LexIndoLLM: Large Language Model untuk Konsultasi Regulasi Daerah di Indonesia

Novianto Rahmadi[1], Arief Setyanto[2]
Teknik Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta, Indonesia
Email: [1]novay@students.amikom.ac.id, [2]arief_s@amikom.ac.id

Abstract. *Large Language Models (LLMs) have the potential to improve access to local regulatory consultation services, yet general-purpose models often produce inaccurate responses when handling Indonesian legal documents that are lengthy, formal, and highly contextual. This study develops LexIndoLLM, a lightweight model based on Llama 3.2-1B, through staged fine-tuning on 393 local regulatory documents from Kutai Kartanegara Regency and the integration of FAISS-based Retrieval-Augmented Generation (RAG). The model was evaluated using RAGAS, perplexity, ROUGE-L, and inference efficiency metrics. The results show that the proposed approach improves answer quality, as indicated by a reduction in perplexity from 9.13 to 1.74, an increase in ROUGE-L from 0.2058 to 0.4429, and faithfulness and answer correctness scores of 0.77 and 0.66, respectively. The system maintains an average response time under 3.4 seconds, suitable for local deployment. These findings indicate that a lightweight model combined with retrieval is feasible for local regulatory consultation in resource-constrained environments.*
Keywords: *Large Language Model, domain fine-tuning, Retrieval-Augmented Generation, regional regulations, RAGAS*

Abstrak. *Large Language Model (LLM) berpotensi meningkatkan akses terhadap layanan konsultasi regulasi daerah, tetapi model generik masih sering menghasilkan jawaban yang kurang akurat pada dokumen hukum Indonesia yang panjang, formal, dan kontekstual. Penelitian ini mengembangkan LexIndoLLM, model ringan berbasis Llama 3.2-1B, melalui fine-tuning bertahap pada 393 dokumen regulasi Kabupaten Kutai Kartanegara dan integrasi Retrieval-Augmented Generation (RAG) berbasis FAISS. Evaluasi dilakukan menggunakan RAGAS, perplexity, ROUGE-L, dan metrik efisiensi inferensi. Hasil menunjukkan bahwa pendekatan yang diusulkan meningkatkan kualitas jawaban, ditandai dengan penurunan perplexity dari 9,13 menjadi 1,74, peningkatan ROUGE-L dari 0,2058 menjadi 0,4429, serta nilai faithfulness 0,77 dan answer correctness 0,66. Waktu respons rata-rata di bawah 3,4 detik sehingga cocok untuk deployment lokal. Temuan ini menunjukkan bahwa model ringan yang dipadukan dengan retrieval layak digunakan untuk konsultasi regulasi daerah pada lingkungan komputasi terbatas.*
Kata Kunci: *Large Language Model, fine-tuning domain, Retrieval-Augmented Generation, regulasi daerah, RAGAS*

1. Pendahuluan

Perkembangan teknologi *Large Language Model* (LLM) membuka peluang besar untuk meningkatkan akses terhadap layanan informasi hukum melalui antarmuka tanya jawab yang lebih alami, cepat, dan mudah digunakan. Kemampuan model ini dalam memahami dan menghasilkan bahasa alami menjadikannya relevan untuk mendukung proses pencarian, penjelasan, dan interpretasi dokumen hukum yang kompleks [1], [2]. Di Indonesia, kebutuhan tersebut semakin penting karena implementasi otonomi daerah telah menghasilkan jumlah regulasi daerah yang sangat besar dan terus bertambah. Hingga tahun 2025, Indonesia memiliki 514 kabupaten dan kota [3] yang secara kumulatif menghasilkan ribuan peraturan daerah, peraturan bupati, dan peraturan wali kota yang masih aktif. Kondisi ini menuntut adanya mekanisme konsultasi regulasi yang akurat, mudah diakses, dan mampu membantu masyarakat maupun aparatur daerah dalam menemukan aturan yang relevan sesuai konteks permasalahan.

Namun, penerapan LLM pada domain hukum Indonesia masih menghadapi tantangan. Bahasa hukum formal Indonesia yang panjang, sarat terminologi teknis, bernuansa interpretatif, dan memiliki struktur redaksional kontekstual sering menyebabkan model generik menghasilkan jawaban tidak tepat atau halusinasi [4], [5]. Pada regulasi daerah, masalah ini menjadi lebih kritis karena pertanyaan pengguna sering menuntut rujukan normatif yang spesifik. Temuan terbaru pada *legal question answering* juga menunjukkan bahwa halusinasi dan ketidakakuratan faktual masih menjadi tantangan utama, sehingga adaptasi domain dan evaluasi terarah diperlukan untuk meningkatkan keandalan model pada konteks hukum [6]. Kondisi ini penting bagi daerah seperti Kabupaten Kutai Kartanegara, terutama karena akses terhadap konsultan hukum profesional masih terbatas di wilayah pedalaman.

