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

Analisis Komparatif Internalisasi Pengetahuan melalui *Supervised Fine-Tuning* dan Penalaran Kontekstual melalui *Retrieval-Augmented Fine-Tuning* pada *Large Language* Model untuk Peraturan Desa di Indonesia

Comparative Analysis of Knowledge Internalization through Supervised Fine-Tuning and Contextual Reasoning through Retrieval-Augmented Fine-Tuning in Large Language Models for Indonesian Village Regulations

[1]Oppir Hutapea*, [2]Syahrial Jeremia Sinaga, [3]Vivaldi Adventus Simangunsong, [4]Yohanssen Pratama*, [5]Adli Abdillah Nababan*
[1,2,3]Program Studi Rekayasa Perangkat Lunak, Fakultas Vokasi, Institut Teknologi Del, Laguboti, Toba, Sumatera Utara, Indonesia
[4]Tokyo International University, Institute for International Strategy and Emerging Technologies, Kawagoe, Japan
[5]Departemen Sistem Informasi, School of Information Systems, Universitas Bina Nusantara, Jakarta 11480, Indonesia
*e-mail: oppir.hutapea@del.ac.id, ypratama@tiu.ac.jp, adli.nababan@binus.ac.id

(***received***: 31 August 2026, ***revised***: 21 September 2026, ***accepted***: 22 September 2026)

Abstrak

Large language model (LLM) rentan menghasilkan halusinasi sehingga keandalannya pada tanya-jawab hukum berbahasa Indonesia belum terjamin, terutama untuk Peraturan Desa yang sebagian besar tersedia sebagai PDF hasil pemindaian. Penelitian ini menyumbangkan perbandingan terkontrol antara internalisasi pengetahuan melalui Supervised Fine-Tuning (SFT) dan penalaran kontekstual melalui Retrieval-Augmented Fine-Tuning (RAFT), masing-masing dengan LoRA dan full fine-tuning dari checkpoint Llama-3.1-8B-Instruct yang sama, serta alur pemrosesan dokumen dan pembuatan dataset tanya-jawab yang tertelusur untuk 55 Peraturan Desa Kabupaten Badung (50 hasil pemindaian dan lima born-digital). Kinerja diukur dengan ROUGE-L, METEOR, BERTScore, dan RAGAS. Pada epoch keempat, RAFT-LoRA mencapai skor agregat tertinggi (0,883), disusul RAFT-Full (0,881), sedangkan SFT terbaik mencapai 0,749. RAFT meningkatkan faithfulness dari 0,640 menjadi 0,872 dan context recall dari 0,637 menjadi 0,907. Struktur data pelatihan berpengaruh lebih besar daripada strategi pembaruan parameter, dan LoRA mempertahankan kinerja RAFT dengan biaya adaptasi yang lebih rendah. Implikasinya, pengetahuan hukum dapat dijaga pada dokumen eksternal yang diperbarui tanpa pelatihan ulang, sementara model dilatih agar menjawab berdasarkan bukti (grounding). Temuan ini masih terbatas pada satu kabupaten dan evaluasi otomatis, sehingga validasi ahli hukum diperlukan sebelum penerapan.

Kata kunci: halusinasi, large language model, tanya-jawab hukum, retrieval-augmented fine-tuning, supervised fine-tuning, peraturan desa

Abstract

Large language models (LLMs) are prone to hallucination, making their reliability in Indonesian-language legal question answering uncertain, particularly for Village Regulations, which are predominantly available as scanned PDF documents. This study provides a controlled comparison between knowledge internalization through Supervised Fine-Tuning (SFT) and contextual reasoning through Retrieval-Augmented Fine-Tuning (RAFT), using both LoRA and full fine-tuning from the same Llama-3.1-8B-Instruct checkpoint. The study also develops a traceable document-processing pipeline and a question-answer dataset derived from 55 Village Regulations from Badung Regency,

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

consisting of 50 scanned and five born-digital documents. Model performance was evaluated using ROUGE-L, METEOR, BERTScore, and RAGAS. At the fourth epoch, RAFT-LoRA achieved the highest aggregate score (0.883), followed by RAFT-Full (0.881), while the best-performing SFT configuration reached 0.749. RAFT improved faithfulness from 0.640 to 0.872 and context recall from 0.637 to 0.907. The results indicate that training data structure had a greater impact on performance than the parameter-update strategy, while LoRA maintained RAFT performance with lower adaptation costs. These findings suggest that legal knowledge can be maintained in updatable external documents without requiring model retraining, while the model can be trained to generate evidence-grounded responses. However, the findings are limited to a single regency and automated evaluation; therefore, validation by legal experts is required before deployment.

1. Pendahuluan

Large Language Models (LLM) merupakan model bahasa berbasis arsitektur Transformer. Model ini dilatih menggunakan korpus teks dalam jumlah sangat besar dan memiliki miliaran hingga triliunan parameter, sehingga mampu memahami dan menghasilkan bahasa alami dengan cara yang menyerupai komunikasi manusia. Kemampuan tersebut bersumber dari pralatihan berskala besar dan arsitektur Transformer, khususnya mekanisme self-attention, yang secara efisien memodelkan hubungan kontekstual jarak jauh [1], [2]. Namun, kefasihan berbahasa saat ini tidak dapat dijadikan bukti keandalan faktual. LLM dapat menghasilkan jawaban yang tersusun rapi, umum, dan sesuai gaya bahasa, tetapi tidak didukung oleh sumber materi atau bahkan bertentangan dengan aturan yang berlaku fenomena yang umum disebut "halusinasi" [3], [4]. Dalam konteks hukum, kegagalan ini memiliki konsekuensi yang lebih serius dibandingkan kesalahan editorial biasa karena jawaban yang dihasilkan dapat ditafsirkan sebagai penjelasan atas hak, kewajiban, prosedur, sanksi, atau kewenangan kelembagaan. Oleh karena itu, dalam penelitian ini, kualitas LLM hukum harus andal dan dievaluasi tidak hanya berdasarkan kemiripan jawabannya dengan jawaban rujukan, tetapi juga apakah setiap jawaban dapat ditelusuri kembali ke dokumen yang mengaturnya.

