mengindikasikan proses pembelajaran yang lebih berhasil. Perbandingan disajikan pada Tabel 4.

Tabel 4. Perbandingan Evaluation Loss pada Data Pelatihan

Model	Skor *Evaluation Loss*
Baseline (Model B)	2,6530
Cakrakarsa (Model A)	1,0208

Hasil menunjukkan penurunan *loss* sebesar 61,52% untuk model Cakrakarsa. Hal ini memvalidasi bahwa model hasil *fine-tuning* tidak hanya belajar secara dangkal (*surface-learn*), melainkan berhasil menginternalisasi pola-pola mendasar dari data pelatihan secara jauh lebih efektif dibandingkan model *baseline*. Secara keseluruhan, kedua hasil kuantitatif ini secara konsisten menunjukkan superioritas yang signifikan dari model Cakrakarsa atas *baseline*-nya, dan membentuk fondasi yang kuat untuk analisis kualitatif berikut.

4.2 Hasil Kualitatif

Meskipun metrik kuantitatif menunjukkan keunggulan secara statistik, penilaian kualitatif sangat penting untuk memahami peningkatan fungsional dan perilaku bernuansa dari model.

Bagian ini menyajikan temuan dari evaluasi *LLM-as-a-Judge*, di mana panel tiga hakim AI melakukan 90 perbandingan *blind-test* independen terhadap 30 *Golden Prompts*.

Temuan 3: Cakrakarsa memenangkan 81,11% perbandingan head-to-head
Hasil agregat disajikan pada Tabel 5.

Tabel 5. Hasil Penilaian Kualitatif Agregat

Hasil	Jumlah (dari 90)	Persentase
Cakrakarsa (Model A) Menang	73	81,11%
Baseline (Model B) Menang	13	14,44%
Seri	4	4,44%

Kemenangan lebih dari 81% dalam perbandingan *head-to-head* mengindikasikan lompatan performa yang fundamental, bukan sekadar peningkatan marginal. Rendahnya jumlah hasil seri (4,44%) lebih lanjut menunjukkan bahwa perbedaan kualitas antara kedua model sangat nyata dan mudah dibedakan dalam sebagian besar kasus.

Temuan 4: Peningkatan merata di seluruh kategori evaluasi

Untuk memahami di mana peningkatan tersebut terjadi, skor rata-rata untuk setiap kategori evaluasi dianalisis sebagaimana dirinci pada Tabel 6.

Tabel 6. Rata-rata Skor per Kategori Evaluasi (Skala 0-5)

Kategori	Cakrakarsa (A)	*Baseline* (B)
Coherence & Fluency	3,76	2,21
Accuracy & Depth of Analysis	2,42	1,07
Creativity & Originality	2,72	1,41
Relevance & Instruction Following	3,30	2,17
Cultural Context Understanding	2,77	1,81
Rata-rata Keseluruhan	**3,10**	**1,86**

Model hasil *fine-tuning* mengungguli *baseline* di setiap kategori. Peningkatan paling dramatis terlihat pada *Coherence & Fluency* (+1,55 poin), mengkonfirmasi bahwa proses *fine-tuning* secara fundamental meningkatkan kemampuan model dalam menghasilkan teks Indonesia yang natural dan terstruktur dengan baik. Peningkatan signifikan juga terlihat pada *Accuracy & Depth of Analysis* dan *Creativity & Originality*, area di mana model *baseline* menunjukkan performa yang sangat rendah.

Temuan 5 (Kritis): Trade-off antara Spesialisasi Domain dan Pengetahuan Umum

Untuk memberikan bukti konkret dan menyoroti temuan kritis, studi kasus berikut memeriksa sebuah *prompt* di mana model *baseline* dinilai lebih unggul. Kasus ini, disajikan pada list yang akan diuaraikan dibawah paragraph ini, menggunakan *prompt* yang dirancang untuk menguji pengetahuan faktual spesifik mengenai legenda lokal Indonesia.