Penelitian sebelumnya pada domain hukum global menunjukkan potensi spesialisasi model, tetapi sebagian besar masih berfokus pada hukum umum berbahasa Inggris atau Cina, menggunakan model berparameter besar, kebutuhan komputasi tinggi, dan belum diarahkan pada karakter regulasi daerah Indonesia [2], [7], [8]. Di sisi lain, *Retrieval-Augmented Generation* (RAG) telah digunakan untuk mengurangi ketergantungan model pada pengetahuan parametrik dengan menambahkan konteks eksternal saat inferensi [9]. Pendekatan ini penting dalam domain hukum karena jawaban tidak cukup hanya relevan secara linguistik, tetapi juga harus terikat pada sumber hukum yang dapat ditelusuri. Di Indonesia, pemanfaatan *chatbot* untuk layanan informasi juga telah dikaji pada konteks organisasi profesional [10]. Namun, penelitian yang menggabungkan LLM, regulasi daerah, dan evaluasi sistematis masih terbatas [11], [12]. Dengan demikian, masih terdapat celah berupa belum tersedianya model *open-source* ringan yang dikembangkan khusus untuk konsultasi regulasi daerah Indonesia dan dievaluasi dari sisi kualitas jawaban, *grounding*, serta efisiensi inferensi.

Berdasarkan celah tersebut, penelitian ini mengembangkan LexIndoLLM sebagai model bahasa ringan berbasis Llama 3.2-1B yang diadaptasi melalui *fine-tuning domain* bertahap pada 393 regulasi Kabupaten Kutai Kartanegara periode 2020-2025. Pemilihan model kecil ini sejalan dengan penelitian sebelumnya yang memanfaatkan Llama-3.2-1B-Instruct terkuantisasi untuk *fine-tuning domain* hukum pada keterbatasan sumber daya komputasi [13]. Model ini kemudian diintegrasikan dengan *retrieval* berbasis FAISS untuk memperkuat *grounding* jawaban terhadap dokumen regulasi [14]. Kombinasi *fine-tuning* dan *retrieval* sesuai untuk domain dengan informasi yang jarang muncul dalam data pelatihan umum, karena *retrieval* membantu model mengakses informasi spesifik yang tidak tersimpan kuat dalam parameter model [15]. Penelitian ini bertujuan untuk: (1) membangun model LLM ringan khusus regulasi daerah Indonesia, (2) menguji kontribusi *retrieval* terhadap *grounding* jawaban, dan (3) mengevaluasi performa model menggunakan RAGAS, *perplexity*, *ROUGE-L*, *Time to First Token*, *latency*, serta *throughput*. Kontribusi utama penelitian ini adalah pengembangan model LexIndoLLM, integrasi *fine-tuning domain* dengan *retrieval* berbasis dokumen regulasi, dan evaluasi kuantitatif pada lingkungan komputasi terbatas.

2. Tinjauan Pustaka

Penelitian mengenai *Large Language Model* pada domain hukum menunjukkan bahwa adaptasi domain dapat meningkatkan kemampuan model dalam memahami terminologi hukum, struktur argumentasi, serta tugas spesifik seperti klasifikasi, ringkasan, dan tanya jawab hukum. SaulLM-7B [2] dikembangkan untuk korpus hukum berbahasa Inggris dan Prancis, sedangkan LawLLM [7] dan Lawma [8] menunjukkan bahwa *fine-tuning domain* dapat meningkatkan performa model pada tugas *reasoning*, anotasi, dan klasifikasi hukum. Meskipun demikian, sebagian besar penelitian tersebut masih berfokus pada sistem hukum non-Indonesia, menggunakan model berparameter besar, dan belum diarahkan pada karakter regulasi daerah Indonesia yang formal, kontekstual, serta menuntut efisiensi komputasi.

Selain adaptasi model, peningkatan kualitas sistem hukum berbasis LLM juga banyak dilakukan melalui integrasi RAG. Pendekatan ini memungkinkan model memperoleh konteks eksternal saat inferensi, sehingga jawaban lebih terikat pada sumber hukum yang dapat ditelusuri [9]. Seo et al. menunjukkan bahwa pendekatan RAG pada sistem tanya jawab hukum Korea yang

menggabungkan teks undang-undang dan preseden dapat mendukung penalaran hukum yang lebih kontekstual [16]. Martín Chozas et al. juga menunjukkan bahwa pengayaan terminologi dan *query expansion* dapat meningkatkan efektivitas *retrieval* pada korpus hukum Spanyol [17]. Temuan tersebut menegaskan bahwa kualitas *retrieval* hukum dipengaruhi oleh kesesuaian representasi *query*, istilah hukum, dan struktur dokumen.

Di Indonesia, penelitian terkait penerapan LLM pada bidang hukum masih relatif terbatas. Beberapa upaya awal telah dilakukan, seperti sistem tanya jawab berbasis RAG untuk akses informasi RANPERDA di Kabupaten Kampar [11] dan penerapan *Graph*-RAG pada Peraturan Daerah Provinsi Banten tentang pajak kendaraan bermotor [12]. Penelitian LawRAG juga menunjukkan bahwa strategi *chunking* yang mempertimbangkan struktur judul dokumen, pasal, dan ayat dapat meningkatkan akurasi jawaban RAG dibandingkan dengan *chunking* sekuensial biasa [18]. Namun, penelitian tersebut umumnya masih mengandalkan model generik, memiliki cakupan data terbatas, dan belum menghasilkan model *open-source* ringan yang diadaptasi secara khusus untuk regulasi daerah Indonesia.