Persoalan keandalan menjadi lebih kompleks pada teks hukum berbahasa Indonesia, karena bahasa Indonesia masih relatif kurang terwakili dalam korpus pelatihan multibahasa berskala besar, sementara dokumen hukum tingkat desa memuat terminologi khusus, ketentuan baku, frasa administratif yang berulang, serta fakta-fakta spesifik. Model yang di-fine-tuning menggunakan instruksi berbahasa Indonesia menunjukkan bahwa adaptasi terhadap bahasa tertentu sangat penting untuk meningkatkan kinerja pada tugas-tugas tingkat lanjut [5]. Akan tetapi, sebagian dokumen Peraturan Desa hanya tersedia dalam bentuk PDF hasil pemindaian beresolusi rendah, bukan teks yang dapat dibaca mesin. Karena itu, keseluruhan alur kerja menghadapi dua tantangan yang saling terkait. Pertama, model belum memiliki pengetahuan yang memadai tentang peraturan-peraturan tersebut. Kedua, teks yang diberikan kepada model mungkin sudah mengandung kesalahan pengenalan karakter atau masalah struktur dokumen. Jika tidak ditangani dengan tepat, kesalahan dari proses pembersihan dokumen hasil pemindaian, penataan halaman, atau segmentasi dokumen dapat memengaruhi proses pembuatan pertanyaan, pengambilan informasi, inferensi, dan evaluasi.

Peraturan Desa merupakan instrumen hukum yang mengikat, yang dibentuk oleh kepala desa bersama badan permusyawaratan desa. Peraturan ini diakui dalam sistem hukum Indonesia apabila ditetapkan berdasarkan kewenangan yang didelegasikan [6], [7]. Penyebarannya secara digital melalui jaringan dokumentasi hukum dimaksudkan untuk meningkatkan akses terhadap informasi hukum yang komprehensif dan akurat [8]. Namun, banyaknya arsip hasil pemindaian dan tingginya kemiripan leksikal antar-peraturan dari desa yang berbeda menciptakan lingkungan tanya-jawab yang kompleks: bagian-bagian yang secara semantik mirip dapat mengacu pada desa, tahun anggaran, pasal hukum, atau kewajiban administratif yang berbeda.

Dalam penelitian ini, kami membandingkan dua paradigma adaptasi yang menangani persoalan tersebut dari sudut pandang berbeda. Paradigma pertama adalah internalisasi pengetahuan (*knowledge internalization*) melalui Supervised Fine-Tuning (SFT), yaitu pelatihan model pralatih dengan pasangan instruksi-jawaban yang dipilih secara cermat sehingga pola jawaban spesifik domain dan asosiasi faktual terenkode ke dalam parameter model. Pengetahuan yang tersimpan dengan cara ini disebut pengetahuan parametrik. Namun, pengetahuan parametrik bersifat statis, sulit diperbarui secara selektif, dan mungkin tidak sepenuhnya dapat dipanggil kembali ketika pertanyaan berbeda

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

dari pola pelatihan yang telah diamati. Fine-tuning terhadap pengetahuan faktual baru juga dapat memicu halusinasi apabila fakta yang dipelajari tidak terepresentasikan secara kuat atau ketika model melakukan ekstrapolasi di luar bukti yang terkandung dalam dataset pelatihan [9], [10]. Definisi istilah yang digunakan secara konsisten dalam manuskrip ini dirangkum pada Bagian 2.A.

Paradigma kedua adalah penalaran kontekstual (*contextual reasoning*) melalui Retrieval-Augmented Fine-Tuning (RAFT). RAFT melatih model dalam kondisi yang menyerupai proses retrieval (pengambilan informasi), di mana sebuah pertanyaan disajikan bersama *oracle context* yang memuat jawaban, serta *distractor context* yang secara topik atau leksikal mirip tetapi secara hukum tidak relevan. Dengan demikian, tujuan pembelajaran ini menggabungkan pembuatan jawaban dengan diskriminasi bukti: model harus mengidentifikasi bagian yang mendukung jawaban, mengabaikan konteks yang menyesatkan, dan menghasilkan jawaban yang berpijak pada bukti yang tersedia (*grounding*) [11]. Perbedaan ini sangat relevan untuk korpus hukum, di mana dua ketentuan dapat menggunakan bahasa administratif yang hampir identik tetapi berlaku untuk desa, tahun, atau instrumen regulasi yang berbeda.

Penelitian ini mengisi kesenjangan tersebut dengan membandingkan dua pendekatan SFT dan RAFT menggunakan model Llama-3.1-8B-Instruct pada dataset Peraturan Desa yang dikumpulkan dari Kabupaten Badung. Kedua pendekatan dievaluasi menggunakan LoRA dan full-parameter fine-tuning untuk memisahkan pengaruh struktur data pelatihan dari pengaruh strategi pembaruan parameter. Penelitian ini menjawab dua pertanyaan utama: (i) apakah RAFT menghasilkan jawaban yang lebih akurat dan berbasis fakta dibandingkan SFT, dan (ii) apakah tujuan pelatihan berbasis retrieval dapat mengurangi halusinasi pada domain hukum Indonesia yang sumber dayanya terbatas.

Penelitian ini menetapkan tiga tujuan utama. Pertama, penelitian ini mengembangkan alur pemrosesan dokumen Peraturan Desa, mencakup berkas hasil pemindaian maupun berkas digital asli. Kedua, penelitian ini melakukan perbandingan terkontrol terhadap empat konfigurasi model SFT-LoRA, SFT-Full, RAFT-LoRA, dan RAFT-Full menggunakan model dasar yang sama. Ketiga, penelitian ini mengevaluasi jawaban yang dihasilkan pada tingkat leksikal, semantik, dan kontekstual menggunakan ROUGE-L, METEOR, BERTScore, dan RAGAS, alih-alih hanya mengandalkan satu metrik kemiripan tunggal.