Prompt: *"Identifikasi tokoh-tokoh kunci dalam 'Legenda Buaya Putih Kali Bekasi' dan jelaskan peran masing-masing dalam konflik utama cerita."*

A. Respons Model A (Cakrakarsa):

"Tokoh-tokoh kunci dalam cerita ini adalah: 1. Kyai Gentayu: Seorang kuda pemberani yang menjadi teman dan pelindung sang pangeran. 2. Prabu Siliwangi: Raja Bekasi yang bijaksana dan pemimpin. 3. Pangeran Diponegoro: Pemimpin perang yang berani dan bijaksana. 4. Kyai Gentayu: Kuda pemberani yang menjadi teman dan pelindung sang pangeran. 5. Prajurit Belanda: Pihak yang menjadi antagonis utama. 6. Kyai Gentayu (Kembali): Kuda pemberani yang kembali menjadi teman dan pelindung sang pangeran."

B. Respons Model B (Baseline):

"Dalam legenda Buaya Putih Kali Bekasi, ada beberapa tokoh-tokoh yang memainkan peranan penting dalam konflik utama cerita. 1. Buaya Putih Kali Bekasi: Buaya putih ini adalah tokoh utama dalam legenda. Dia merupakan pembela masyarakat... 2. Raja Kawasan: Raja kawasan adalah tokoh negatif dalam legenda..."

Kasus ini mengungkap *trade-off* kritis antara spesialisasi domain dan pengetahuan umum yang luas dimana sebuah temuan yang akan dibahas secara mendalam pada bagian berikut.

4.3 Pembahasan

4.3.1 Interpretasi Hasil: Efektivitas Fine-Tuning Domain Spesifik

Hasil kuantitatif dan kualitatif menyajikan narasi yang koheren dan saling memperkuat: proses *fine-tuning* yang ditargetkan terbukti sangat efektif dalam mentransformasi model *baseline* generik menjadi asisten terspesialisasi dengan kompetensi superior di domain literasi dan kreativitas Indonesia. Penurunan *Perplexity* sebesar 67,11% merupakan alasan statistik di balik peningkatan *fluency* dan koherensi yang terobservasi secara kualitatif. Temuan ini sejalan dengan penelitian Wei et al. yang membuktikan bahwa model yang di-*fine-tune* pada instruksi domain spesifik mampu menggeneralisasi kemampuannya secara signifikan dibandingkan model generik [6]. Senada dengan itu, penurunan *Evaluation Loss* sebesar 61,52% mengkonfirmasi internalisasi materi pelatihan yang mendalam, yang termanifestasi sebagai pemahaman konteks budaya dan nuansa tematik yang lebih kuat.

Keberhasilan ini berakar langsung pada komposisi strategis dataset, di mana setiap komponen seperti folklor, sastra, puisi, dan tugas penalaran linguistic dapat berkontribusi pada kemampuan spesifik model akhir. Hal ini konsisten dengan temuan Jeong yang mendemonstrasikan bahwa komposisi dataset yang cermat dan relevan secara domain adalah faktor penentu utama keberhasilan spesialisasi LLM [7].

4.3.2 Trade-off Spesialisasi: Temuan Kritis

Namun demikian, analisis komprehensif harus pula mengkaji anomali yang terungkap dalam studi kasus "Buaya Putih Kali Bekasi". Kegagalan model Cakrakarsa dan halusinasi yang ditemukan berbanding terbalik dengan respons *baseline* yang lebih faktual meski tidak lengkap menyoroti adanya *trade-off* kritis antara spesialisasi domain yang mendalam dan pengetahuan umum yang luas. Proses *fine-tuning* intensif berhasil menciptakan "pakar" domain, namun spesialisasi ini tampaknya berdampak pada berkurangnya kemampuan model untuk mengakses basis pengetahuan pra-latih yang lebih luas ketika menghadapi topik di luar data pelatihannya.