Dari sisi metodologis, evaluasi sistem RAG dapat didukung oleh RAGAS yang menyediakan metrik *faithfulness*, *answer relevancy*, *context precision*, *context recall*, dan *answer correctness* untuk menilai kualitas jawaban dan *retrieval* [19]. Teknik *Parameter-Efficient Fine-Tuning* seperti LoRA dan QLoRA juga memungkinkan adaptasi domain dengan biaya komputasi lebih rendah [20], [21]. Pada konteks bahasa Indonesia, *transfer learning* menggunakan LoRA+ pada Llama 3.2 menunjukkan potensi adaptasi model ringan untuk meningkatkan kualitas respons percakapan berbahasa Indonesia [22]. Kombinasi *fine-tuning* efisien, *retrieval* berbasis dokumen hukum, dan evaluasi terstruktur menjadi landasan teknis bagi pengembangan model hukum ringan pada lingkungan komputasi terbatas.

Tabel 1 menyajikan perbandingan beberapa model LLM hukum dengan LexIndoLLM yang diusulkan dalam penelitian ini.

Tabel 1. Perbandingan Model LLM Hukum

Model	Parameter	Fokus Utama
SaulLM-7B [2]	7B	Hukum umum (Inggris dan Prancis)
LawLLM [7]	-	*Reasoning* hukum Cina
Lawma [8]	Kecil	Anotasi/klasifikasi hukum
LexIndoLLM**	1B	Regulasi daerah Indonesia

Catatan: **Usulan dari penelitian ini

Berdasarkan tinjauan pustaka tersebut, celah penelitian yang diidentifikasi adalah belum tersedianya model LLM ringan, *open-source*, dan secara khusus diadaptasi untuk regulasi daerah Indonesia yang berbahasa formal serta kontekstual. Penelitian ini mengisi celah tersebut melalui pengembangan model berbasis Llama 3.2-1B dengan *fine-tuning domain* bertahap, integrasi *retrieval* berbasis FAISS, serta evaluasi yang mencakup kualitas jawaban dan efisiensi inferensi.

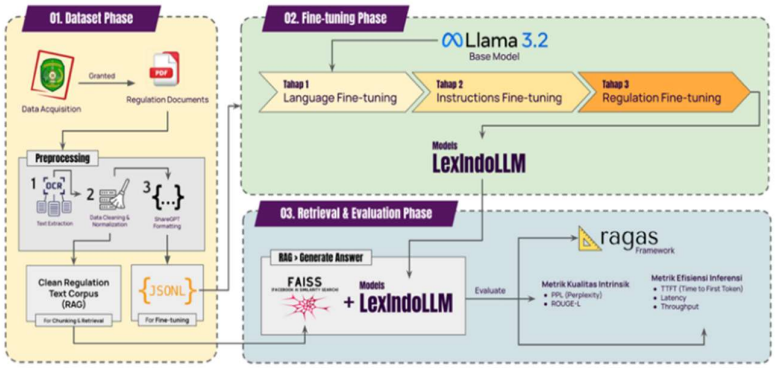
3. Metodologi Penelitian

Penelitian ini mengadopsi pendekatan *design science research* untuk menghasilkan artefak utama berupa model LLM khusus regulasi daerah yang dapat dijalankan pada lingkungan komputasi terbatas. Artefak yang dikembangkan adalah LexIndoLLM beserta mekanisme *retrieval* berbasis FAISS yang dievaluasi secara kuantitatif menggunakan metrik kualitas dan efisiensi.

3.1. Alur Penelitian

Pengembangan LexIndoLLM dalam penelitian ini dilakukan melalui tiga fase utama yang saling terhubung, sebagaimana ditunjukkan pada Gambar 1. Fase pertama adalah *Dataset Phase*, yang mencakup akuisisi dokumen regulasi daerah, ekstraksi teks dengan *Optical Character Recognition* (OCR), pembersihan data, dan transformasi data ke format yang sesuai untuk kebutuhan eksperimen. Fase kedua adalah *Fine-tuning Phase*, yaitu tahap adaptasi model dasar Llama 3.2-1B secara bertahap melalui *language fine-tuning*, *instruction fine-tuning*, dan

regulation fine-tuning hingga menghasilkan model LexIndoLLM. Fase ketiga adalah *Retrieval and Evaluation Phase*, yang mencakup pembentukan indeks FAISS, *retrieval* konteks regulasi, generasi jawaban oleh model, serta evaluasi keluaran menggunakan *framework* RAGAS dan metrik pendukung lainnya.



Gambar 1. Alur Pengembangan LexIndoLLM

Alur ini dirancang untuk mendukung tujuan penelitian secara langsung. Pada fase pertama, korpus regulasi dipersiapkan untuk kebutuhan *fine-tuning* dan *retrieval*. Pada fase kedua, model dasar diadaptasi secara bertahap agar sesuai dengan bahasa Indonesia, pola instruksi, dan domain regulasi daerah. Pada fase ketiga, performa LexIndoLLM dibandingkan pada konfigurasi tanpa *retrieval* dan dengan *retrieval* berbasis FAISS untuk mengukur dampak integrasi RAG secara kuantitatif.