Penelitian ini berbeda dari studi terdahulu dalam tiga hal. Pertama, RAFT [11] dievaluasi pada korpus berbahasa Inggris seperti PubMed, HotpotQA, dan dokumentasi API; kajian injeksi pengetahuan [10] membandingkan fine-tuning dengan retrieval tanpa melatih model membedakan bukti relevan dari pengecoh; studi AI hukum berbasis RAG [12] berfokus pada pengurangan halusinasi melalui retrieval; dan adaptasi bahasa Indonesia [5] tidak menyasar teks hukum tingkat desa. Kedua, sepengetahuan penulis, belum ada studi yang membandingkan SFT dan RAFT secara terkontrol dengan LoRA maupun full-parameter fine-tuning dari checkpoint yang sama pada dataset Peraturan Desa berbahasa Indonesia yang sebagian besar berasal dari dokumen hasil pemindaian. Ketiga, evaluasi tidak bertumpu pada satu metrik, melainkan menggabungkan metrik leksikal, semantik, dan berbasis konteks (RAGAS) dengan catatan interpretasi yang eksplisit untuk kondisi inferensi tanpa retrieval. Secara konkret, kontribusi penelitian ini adalah:

1) Desain eksperimen faktorial terkontrol (SFT/RAFT × LoRA/full fine-tuning × epoch × ukuran himpunan uji) yang memisahkan pengaruh struktur data pelatihan dari pengaruh strategi pembaruan parameter.
2) Alur pemrosesan Peraturan Desa klasifikasi dokumen born-digital dan hasil pemindaian, OCR, pembersihan, chunking berbasis struktur pasal, serta pembuatan pasangan tanya-jawab dengan validasi ground-truth dan metadata sumber yang dapat direplikasi untuk korpus hukum berbahasa Indonesia lainnya.
3) Bukti empiris bahwa pelatihan dengan oracle context dan distractor context meningkatkan faithfulness dari 0,640 menjadi 0,872 dan skor agregat terbaik dari 0,749 (SFT) menjadi 0,883 (RAFT) pada epoch keempat, serta bahwa struktur data pelatihan berpengaruh lebih besar daripada strategi pembaruan parameter.
4) Implikasi praktis: RAFT-LoRA mempertahankan kinerja RAFT dengan biaya adaptasi yang lebih rendah dan memungkinkan pembaruan pengetahuan hukum melalui indeks dokumen tanpa pelatihan ulang model.

Bagian 2 memuat definisi istilah dan penelitian terkait, Bagian 3 menguraikan metode, Bagian 4 menyajikan hasil dan pembahasan, dan Bagian 5 memuat kesimpulan.

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

2. Tinjauan Literatur

A. Definisi Istilah

Agar penggunaan istilah konsisten, definisi berikut dipakai di seluruh manuskrip.

Internalisasi pengetahuan (*knowledge internalization*): proses penyerapan fakta, terminologi, dan pola jawaban suatu domain ke dalam parameter model melalui pelatihan. Pengetahuan yang tersimpan dengan cara ini disebut pengetahuan parametrik. Pada penelitian ini, internalisasi pengetahuan dioperasionalkan melalui SFT.

Penalaran kontekstual (*contextual reasoning*): kemampuan model memilih bukti yang relevan dari beberapa konteks yang diberikan, mengabaikan konteks pengecoh, dan menurunkan jawaban dari bukti tersebut. Pada penelitian ini, penalaran kontekstual dioperasionalkan melalui RAFT.

Oracle context dan distractor context: *oracle context* adalah segmen peraturan yang memuat bukti jawaban, sedangkan *distractor context* adalah segmen yang mirip secara topik atau leksikal tetapi tidak memuat bukti yang diperlukan.

Closed-book dan open-book: dua kondisi inferensi. Pada kondisi *closed-book*, model menjawab tanpa langkah retrieval sehingga bergantung pada pengetahuan parametrik; pada kondisi *open-book*, model menjawab dengan bagian peraturan hasil retrieval (pengambilan informasi) yang disertakan pada prompt. Dalam penelitian ini, SFT dievaluasi pada kondisi closed-book dan RAFT pada kondisi open-book.

Grounding: sejauh mana klaim dalam jawaban dapat ditelusuri ke, dan didukung oleh, konteks yang diberikan. Grounding diukur melalui faithfulness, context recall, dan context precision.

Halusinasi: jawaban yang tidak didukung oleh sumber materi atau bertentangan dengan peraturan yang berlaku, meskipun tersusun fasih.

B. Penelitian Terkait dan Posisi Penelitian

Penelitian-penelitian terdahulu memperkenalkan Retrieval-Augmented Generation (RAG) sebagai pendekatan yang menggabungkan pengetahuan parametrik dengan memori eksternal non-parametrik [13]. Kajian injeksi pengetahuan yang membandingkan fine-tuning dengan retrieval menemukan bahwa retrieval umumnya lebih unggul dalam menambahkan pengetahuan faktual baru ke dalam LLM [10]. Terdapat pula penelitian yang berfokus pada RAFT, yang menunjukkan bahwa pelatihan model menggunakan dokumen oracle dan distractor dapat secara nyata meningkatkan kinerja pada pertanyaan terbuka di domain spesifik [11]. Selain itu, metode information retrieval terbukti efektif mengurangi halusinasi karena memberikan informasi eksternal yang dapat diverifikasi selama model menghasilkan jawaban [14], dan pendekatan berbasis RAG juga telah diusulkan untuk mengurangi halusinasi pada AI hukum [12].

Meskipun demikian, tidak dapat diasumsikan begitu saja bahwa seluruh pendekatan tersebut akan bekerja optimal untuk Peraturan Desa di Indonesia. Hal ini disebabkan dataset yang digunakan relatif kecil, banyak memuat pernyataan hukum yang repetitif, hanya berfokus pada satu wilayah tertentu, dan sebagian besar berasal dari dokumen hasil pemindaian. Oleh karena itu, diperlukan perbandingan yang cermat untuk mengetahui pendekatan mana yang lebih efektif dalam mengurangi halusinasi sekaligus menjaga konsistensi sumber, dengan menggunakan model dasar yang sama: apakah lebih baik melakukan internalisasi pengetahuan ke dalam parameter model (SFT), ataukah melatih penalaran kontekstual berbasis bukti (RAFT).