Fenomena ini, yang dikenal sebagai *catastrophic forgetting* dalam literatur pembelajaran mesin, merupakan tantangan yang diakui secara luas dalam adaptasi model [6]. Temuan ini memberikan wawasan krusial tentang konsekuensi bernuansa dari adaptasi model, menegaskan bahwa spesialisasi bukan tanpa potensi kerugian. Oleh karena itu, penelitian mendatang perlu mengeksplorasi strategi mitigasi seperti integrasi *Retrieval-Augmented Generation* (RAG) untuk mempertahankan akses ke pengetahuan umum sekaligus mempertahankan keahlian domain.

4.3.3 Implikasi Teoritis dan Implementasi

Penelitian ini memberikan beberapa implikasi penting, baik secara teoritis maupun praktis.

Implikasi Teoritis: Penelitian ini memberikan validasi empiris bahwa adaptasi LLM *open-source* ke domain budaya yang bernuansa dan non-Inggris adalah layak dan efektif, melampaui sekadar penerjemahan sederhana untuk menanamkan pemahaman kontekstual yang genuine. Temuan ini memperkuat dan memperluas argumen Tao et al. bahwa bias budaya dalam LLM dapat diatasi melalui intervensi data yang ditargetkan [3], sekaligus menambahkan dimensi baru berupa bukti empiris dalam konteks bahasa Indonesia yang selama ini masih terbatas dalam literatur.

Implikasi Implementasi: Dari perspektif praktis, penelitian ini membuktikan bahwa riset adaptasi LLM yang bermakna adalah layak tanpa infrastruktur komputasi masif, melalui pemanfaatan teknik efisien seperti QLoRA [10] dan pustaka akselerasi seperti Unsloth [16]. Metodologi yang terdokumentasi secara lengkap dalam penelitian ini berfungsi sebagai cetak biru metodologis yang dapat direplikasi oleh peneliti dan pengembang yang bertujuan mengadaptasi LLM untuk konteks lokal lain dengan sumber daya terbatas seperti khususnya untuk bahasa-bahasa daerah di Indonesia yang belum terwakili dengan baik dalam model-model LLM global.

4.3.4 Keterbatasan Penelitian

Meskipun temuan ini signifikan, penting untuk mengakui keterbatasan penelitian. Pertama, penelitian ini merupakan studi kasus yang berfokus pada satu model dasar (Mistral-7B), sehingga temuan mungkin tidak langsung dapat digeneralisasi ke arsitektur lain. Kedua, *proxy test set* yang digunakan untuk *Perplexity*, yang hanya terdiri dari 30 entri, berfungsi sebagai indikator yang kuat namun kurang memiliki kekuatan statistik dari *test set* yang lebih besar. Ketiga, meskipun telah diterapkan strategi mitigasi, metode *LLM-as-a-Judge* tetap memiliki potensi bias inheren dan bukan merupakan pengganti sempurna untuk evaluasi pakar manusia, khususnya untuk tugas-tugas yang bersifat subjektif [11], [12]. Keterbatasan-keterbatasan ini memberikan arah yang jelas bagi penelitian masa depan untuk membangun dan mengembangkan temuan penelitian ini.

5. KESIMPULAN

Penelitian ini berhasil mendemonstrasikan bahwa *fine-tuning* model Mistral-7B-Instruct *open-source* menggunakan dataset konten lokal yang dikurasi secara efektif mentransformasikannya menjadi asisten AI terspesialisasi 'Cakrakarsa' dengan performa superior dalam domain literasi dan kreativitas Indonesia. Berikut adalah simpulan penelitian secara rinci:

1. *Fine-tuning* QLoRA pada dataset 11.330 entri konten lokal Indonesia menghasilkan penurunan *Perplexity* sebesar 67,11% (293,77 → 96,62) dan *Evaluation Loss* sebesar 61,52% (2,6530 → 1,0208).
2. Secara kualitatif, Cakrakarsa memenangkan 81,11% dari 90 perbandingan *blind-test*, dengan peningkatan terbesar pada *Coherence & Fluency* (+1,55 poin).
3. Teridentifikasi *trade-off* kritis: model unggul pada konten domain pelatihannya, namun cenderung berhalusinasi pada topik di luarnya, sebagaimana terungkap dalam studi kasus "Legenda Buaya Putih Kali Bekasi".
4. Seluruh proses *fine-tuning* berhasil dilakukan pada GPU Tunggal via QLoRA dan Unsloth, membuktikan kelayakan adaptasi LLM tanpa infrastruktur komputasi masif.
5. Metodologi penelitian — dari kurasi dataset hingga protocol *LLM-as-a-Judge* — berfungsi sebagai cetak biru yang dapat direplikasi untuk adaptasi LLM pada bahasa dan budaya lokal lainnya.
6. Penelitian terbatas pada satu arsitektur model (Mistral-7B) dengan *proxy test set* hanya 30 entri, sehingga generalisabilitas temuan masih terbatas dan memerlukan validasi lebih lanjut.
7. Penelitian mendatang disarankan mengintegrasikan *Retrieval-Augmented Generation* (RAG) untuk memitigasi *trade-off* pengetahuan, serta memperluas dataset dan pengujian lintas-arsitektur (Llama, Qwen) untuk memperkuat generalisabilitas metodologi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] H. Naveed *et al.*, “A Comprehensive Overview of Large Language Models,” *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, Aug. 2025, doi: 10.1145/3744746.

[2] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, “Cultural Bias and Cultural Alignment of Large Language Models,” Jun. 2024, doi: 10.1093/pnasnexus/pgae346.

[4] WIPO, “Global Innovation Index 2023,” 2023.

[5] OECD, “PISA 2022 Results Factsheets Indonesia PUBE,” 2023. [Online]. Available: https://oecdch.art/a40de1dbaf/C108.

[6] J. Wei *et al.*, “Finetuned Language Models Are Zero-Shot Learners,” in *International Conference on Learning Representations (2022)*, 2022. [Online]. Available: https://openreview.net/forum?id=gEZrGCozdqR

[7] C. Jeong, “Domain-specialized LLM: Financial fine-tuning and utilization method

using Mistral 7B," *Journal of Intelligence and Information Systems*, vol. 30, pp. 93–120, Sep. 2024, doi: 10.13088/jiis.2024.30.1.093.

[8] Y. Zhao *et al.*, "Assessing and Understanding Creativity in Large Language Models," *Machine Intelligence Research*, vol. 22, no. 3, pp. 417–436, 2025, doi: 10.1007/s11633-025-1546-4.

[9] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations*, 2022. Accessed: Sep. 07, 2025. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[10] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 10088–10115. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf

[11] S. Liu *et al.*, "Judge as A Judge: Improving the Evaluation of Retrieval-Augmented Generation through the Judge-Consistency of Large Language Models," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 5788–5807. doi: 10.18653/v1/2025.findings-acl.301.

[12] G. H. Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang, "Humans or LLMs as the Judge? A Study on Judgement Bias," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8301–8327. doi: 10.18653/v1/2024.emnlp-main.474.

[13] R. Y. Pratama and Supatman, "Pengembangan Sistem Chatbot Cerdas Berbasis Natural Language Processing (NLP) untuk Peningkatan Layanan Informasi Hotel," *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, vol. 14, no. 1, pp. 1350–1362, 2025, doi: 10.23960/jitet.v14i1.8528.

[14] W.-L. Chiang *et al.*, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," in *Proceedings of the 41st International Conference on Machine Learning*, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., Vienna, Austria: JMLR.org, 2024, pp. 8359–8388. [Online]. Available: https://proceedings.mlr.press/v235/chiang24b.html

[15] A. Malik, S. Mayhew, C. Piech, and K. Bicknell, "From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15670–15693. doi: 10.18653/v1/2024.findings-acl.926.

[16] Daniel Han, Michael Han, and Team, "Unsloth," 2023. Accessed: Feb. 14, 2024. [Online]. Available: https://github.com/unslothai/unsloth