3.2. *Dataset Phase*

3.2.1. Akuisisi Data

Data primer penelitian ini bersumber dari dokumen regulasi daerah Kabupaten Kutai Kartanegara periode 2020–2025 yang terdiri atas Peraturan Daerah (Perda) dan Peraturan Bupati (Perbup). Proses akuisisi diawali dengan pengumpulan 1.400 dokumen PDF, kemudian dilakukan kurasi ketat untuk memastikan kelengkapan dan kualitas data. Setiap dokumen yang dipertahankan harus memenuhi atribut minimum sebagaimana ditunjukkan pada Tabel 2.

Tabel 2. Atribut Dokumen Regulasi Daerah

No	Atribut	Keterangan
1	Jenis	Kategori dokumen (Perda dan Perbup)
2	Tahun	Tahun penerbitan dokumen
3	Judul	Nomor dan judul resmi dokumen sebagai identitas rujukan
4	Isi	Konten utama (pasal/ayat)
5	Status	Status keberlakuan (masih berlaku, perubahan dan dicabut)

Setelah proses kurasi, diperoleh 393 dokumen final dengan distribusi jumlah regulasi per tahun yang ditampilkan pada Tabel 3.

Tabel 3. Sebaran Regulasi Daerah

Tahun	Peraturan Bupati	Peraturan Daerah	Total
2020	61	13	74
2021	72	8	80
2022	42	5	47
2023	64	6	70
2024	47	16	63
2025	50	9	59
Total	**336**	**57**	**393**

3.2.2. *Pre-processing*

Tahap *pre-processing* dilakukan untuk mengubah dokumen regulasi daerah berbentuk PDF menjadi korpus teks yang bersih, terstruktur, dan siap digunakan dalam proses pemodelan. Proses ini dimulai dengan ekstraksi teks menggunakan *Optical Character Recognition* (OCR) pada dokumen yang belum memiliki lapisan teks digital, kemudian dilanjutkan dengan pembersihan data untuk menghapus elemen yang tidak relevan, seperti *header*, *footer*, nomor halaman, karakter hasil OCR yang keliru, spasi berlebih, dan simbol yang tidak diperlukan. Teks yang telah dibersihkan selanjutnya dinormalisasi agar memiliki format yang lebih konsisten. Untuk menjaga kualitas korpus, setiap dokumen juga melalui pemeriksaan manual terbatas agar hasil OCR dan pembersihan tidak menghilangkan substansi hukum yang penting.

Hasil *pre-processing* dibagi menjadi dua keluaran dengan fungsi berbeda. Keluaran pertama adalah *dataset* instruksi jawaban berformat ShareGPT JSONL yang digunakan untuk *fine-tuning* model. Keluaran kedua adalah teks regulasi terstruktur yang dipertahankan dalam bentuk dokumen hukum untuk kebutuhan *chunking*, *embedding*, dan indeksasi pada *pipeline retrieval* berbasis FAISS. Pemisahan ini penting karena *fine-tuning* membutuhkan data percakapan terstruktur, sedangkan *retrieval* membutuhkan representasi dokumen yang tetap mempertahankan konteks pasal, ayat, dan ketentuan hukum. *Dataset fine-tuning* akhir menghasilkan 11.190 pasangan tanya jawab, yang dibagi dengan rasio 80% untuk *training set*, 10% untuk *validation set*, dan 10% untuk *test set*, dengan total 1.119 entri pada data uji.

3.3. *Fine-tuning Phase*

Fine-tuning dilakukan secara bertahap dalam tiga tahap. Tahap pertama adalah *Language Fine-tuning*, yaitu adaptasi model terhadap kosakata dan terminologi hukum berbahasa Indonesia dengan menggunakan sumber terminologi hukum yang relevan untuk menyesuaikan model dengan karakter bahasa regulasi formal. Tahap kedua adalah *Instruction Fine-tuning*, yaitu pelatihan model agar mampu mengikuti pola instruksi dan format tanya jawab dalam bahasa Indonesia. Tahap ketiga adalah *Regulation Fine-tuning*, yaitu adaptasi spesifik pada korpus regulasi daerah Kabupaten Kutai Kartanegara yang disusun dalam penelitian ini.

Proses ini memanfaatkan teknik *Parameter-Efficient Fine-Tuning* (PEFT) dengan LoRA dan QLoRA serta kuantisasi 4-bit untuk menjaga efisiensi komputasi. Seluruh eksperimen dijalankan pada Google Colab menggunakan GPU NVIDIA Tesla T4 dengan memanfaatkan Hugging Face Transformers [23]. Proses *fine-tuning* dipercepat menggunakan kerangka pelatihan Unsloth. Unsloth dipilih karena mendukung proses *fine-tuning* yang lebih efisien pada lingkungan GPU terbatas, sehingga sesuai dengan skenario komputasi penelitian ini. Konfigurasi parameter pelatihan secara lengkap disajikan pada Tabel 4 dan Tabel 5.