Penelitian ini juga perlu membedakan pengaruh pendekatan adaptasi dari strategi pembaruan parameter. SFT dan RAFT menentukan bagaimana data pelatihan disusun dan apa yang diharapkan dipelajari oleh model, sedangkan Low-Rank Adaptation (LoRA) dan penyempurnaan parameter penuh (full fine-tuning) menentukan bagaimana parameter model diperbarui selama pelatihan. Dengan memulai seluruh eksperimen dari model Llama-3.1-8B-Instruct sebagai baseline yang sama dan menerapkan kedua strategi pembaruan parameter pada kedua dataset, penelitian ini dapat mengidentifikasi apakah perbedaan kinerja terutama berasal dari tujuan pelatihan atau dari jumlah parameter yang dapat dilatih. Bagian akhir tinjauan ini menunjukkan bahwa penelitian terdahulu belum membandingkan secara terkontrol pengaruh struktur data pelatihan (SFT dibandingkan RAFT) dan strategi pembaruan parameter (LoRA dibandingkan full fine-tuning) pada domain hukum desa berbahasa Indonesia yang datanya terbatas dan sebagian besar berasal dari dokumen hasil pemindaian celah inilah yang menjadi fokus utama penelitian ini.

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

3. Metode Penelitian

A. Desain Eksperimen

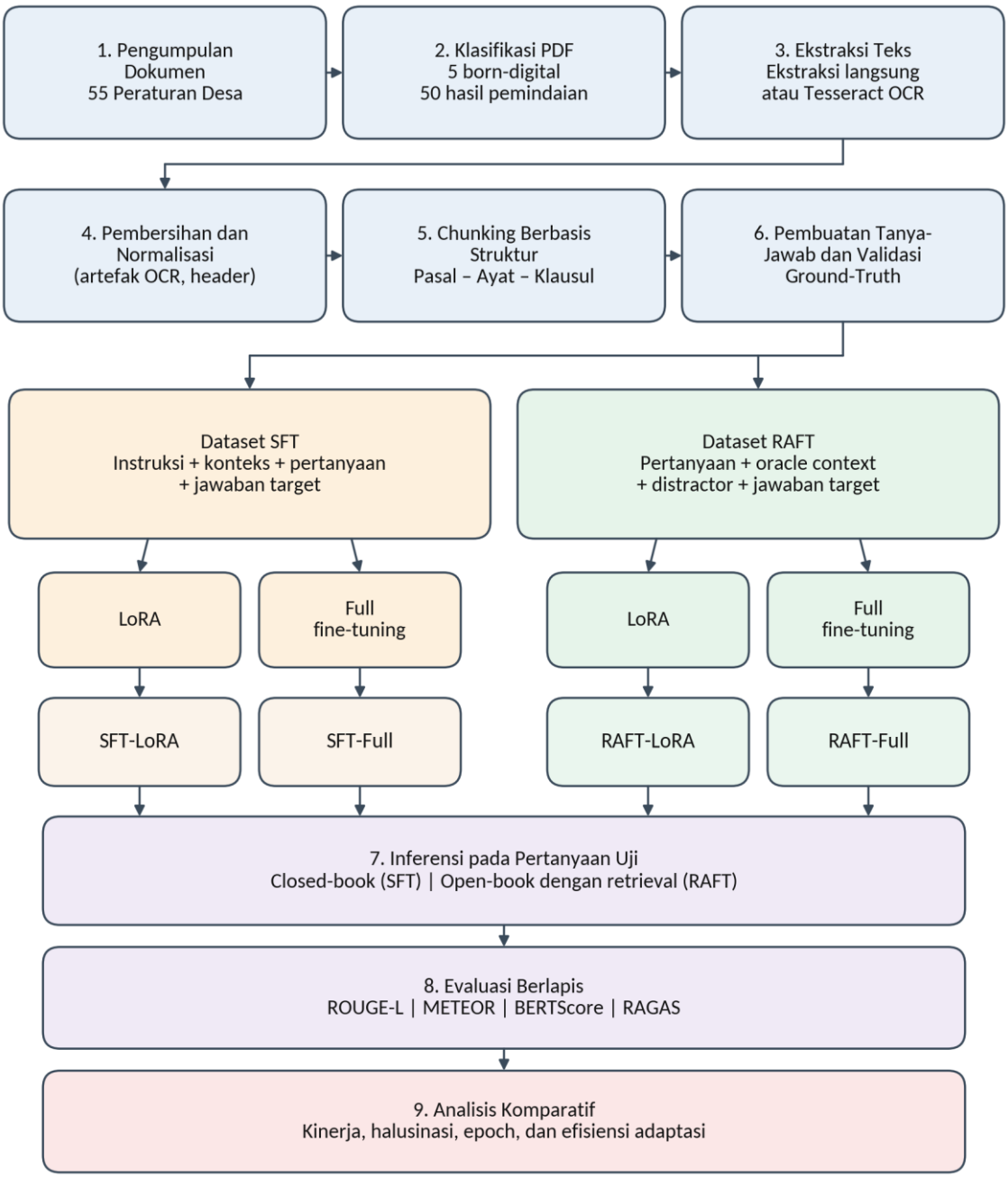
Dalam penelitian ini, kami menggunakan desain eksperimental kuantitatif, dengan metode adaptasi dan prosedur pembaruan parameter sebagai variabel bebas utama. Metode adaptasi yang digunakan meliputi SFT dan RAFT, sedangkan prosedur pembaruan parameter terdiri atas LoRA dan penyempurnaan parameter penuh. Variabel terikatnya adalah skor evaluasi leksikal, semantik, dan RAGAS yang diperoleh dari pertanyaan yang disisihkan untuk pengujian. Untuk menjaga validitas internal, setiap konfigurasi model dimulai dari model dasar yang sama, yaitu Llama-3.1-8B-Instruct, menggunakan dataset hukum yang sama, prosedur pemisahan dokumen yang sama, dan diuji pada dataset pengujian yang sama. Pengaturan ini memastikan bahwa perbedaan kinerja berasal dari pendekatan pelatihan dan strategi pembaruan parameter, bukan dari perbedaan model dasar atau inkonsistensi data evaluasi.

Kami memilih Llama-3.1-8B-Instruct karena, sebagaimana ditunjukkan pada beberapa penelitian, model ini memiliki arsitektur decoder-only dengan sekitar delapan miliar parameter, kemampuan multibahasa, dan kemampuan memodelkan jendela konteks yang panjang, sekaligus tetap sesuai untuk eksperimen dengan sumber daya terbatas [15]. Varian "instruct" juga menyediakan inisialisasi yang sesuai untuk sistem tanya-jawab berbasis prompt, karena telah di-fine-tuning untuk mengikuti instruksi dalam bahasa alami. Dalam penelitian ini, model-model tersebut tidak terlebih dahulu diubah menjadi model SFT lalu disempurnakan dengan LoRA. Sebaliknya, setiap cabang eksperimen dimulai secara independen dari checkpoint dasar yang sama: dataset SFT dilatih menggunakan LoRA atau penyempurnaan penuh, dan dataset RAFT dilatih menggunakan LoRA atau penyempurnaan penuh.

Empat konfigurasi model dikembangkan: SFT-LoRA, SFT dengan full-parameter fine-tuning, RAFT-LoRA, dan RAFT dengan full-parameter fine-tuning. Perilaku model diamati pada epoch 1, 3, dan 4, serta pada ukuran evaluasi kumulatif sebanyak 10, 20, 30, 40, dan 50 pertanyaan. Karena pengamatan ini dilakukan berulang, analisis dapat berfokus tidak hanya pada skor akhir tetapi juga pada arah perubahan kinerja seiring berjalannya pelatihan dan semakin beragamnya himpunan evaluasi. Kombinasi dua strategi pembaruan parameter, tiga epoch yang diamati, dan lima ukuran pengujian menghasilkan 30 konfigurasi pada masing-masing paradigma.

Alur kerja lengkap diilustrasikan pada Gambar 1. Proses dimulai dengan akuisisi dokumen dan klasifikasi PDF, dilanjutkan dengan ekstraksi, pembersihan, normalisasi, pengelompokan struktural, dan konstruksi ground-truth, kemudian bercabang menjadi dataset SFT dan RAFT. Setiap cabang dilatih menggunakan LoRA dan full fine-tuning untuk menghasilkan empat model yang dapat dibandingkan secara langsung. Hasil inferensi kemudian dievaluasi menggunakan metrik leksikal, semantik, dan berorientasi grounding sebelum dilakukan analisis perbandingan.

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



Gambar 1 Alur penelitian

Langkah 1–6 memproses dokumen hingga menghasilkan pasangan tanya-jawab tervalidasi; alur kemudian bercabang menjadi dataset SFT (oranye) dan RAFT (hijau), masing-masing dilatih dengan LoRA atau full fine-tuning dari checkpoint yang sama sehingga menghasilkan empat model. Langkah 7–9 mencakup inferensi pada pertanyaan uji, evaluasi berlapis, dan analisis komparatif.

Desain ini juga membedakan antara variabel yang dikontrol dan variabel yang sengaja divariasikan. Cakupan korpus, model dasar, domain tugas, pertanyaan yang disisihkan, dan metrik evaluasi dijaga tetap konstan. Sebaliknya, struktur contoh pelatihan dan kapasitas model yang dapat dilatih divariasikan. Pemisahan ini penting karena skor RAFT yang lebih tinggi tidak dapat diinterpretasikan sebagai bukti penalaran kontekstual apabila kondisi RAFT sekaligus menggunakan model dasar yang berbeda, dataset pengujian yang lebih besar, atau prosedur evaluasi yang lebih menguntungkan.

Matriks faktorial lengkap dirangkum pada Tabel 1. Desain ini menghasilkan 30 konfigurasi untuk setiap paradigma adaptasi dan 60 konfigurasi yang dievaluasi secara keseluruhan, dengan tetap mempertahankan checkpoint dasar dan protokol evaluasi yang sama.

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

Tabel 1 Matriks skenario eksperimen

Faktor	Level / Pengaturan	Peran Eksperimental
Paradigma adaptasi	SFT; RAFT	Variabel bebas
Pembaruan parameter	LoRA; full fine-tuning	Variabel bebas
Epoch yang diamati	1; 3; 4	Checkpoint pelatihan
Ukuran uji kumulatif	10; 20; 30; 40; 50 pasangan QA	Skala himpunan evaluasi
Model dasar	Llama-3.1-8B-Instruct	Variabel terkontrol
Metrik evaluasi	ROUGE-L; METEOR; BERTScore F1; answer relevancy; faithfulness; context recall; context precision	Variabel terikat
Jumlah konfigurasi	30 per paradigma; 60 total	2×2×3×5 observasi

B. Pengumpulan Dataset

Dataset penelitian kami dikumpulkan dari situs Jaringan Dokumentasi dan Informasi Hukum Kabupaten Badung, dengan hanya menggunakan Peraturan Desa resmi dalam format PDF, sebagaimana lazim pada pembangunan korpus dokumen dari web [16]. Setelah proses pengumpulan data, sebanyak 55 peraturan dalam format PDF berhasil dipertahankan untuk dianalisis. Sebagaimana ditunjukkan pada Tabel 2, dokumen-dokumen tersebut dikelompokkan ke dalam lima kategori: Anggaran Pendapatan dan Belanja Desa (APBDes), Perubahan APBDes, Rencana Kerja Pemerintah Desa (RKPDes), laporan pertanggungjawaban, dan Kewenangan Desa.

Meskipun memiliki tujuan yang berbeda-beda, gaya penulisan hukum pada dokumen-dokumen tersebut sebagian besar serupa. Dokumen-dokumen ini memiliki struktur pasal yang mirip, menggunakan terminologi hukum yang konsisten, dan mengikuti kaidah penyusunan yang baku. Sebagian besar perbedaan hanya terbatas pada nama desa, nomor peraturan, tahun anggaran, anggaran, dan pejabat penandatangan. Karena kemiripan ini, dokumen dari kategori berbeda berfungsi baik sebagai distractor karena sangat mirip dengan peraturan yang relevan tetapi memiliki konteks hukum yang berbeda. Oleh karena itu, model retrieval harus mampu mengidentifikasi dokumen yang berkaitan erat, bukan sekadar mengenali perbedaan tekstual yang jelas.