Tabel 4. Konfigurasi Lingkungan Eksperimen

Komponen	Keterangan
Perangkat Komputasi	Google Colab
GPU	NVIDIA Tesla T4
Framework Pelatihan	Unsloth
Library Model	Hugging Face Transformers

Tabel 5. Konfigurasi LoRA/QLoRA

Parameter	Nilai
Model dasar	Llama 3.2-1B
Metode *fine-tuning*	LoRA / QLoRA
Kuantisasi	4-bit
Konteks Maksimal	2048 token
LoRA r / α / dropout	16 / 32 / 0.05
Batch size	16
Learning rate	2×10^{-4}
Epoch	2
Optimizer	AdamW 8-bit

3.4. *Retrieval and Evaluation Phase*

Retrieval and Evaluation Phase merupakan tahap akhir dalam pengembangan LexIndoLLM. Pada tahap ini, model diuji pada dua konfigurasi, yaitu tanpa *retrieval* dan dengan *retrieval* berbasis FAISS, untuk menilai pengaruh penambahan konteks regulasi terhadap kualitas, ketepatan, dan keterlacakan jawaban. Selain membangun *pipeline retrieval* untuk mengambil potongan dokumen regulasi yang relevan, tahap ini juga mencakup evaluasi keluaran model menggunakan *framework* RAGAS serta metrik kualitas dan efisiensi lainnya. Dengan demikian, *retrieval* dan evaluasi diposisikan sebagai satu rangkaian eksperimen untuk mengukur dampak integrasi RAG terhadap performa LexIndoLLM.

3.4.1. *Retrieval* Berbasis FAISS

Retrieval-Augmented Generation diimplementasikan menggunakan FAISS sebagai indeks vektor lokal untuk pencarian kemiripan semantik [14]. Dokumen regulasi di-*chunk* dengan ukuran 512 token dan *overlap* 128 token. Representasi vektor (*embedding*) dihasilkan menggunakan model multilingual-e5-large-instruct. Pemilihan model ini didasarkan pada kemampuan *embedding* berbasis instruksi dalam meningkatkan kualitas representasi semantik untuk tugas *retrieval* [24]. Selain itu, penggunaan *multilingual-e5-large-instruct* juga sejalan dengan penelitian RAG berbahasa Indonesia sebelumnya yang memanfaatkan model tersebut untuk mengindeks dokumen dan mendukung pencarian kesamaan semantik [25]. Pada tahap inferensi, sistem mengambil lima *chunk* paling relevan (*top*-5). Nilai *similarity threshold* 0,75 digunakan untuk menjaga keseimbangan antara relevansi konteks yang diambil dan ketersediaan konteks bagi proses generasi jawaban. Pendekatan ini memungkinkan jawaban model disertai rujukan pasal dan nomor regulasi yang relevan sebagai dasar hukum yang dapat ditelusuri.

3.4.2. Evaluasi Model

Evaluasi model dilakukan untuk menilai performa LexIndoLLM pada dua konfigurasi, yaitu tanpa *retrieval* dan dengan *retrieval* berbasis FAISS. Penilaian difokuskan pada tiga aspek utama, yaitu kualitas generasi jawaban, kualitas *retrieval* dan *grounding* jawaban, serta efisiensi inferensi. Dengan susunan ini, evaluasi tidak hanya menunjukkan kemampuan model dalam menghasilkan jawaban, tetapi juga menunjukkan keterkaitan jawaban dengan konteks regulasi yang digunakan serta kelayakan model pada lingkungan komputasi terbatas. Dari sisi kualitas generasi, penelitian ini menggunakan *Perplexity* dan *ROUGE-L*. *Perplexity* digunakan untuk mengukur tingkat ketidakpastian model dalam memprediksi token pada data evaluasi, sehingga nilai yang lebih rendah menunjukkan model lebih stabil dalam menghasilkan keluaran. Sementara itu, *ROUGE-L* digunakan untuk menilai tingkat kemiripan jawaban model terhadap jawaban referensi berdasarkan *longest common subsequence*. Kedua metrik ini digunakan untuk melihat peningkatan kualitas model secara intrinsik setelah proses *fine-tuning*.

Untuk mengevaluasi kualitas jawaban pada konfigurasi RAG, penelitian ini menggunakan RAGAS sebagai *framework* evaluasi utama. *Faithfulness* digunakan untuk menilai konsistensi faktual jawaban terhadap konteks hasil *retrieval*, sedangkan *Answer Relevancy* digunakan untuk menilai sejauh mana jawaban benar-benar menjawab pertanyaan pengguna. Kedua metrik ini penting karena sistem yang baik tidak hanya harus memberikan jawaban yang relevan, tetapi juga harus tetap terikat pada sumber regulasi yang digunakan sebagai konteks. Kualitas *retrieval* dievaluasi menggunakan *Context Precision* dan *Context Recall*. *Context Precision* mengukur ketepatan konteks yang diambil, sedangkan *Context Recall* mengukur kelengkapan informasi penting yang berhasil dicakup dalam *retrieved context*. Selain itu, *Answer Correctness* digunakan untuk menilai tingkat kebenaran jawaban model terhadap jawaban referensi berdasarkan kesamaan faktual. Perlu ditegaskan bahwa *Faithfulness*, *Context Precision*, dan *Context Recall* hanya dihitung pada konfigurasi dengan RAG karena ketiga metrik tersebut memerlukan *retrieved context* sebagai dasar evaluasi. Oleh karena itu, pada konfigurasi tanpa *retrieval*, nilai ketiga metrik tersebut dinyatakan sebagai N/A.