Selama pengumpulan data, kami berfokus pada kualitas dokumen dan keabsahan hukum, bukan pada perolehan sampel dalam jumlah besar. Untuk memastikan sumber pengetahuan berasal dari teks hukum yang sah, kami hanya menyertakan Peraturan Desa resmi dan mengecualikan ringkasan tidak resmi, laporan berita, dan materi sekunder lainnya. Selain itu, kami membatasi dataset pada satu wilayah untuk kepentingan eksperimen, sehingga memungkinkan kami menguji efektivitas pendekatan adaptasi secara lebih terfokus.

Tabel 2 Komposisi korpus peraturan desa kabupaten badung

Kategori	Dokumen	Persentase
Anggaran Pendapatan dan Belanja Desa (APBDes)	6	10,9%
Perubahan APBDes	8	14,5%
Rencana Kerja Pemerintah Desa (RKPDes)	16	29,1%
Laporan Pertanggungjawaban	11	20,0%
Kewenangan Desa	14	25,5%
Total	55	100%

Setiap dokumen PDF diklasifikasikan sebagai dokumen digital asli atau dokumen hasil pemindaian. Klasifikasi ini didasarkan pada ada tidaknya lapisan teks yang dapat digunakan pada dokumen serta apakah karakter hasil ekstraksi membentuk teks hukum yang jelas dan konsisten. Lima dokumen diidentifikasi sebagai dokumen digital asli dan dapat diproses secara langsung, sedangkan 50 dokumen lainnya berupa hasil pemindaian berbasis gambar.

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

Klasifikasi ini penting karena kedua jenis dokumen memerlukan metode ekstraksi yang berbeda. Ekstraksi teks langsung dari halaman yang hanya berisi gambar akan menghasilkan keluaran kosong atau tidak dapat digunakan. Sebaliknya, penerapan Optical Character Recognition (OCR) pada dokumen digital asli yang bersih dapat menimbulkan kesalahan pengenalan yang tidak perlu. Oleh karena itu, klasifikasi ini berfungsi sebagai mekanisme perutean untuk menentukan metode ekstraksi yang paling sesuai pada tahap berikutnya.

C. OCR, Pembersihan, dan Normalisasi

Berkas PDF digital asli diproses menggunakan ekstraksi teks langsung, sedangkan berkas PDF hasil pemindaian diubah menjadi teks yang dapat dibaca mesin menggunakan Tesseract OCR. Proses dua jalur ini dirancang karena dokumen hasil pemindaian dan dokumen digital memerlukan metode ekstraksi teks yang berbeda [17], [18]. Seluruh 55 dokumen berhasil diproses menggunakan metode ekstraksi yang sesuai, menghasilkan tingkat ekstraksi dokumen sebesar 100%. Artinya, setiap dokumen menghasilkan teks yang dapat digunakan untuk pemrosesan lebih lanjut. Namun, hal ini tidak berarti setiap karakter dikenali dengan benar. Pada dokumen hasil pemindaian, tanda tangan, stempel, halaman miring, kualitas gambar rendah, kop surat dekoratif, dan gambar terkompresi masih dapat menyebabkan kata hilang, karakter salah, atau baris terputus. Oleh karena itu, masalah-masalah ini diperlakukan sebagai persoalan kualitas data dan ditangani pada tahap pembersihan (cleaning) dan normalisasi, bukan diabaikan setelah proses ekstraksi.

Tahap pembersihan menghilangkan header dan footer yang berulang, nomor halaman, spasi kosong berulang, artefak hasil pemindaian, serta pemenggalan baris yang memisahkan kalimat pada baris PDF. Kamus koreksi diterapkan untuk kesalahan OCR yang umum terjadi, termasuk kasus huruf "l" yang terbaca sebagai angka satu pada penomoran pasal. Normalisasi leksikal ringan menstandarkan ekspresi numerik serta representasi bab, pasal, ayat, dan klausul. Prosedur ini sengaja dirancang secara konvensional: artefak pemformatan dihilangkan, tetapi kata-kata yang bermakna hukum, penomoran, dan penanda hierarki tetap dipertahankan. Operasi ini penting karena elemen tata letak dan struktur hukum tertentu dapat salah diklasifikasikan sebagai konten substantif, sementara normalisasi yang berlebihan dapat mengubah makna hukum yang seharusnya dipelajari oleh model [19], [20].

Kontrol kualitas dilakukan pada tingkat unit struktural, bukan hanya pada tingkat dokumen secara keseluruhan. Unit kosong, fragmen duplikat, ketentuan yang tidak lengkap, header, footer, tanda tangan, dan elemen non-substantif lainnya dikeluarkan dari proses pembuatan dataset. Menjaga metadata sumber selama proses ini memastikan setiap segmen yang diterima tetap terhubung dengan dokumen sumber, rentang halaman, bab, bagian, ayat, dan klausulnya. Ketertelusuran ini diperlukan untuk validasi ground-truth di masa mendatang serta untuk menentukan apakah jawaban yang dihasilkan didukung oleh peraturan yang benar.

D. Chunking Berbasis Struktur

Peraturan dibagi berdasarkan struktur hukumnya, bukan berdasarkan batas karakter yang tetap. Pasal digunakan sebagai unit semantik utama karena setiap pasal umumnya memuat satu ketentuan regulasi yang lengkap dan koheren. Jika suatu pasal melebihi batas yang ditentukan, proses segmentasi dilanjutkan ke struktur tingkat yang lebih rendah, seperti ayat dan klausul berhuruf. Metode rekursif hanya digunakan ketika unit struktural yang lengkap masih terlalu panjang.

Ukuran maksimum setiap segmen ditetapkan sebesar 512 token, dengan overlap sebesar 50 token. Overlap ini membantu menjaga hubungan kontekstual ketika ketentuan yang saling berkaitan muncul di dekat batas antara dua segmen. Setiap segmen juga menyimpan metadata kontekstual, termasuk judul peraturan, bab, nomor pasal, rentang halaman, dan dokumen sumber.

Dengan mempertimbangkan struktur konten, pendekatan ini mengurangi risiko terpisahnya elemen hukum penting, seperti persyaratan, kewajiban, pengecualian, sanksi, dan pihak yang bertanggung jawab. Pada saat yang sama, pendekatan ini menghasilkan segmen yang sesuai untuk proses pengindeksan dan pencarian. Desain ini sejalan dengan penelitian terdahulu yang menyatakan bahwa proses chunking harus disesuaikan dengan struktur dokumen, bukan dipandang memiliki satu konfigurasi optimal yang berlaku universal [21].