Selain kualitas jawaban, penelitian ini juga mengevaluasi efisiensi inferensi menggunakan *Time to First Token* (TTFT), *latency*, dan *throughput*. TTFT mengukur waktu

hingga token pertama diterima, *latency* mengukur total waktu hingga respons selesai dihasilkan, dan *throughput* menunjukkan jumlah token keluaran per satuan waktu. Ketiga metrik ini digunakan untuk menangkap *trade-off* antara kualitas jawaban dan biaya komputasi. Seluruh metrik dihitung pada *test set* yang sama untuk menjaga konsistensi perbandingan antara model dasar Llama 3.2-1B, LexIndoLLM setelah *fine-tuning*, dan LexIndoLLM dengan integrasi RAG.

4. Hasil dan Diskusi

4.1. Pra-seleksi *Base Model*

Untuk menentukan *base model* yang paling sesuai pada perangkat lokal dengan sumber daya terbatas, dilakukan pra-seleksi terhadap beberapa model LLM *open-source* berukuran kecil hingga sedang. Seluruh pengujian awal dijalankan pada lingkungan Google Colab menggunakan GPU Tesla T4 dengan konfigurasi yang sama seperti tahap *fine-tuning*. Hasil pra-seleksi disajikan pada Tabel 6.

Tabel 6. Hasil Pra-Seleksi *Base Model*

Model	TTFT (s)	*Latency* (s)	*Throughput* (t/s)	PPL
Qwen 3-0.6B	0.11	10.93	0.71	1.49
Qwen 3-1.7B	0.11	23.53	0.86	1.29
Llama 3.2-1B	**0.06**	**6.91**	**41.35**	1.40
Llama 3.2-3B	3.99	10.02	26.54	**1.16**
Nusantara-1.8B	0.09	12.83	29.83	1.54
Gemma 3-1B	0.18	58.86	8.70	2.17
Phi 3.5 Mini	3.30	20.65	24.79	1.26

Dari Tabel 6 terlihat bahwa Llama 3.2-1B memberikan keseimbangan terbaik antara efisiensi inferensi dan kualitas model awal. Model ini memiliki nilai *Time to First Token* (TTFT) dan *latency* yang rendah, *throughput* tertinggi pada kelompok model 1B, serta nilai *perplexity* yang tetap kompetitif. Oleh karena itu, Llama 3.2-1B dipilih sebagai *base model* untuk tahap *fine-tuning* berikutnya pada lingkungan komputasi terbatas.

4.2. Perbandingan *Base Model* dan LexIndoLLM

Proses *fine-tuning domain* bertahap menghasilkan peningkatan yang jelas pada metrik kualitas model, meskipun disertai penurunan efisiensi inferensi pada beberapa aspek. Perbandingan performa inferensi antara *base model* Llama 3.2-1B dan LexIndoLLM setelah *fine-tuning* ditunjukkan pada Tabel 7.

Tabel 7. Perbandingan Performa Inferensi

Metrik	Llama 3.2-1B	LexIndoLLM
Perplexity	9.13	**1.74**
TTFT (s)	**0.084**	0.135
Latency (s)	**1.57**	3.39
Throughput (t/s)	188.95	**86.49**

Perplexity menurun dari 9,13 menjadi 1,74, yang menunjukkan bahwa LexIndoLLM memiliki tingkat ketidakpastian yang lebih rendah pada data evaluasi regulasi daerah. Hasil ini mengindikasikan bahwa model menjadi lebih sesuai dengan karakter bahasa dan struktur dokumen pada domain yang dituju. Di sisi lain, *latency* meningkat dari 1,57 detik menjadi 3,39 detik dan *throughput* menurun dari 188,95 menjadi 86,49 token/s. Meskipun demikian, waktu respons tersebut masih berada pada kisaran yang dapat diterima untuk skenario konsultasi hukum pada perangkat lokal. Hasil ini menunjukkan adanya *trade-off* antara kualitas domain dan efisiensi inferensi. Perlu dicatat bahwa nilai *perplexity* pada Tabel 7 berasal dari evaluasi domain spesifik, sehingga tidak dibandingkan langsung dengan nilai pada Tabel 6 yang digunakan untuk pra-seleksi *base model*.

Tabel 8. Perbandingan Kualitas Jawaban

Metrik	Llama 3.2-1B	LexIndoLLM
ROUGE-L	0.2058	**0.4429**

Kualitas jawaban model setelah *fine-tuning* juga meningkat, sebagaimana ditunjukkan pada Tabel 8. Nilai *ROUGE-L* meningkat dari 0,2058 menjadi 0,4429, yang menunjukkan bahwa jawaban LexIndoLLM semakin selaras dengan jawaban referensi. Hasil ini mengindikasikan bahwa adaptasi domain tidak hanya menurunkan ketidakpastian model, tetapi juga meningkatkan kualitas generasi jawaban pada tugas konsultasi regulasi daerah.

4.3. Dampak Integrasi RAG

Integrasi RAG berbasis FAISS menunjukkan peningkatan pada beberapa aspek kualitas jawaban LexIndoLLM, terutama pada *grounding* jawaban terhadap konteks regulasi dan ketepatan substansi jawaban. Hasil evaluasi menggunakan RAGAS disajikan pada Tabel 9.