Overlap ini tidak dimaksudkan untuk menduplikasi sebagian besar korpus, melainkan untuk menjaga informasi pada batas segmen ketika sebuah kalimat, daftar, atau ketentuan tambahan tidak dapat

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

ditampung sepenuhnya dalam satu chunk. Penyimpanan metadata tingkat bagian juga memungkinkan sistem membedakan bagian-bagian yang secara semantik mirip tetapi berasal dari instrumen hukum yang berbeda.
Perbedaan ini sangat penting pada cabang RAFT. Sebuah distractor harus cukup relevan dan meyakinkan untuk menguji model, tetapi tidak boleh memuat bukti yang diperlukan untuk menghasilkan jawaban target.

E. Pembuatan Pertanyaan-Jawaban dan Validasi Ground-Truth

Pasangan pertanyaan-jawaban dihasilkan secara otomatis menggunakan LLM untuk membuat dataset sintetis. Pendekatan ini dipilih karena anotasi manual terhadap seluruh korpus hukum spesifik domain akan sangat memakan waktu. Untuk setiap segmen teks yang memenuhi syarat, LLM menghasilkan pertanyaan yang jawabannya harus secara eksplisit didukung oleh segmen teks tersebut. Pada tahap berikutnya, sistem menghasilkan kandidat jawaban beserta bukti pendukungnya.
Tahap pascapemrosesan digunakan untuk menghapus data kosong, data duplikat, pertanyaan ambigu, jawaban yang tidak didukung oleh konteks, serta contoh yang buktinya tidak dapat ditelusuri kembali ke teks hukum asli. Oleh karena itu, ground truth tidak diperlakukan sebagai jawaban yang berdiri sendiri, melainkan direpresentasikan sebagai hubungan terstruktur antara konteks, pertanyaan, jawaban, bukti, dan metadata sumber. Penelitian ini mengikuti prinsip umum dalam penyusunan dataset tanya-jawab, di mana contoh yang dihasilkan harus disaring untuk memastikan konsistensi terhadap sumber dan fakta [22], [23], [24]. Pandangan praktisi juga menempatkan akurasi, kebersihan, dan konsistensi data pelatihan sebagai penentu kualitas LLM [25].
Validasi jawaban yang benar merupakan bagian penting dari proses ini karena data hukum dihasilkan secara otomatis. Jawaban sintetis mungkin tampak jelas dan meyakinkan, tetapi tetap dapat mengulang kesalahan yang terjadi selama proses OCR atau memuat informasi yang tidak dinyatakan secara eksplisit dalam peraturan. Karena itu, proses validasi berfokus pada apakah pertanyaan dapat dijawab berdasarkan konteks yang diberikan, apakah jawaban selaras dengan bukti pendukung, dan apakah rujukan hukum yang disebutkan spesifik dan akurat.
Contoh tidak diterima sebagai target evaluasi yang andal apabila bergantung pada pengetahuan hukum di luar konteks yang diberikan, menggunakan rujukan yang tidak jelas seperti "bagian terkait" yang tidak spesifik, atau memiliki metadata sumber yang tidak lengkap.
Data juga disusun berdasarkan dokumen sumber, bukan didistribusikan secara acak antar pasangan pertanyaan-jawaban. Pengorganisasian pada tingkat dokumen ini penting secara metodologis karena pertanyaan yang berasal dari segmen berbeda pada peraturan yang sama dapat memiliki kata, entitas bernama, nilai numerik, dan struktur hukum yang serupa.

F. Konstruksi Dataset SFT dan RAFT

Dataset SFT dibuat untuk mendorong internalisasi pengetahuan hukum ke dalam parameter model. Setiap contoh pelatihan terdiri atas instruksi, konteks hukum, pertanyaan, dan jawaban rujukan. Selama pelatihan, loss hanya dihitung pada token jawaban, sedangkan instruksi dan konteks berfungsi sebagai input. Model kemudian secara bertahap menginternalisasi istilah hukum, gaya jawaban, dan pengetahuan faktual dari Peraturan Desa sebagai pengetahuan parametrik. Namun, SFT tidak secara khusus mengajarkan model untuk membedakan bukti yang relevan dari yang tidak relevan; tujuan pembelajarannya adalah memetakan input yang diberikan ke jawaban rujukan.
Sebaliknya, dataset RAFT menekankan pada penalaran kontekstual. Setiap contoh pertanyaan memiliki satu oracle context yang memuat informasi yang benar dan satu atau lebih distractor context yang membahas topik hukum serupa. Beberapa contoh juga sengaja menghilangkan informasi tertentu untuk mensimulasikan kegagalan proses pengambilan informasi. Pengaturan ini mendorong model untuk hanya memercayai bukti yang benar-benar mendukung jawaban, sekaligus mengabaikan informasi yang terkait tetapi tidak relevan atau kontradiktif, sehingga lebih mencerminkan situasi pengambilan jawaban di dunia nyata [11]. Pendekatan ini berbeda dari penyempurnaan instruksi konvensional, yang umumnya melatih model hanya menggunakan contoh positif tanpa secara eksplisit menyertakan bukti negatif atau menyesatkan.
Oleh karena itu, perbedaan utama antara dataset SFT dan RAFT terletak pada perilaku pembelajaran yang didorong, bukan pada model bahasa yang mendasarinya. Kedua pendekatan menggunakan Llama-3.1-8B-Instruct sebagai model dasar pralatih yang sama. SFT berfokus pada peningkatan

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

kemampuan model menghasilkan jawaban spesifik domain melalui internalisasi pengetahuan hukum. Sebaliknya, RAFT berfokus membantu model memilih bukti yang tepat dan menarik kesimpulan darinya dalam kondisi open-book (Bagian 2.A).
Apabila RAFT menghasilkan performa yang lebih baik, hal ini dapat mengindikasikan bahwa supervisi yang berorientasi pada pengambilan informasi lebih efektif untuk menjawab pertanyaan hukum dibandingkan hanya mengandalkan internalisasi pengetahuan.