Tabel 9. Evaluasi RAGAS *Framework*

Metrik	Tanpa RAG	Dengan RAG	Perubahan (Δ)
Faithfulness	N/A	**0.7523**	N/A
Answer Relevancy	0.8712	0.8733	**+0.0021**
Context Precision	N/A	**0.7007**	N/A
Context Recall	N/A	**0.5035**	N/A
Answer Correctness	0.4310	0.6603	**+0.2293**

Catatan: *Faithfulness*, *Context Precision*, dan *Context Recall* hanya dihitung pada konfigurasi dengan RAG karena ketiga metrik tersebut memerlukan *retrieved context* sebagai dasar evaluasi. Oleh karena itu, nilai untuk konfigurasi tanpa RAG dinyatakan sebagai N/A.

Hasil pada Tabel 9 menunjukkan bahwa konfigurasi dengan RAG memperoleh nilai *Faithfulness* sebesar 0,7523. Selain itu, *Context Precision* sebesar 0,7007 menunjukkan bahwa *chunk* yang diambil oleh *retriever* cenderung relevan terhadap pertanyaan, sedangkan *Context Recall* sebesar 0,5035 menunjukkan bahwa sebagian informasi penting telah berhasil dicakup, walaupun aspek kelengkapan *retrieval* masih dapat ditingkatkan. Temuan ini menunjukkan bahwa integrasi FAISS berkontribusi terhadap proses *grounding* jawaban model pada sumber regulasi yang digunakan. Pada metrik yang dapat dibandingkan secara langsung, *Answer Correctness* meningkat dari 0,4310 menjadi 0,6603. Sebaliknya, *Answer Relevancy* hanya berubah secara marginal dari 0,8712 menjadi 0,8733, yang menunjukkan bahwa model pada dasarnya sudah cukup relevan dalam menjawab pertanyaan bahkan sebelum *retrieval* ditambahkan.

4.4. Diskusi

Hasil penelitian menunjukkan bahwa *fine-tuning* bertahap dan integrasi *retrieval* memberikan kontribusi yang saling melengkapi. *Fine-tuning* meningkatkan kemampuan model dalam menyesuaikan diri dengan bahasa, struktur, dan terminologi regulasi daerah, sebagaimana terlihat pada penurunan *perplexity* dari 9,13 menjadi 1,74 dan peningkatan *ROUGE-L* dari 0,2058 menjadi 0,4429. Sementara itu, RAG berbasis FAISS memperkuat *grounding* jawaban terhadap dokumen sumber, yang terlihat dari nilai *Faithfulness* sebesar 0,7523 dan peningkatan *Answer Correctness* dari 0,4310 menjadi 0,6603.

Peningkatan kualitas tersebut disertai *trade-off* pada efisiensi inferensi. Setelah *fine-tuning*, *latency* meningkat dan *throughput* menurun dibandingkan dengan *base model*. Namun, waktu respons masih berada pada kisaran yang dapat diterima untuk konsultasi hukum lokal. Hal ini menunjukkan bahwa LexIndoLLM tetap layak digunakan pada lingkungan komputasi terbatas, terutama ketika ketepatan jawaban lebih diprioritaskan daripada kecepatan respons.

Dari perspektif praktis, hasil penelitian ini menunjukkan potensi LexIndoLLM untuk mendukung layanan informasi regulasi daerah pada lingkungan komputasi terbatas. Nilai *Faithfulness* sebesar 0,7523 menunjukkan bahwa jawaban model secara umum cukup selaras dengan konteks regulasi yang digunakan pada saat inferensi, sehingga lebih terikat pada sumber

hukum yang dirujuk. Dengan model yang relatif ringan dan *retrieval* berbasis dokumen regulasi, pendekatan ini membuka peluang pengembangan layanan konsultasi yang lebih terarah, lebih terjangkau, dan lebih mudah direproduksi dalam konteks lokal. Meski demikian, penelitian ini masih terbatas pada korpus regulasi dari satu kabupaten dan evaluasi otomatis tanpa validasi langsung oleh pakar hukum daerah. Oleh karena itu, penelitian lanjutan perlu memperluas korpus, melibatkan ahli hukum, serta mengoptimalkan strategi *retrieval* dan pembaruan indeks agar sistem dapat mengikuti perubahan regulasi.

5. Kesimpulan dan Saran

Penelitian ini mengembangkan LexIndoLLM sebagai model bahasa ringan berbasis Llama 3.2-1B untuk konsultasi regulasi daerah di Indonesia. Model dikembangkan melalui *fine-tuning* bertahap pada 393 regulasi Kabupaten Kutai Kartanegara periode 2020-2025 dan diintegrasikan dengan *retrieval* berbasis FAISS. Hasil evaluasi menunjukkan bahwa *fine-tuning* meningkatkan kualitas model, sedangkan *retrieval* memperkuat *grounding* jawaban dan ketepatan substansi respons. Meskipun masih terbatas pada satu kabupaten dan belum melibatkan validasi langsung oleh pakar hukum daerah, temuan ini dapat menjadi dasar empiris bagi pengembangan layanan konsultasi regulasi daerah pada lingkungan komputasi terbatas. Pengembangan selanjutnya dapat diarahkan pada perluasan korpus, evaluasi pakar, optimasi *retrieval*, dan pembaruan indeks regulasi secara berkala.

6. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada pengelola dokumen regulasi daerah Kutai Kartanegara, dosen pembimbing, serta semua pihak yang telah memberikan masukan, dukungan teknis, dan saran selama proses penelitian dan pengembangan sistem ini.

Referensi

[1] T. B. Brown *et al.*, “Language models are few-shot learners,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, Jul. 2020.

[2] P. Colombo *et al.*, “SaulLM-7B: A Pioneering Large Language Model for Law,” *arXiv preprint arXiv:2403.03883*, 2024.

[3] Badan Pusat Statistik, *Statistik Indonesia 2025*. Jakarta, Indonesia: Badan Pusat Statistik, 2025.

[4] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, “Large Language Models in Law: A Survey,” *AI Open*, vol. 5, pp. 181–196, 2024, doi: 10.1016/j.aiopen.2024.09.002.

[5] Y. Wu, C. Wang, E. Gumusel, and X. Liu, “Knowledge-Infused Legal Wisdom: Navigating LLM Consultation through the Lens of Diagnostics and Positive-Unlabeled Reinforcement Learning,” in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 2024, pp. 15542–15555. doi: 10.18653/v1/2024.findings-acl.918.

[6] Y. Hu, L. Gan, W. Xiao, K. Kuang, and F. Wu, “Fine-tuning Large Language Models for Improving Factuality in Legal Question Answering,” in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 4410–4427.

[7] S. Yue *et al.*, “LawLLM: Intelligent Legal System with Legal Reasoning and Verifiable Retrieval,” in *The 29th International Conference on Database Systems for Advanced Applications (DASFAA 2024)*, 2024, pp. 304–321. doi: 10.1007/978-981-97-5569-1_19.

[8] R. Dominguez-Olmedo *et al.*, “Lawma: The Power of Specialization for Legal Annotation,” in *International Conference on Learning Representations (ICLR)*, 2025.

[9] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

[10] N. Setiawan, M. Akbar, and A. A. T. Susilo, “Pengembangan Chatbot Untuk Layanan Informasi Keanggotaan Guru Metode Support Vector Machine,” *Jurnal Buana Informatika*, vol. 16, no. 2, pp. 166–175, Oct. 2025.

[11] B. Arham and Sukasih, "Sistem Tanya Jawab Berbasis Artificial Intelligence untuk Akses Informasi RANPERDA di Kabupaten Kampar," *JEKIN - Jurnal Teknik Informatika*, vol. 5, no. 2, pp. 928–935, Aug. 2025, doi: 10.58794/jekin.v5i2.1632.

[12] A. Prihartono and A. U. Priantoro, "Graph RAG untuk memahami peraturan tentang pajak kendaraan bermotor di Provinsi Banten," *Al-Ihtiram: Multidisciplinary Journal of Counseling and Social Research*, vol. 4, no. 1, pp. 293–300, 2025.

[13] R. Al-Qaesm, M. Hendi, and B. Tantour, "Alkafi-llama3: fine-tuning LLMs for precise legal understanding in Palestine," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 107, Jun. 2025, doi: 10.1007/s44163-025-00313-w.

[14] M. Douze *et al.*, "The Faiss Library," *IEEE Trans. Big Data*, vol. 12, no. 2, pp. 346–361, Apr. 2026, doi: 10.1109/TBDATA.2025.3618474.

[15] H. Soudani, E. Kanoulas, and F. Hasibi, "Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge," in *SIGIR-AP 2024 - Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, Tokyo, Japan: Association for Computing Machinery, Inc, Dec. 2024, pp. 12–22. doi: 10.1145/3673791.3698415.

[16] K. Seo and T. Utsuro, "RAG based Question Answering of Korean Laws and Precedents," in *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, 2025, pp. 91–100. doi: 10.18653/v1/2025.fever-1.7.

[17] P. Martín-Chozas, P. Calleja, and C. R. Limón, "Terminology Enhanced Retrieval Augmented Generation for Spanish Legal Corpora," in *Proceedings of the 5th Conference on Language, Data and Knowledge*, 2025, pp. 147–152.

[18] A. Fadillah, N. Athahirah, and K.-T. Lai, "LawRAG: Indonesian legal document retrieval-augmented generation with specialized chunking and reranking strategies," *Data Technologies and Applications*, vol. 60, no. 2, pp. 330–347, Apr. 2026, doi: 10.1108/DTA-03-2025-0195.

[19] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.

[20] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2022.

[21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in *Advances in Neural Information Processing Systems*, 2023. doi: 10.5555/3666122.3666563.

[22] F. Kautsar *et al.*, "Transfer Learning Menggunakan LoRA+ pada Llama 3.2 untuk Percakapan Bahasa Indonesia," *Techno.Com*, vol. 24, no. 2, pp. 332–343, May 2025, doi: 10.62411/tc.v24i2.12508.

[23] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[24] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Improving Text Embeddings with Large Language Models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, 2024, pp. 11897–11916. doi: 10.18653/v1/2024.acl-long.642.

[25] H. Lijaya, P. Ho, and H. Santoso, "Comparative Analysis of RAG-Based Open-Source LLMs for Indonesian Banking Customer Service Optimization Using Simulated Data," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 14, no. 3, pp. 330–341, Jul. 2025, doi: 10.32736/sisfokom.v14i3.2383.