G. Fine-Tuning

Pada eksperimen LoRA, bobot model pralatih dibekukan, dan hanya adapter low-rank yang ditambahkan pada modul q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, dan down_proj yang dilatih. Eksperimen menggunakan rank 8, alpha 32, dropout 0,05, laju pembelajaran 2×10^{-4}, dan delapan langkah akumulasi gradien, sehingga menghasilkan ukuran batch efektif sebesar 16. Karena hanya lapisan adapter yang diperbarui, LoRA mengurangi jumlah parameter yang dapat dilatih dan penggunaan memori, sekaligus mempertahankan pengetahuan pralatih model [26], [27].
Sebaliknya, full-parameter fine-tuning memperbarui seluruh parameter model Llama. Pendekatan ini memberikan fleksibilitas yang lebih besar bagi model untuk belajar dari data pelatihan, tetapi juga membutuhkan sumber daya komputasi yang jauh lebih tinggi. Kedua strategi pembaruan diterapkan secara independen pada dataset SFT maupun RAFT. LoRA tidak dilakukan setelah tahap SFT penuh selesai; sebaliknya, SFT-LoRA berarti pelatihan checkpoint dasar pada dataset SFT dengan hanya parameter LoRA yang diperbarui. Demikian pula, RAFT-LoRA berarti pelatihan checkpoint dasar yang sama pada dataset RAFT melalui LoRA. Varian full fine-tuning dimulai dari checkpoint yang sama tetapi mengizinkan seluruh parameter menerima gradien. Desain paralel ini menjaga keterbandingan dan menghindari kerancuan akibat kelanjutan fine-tuning dari model yang sebelumnya telah diadaptasi.
Pada tahap inferensi, SFT dijalankan pada kondisi closed-book, yaitu jawaban dihasilkan terutama melalui perilaku prompt-response hasil adaptasi tanpa langkah retrieval, sedangkan RAFT dijalankan pada kondisi open-book di mana pertanyaan dikaitkan dengan bagian hukum hasil retrieval sebelum jawaban dihasilkan. Untuk evaluasi RAGAS pada SFT, konteks peraturan asal diberikan sebagai referensi evaluasi agar dukungan terhadap klaim yang dihasilkan dapat dinilai. Oleh karena itu, skor context recall dan context precision pada SFT harus dimaknai sebagai keselarasan dengan konteks referensi yang diberikan, bukan sebagai kinerja retriever SFT yang sesungguhnya. Inferensi RAFT lebih langsung mencerminkan penerapan retrieval-augmented: chunk yang relevan diambil dan diberikan kepada model, yang diharapkan memilih bukti pendukung dan mengabaikan distractor. Seed acak dan lingkungan komputasi yang sama digunakan pada seluruh proses yang dapat dibandingkan sesuai dokumen sumber.

H. Prosedur Evaluasi

Evaluasi dilakukan menggunakan tiga kelompok metric leksikal (ROUGE-L dan METEOR), semantik (BERTScore), dan berbasis konteks (RAGAS) karena satu metrik saja tidak cukup untuk mengukur kualitas jawaban atas pertanyaan hukum secara menyeluruh [28]. ROUGE-L mengukur kemiripan leksikal dan struktural berdasarkan longest matching subsequence antara jawaban yang dihasilkan dan jawaban rujukan [29]. Skor ROUGE-L yang tinggi menunjukkan bahwa jawaban yang dihasilkan mengikuti urutan kata dan struktur kalimat jawaban rujukan, tetapi dapat memberikan skor lebih rendah pada jawaban yang benar namun ditulis dengan susunan kata berbeda. METEOR menyediakan pendekatan yang lebih fleksibel dalam pencocokan kata dengan menyertakan penalti fragmentasi, sehingga mampu menangani variasi bentuk kata dan urutan kata secara lebih baik dibandingkan metode evaluasi yang hanya mengandalkan pencocokan n-gram [30]. BERTScore mengukur kemiripan semantik dengan membandingkan contextual token embeddings, sehingga mampu mengenali jawaban yang benar meskipun menggunakan frasa yang berbeda [31]. Secara bersama-sama, ROUGE-L, METEOR, dan BERTScore mengevaluasi kemiripan tekstual dan semantik terhadap jawaban rujukan, bukan hanya mengandalkan satu jenis pengukuran.
RAGAS digunakan untuk menganalisis dimensi yang berkaitan dengan halusinasi dan hasil generasi berbasis retrieval [32]. Sebagai contoh, answer relevancy menilai seberapa baik jawaban menjawab pertanyaan pengguna. Selanjutnya, faithfulness mengukur apakah jawaban didukung oleh konteks yang diberikan, sedangkan context recall dan context precision menilai kelengkapan dan relevansi

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

bukti yang diambil. Seluruh metrik dihitung untuk setiap pasangan pertanyaan-jawaban kemudian dirata-ratakan untuk setiap konfigurasi eksperimen. Skor agregat pada Persamaan (17) diperoleh dengan merata-ratakan answer relevancy, faithfulness, context recall, context precision, ROUGE-L, METEOR, dan BERTScore F1. Meskipun skor agregat ini memberikan ringkasan kinerja secara keseluruhan, metrik individual tetap sangat penting untuk mengidentifikasi kelemahan spesifik, seperti perbedaan leksikal, kesalahan semantik, klaim yang tidak didukung, atau pengambilan bukti yang tidak efektif.

Pada persamaan berikut, X menyatakan jawaban yang dihasilkan model (kandidat) dan Y menyatakan jawaban rujukan, masing-masing dipandang sebagai barisan token. ROUGE-L mengevaluasi kemiripan keduanya berdasarkan longest common subsequence (LCS) [29]: recall dihitung pada Persamaan (1), precision pada Persamaan (2), dan rata-rata harmoniknya (F-measure), yang dilaporkan sebagai skor ROUGE-L, pada Persamaan (3).

$$R_{\mathrm{LCS}} = \frac{\mathrm{LCS}(X,Y)}{|Y|} \tag{1}$$

$$P_{\mathrm{LCS}} = \frac{\mathrm{LCS}(X,Y)}{|X|} \tag{2}$$

$$F_{\mathrm{LCS}} = \frac{(1+\beta^2)R_{\mathrm{LCS}}P_{\mathrm{LCS}}}{R_{\mathrm{LCS}} + \beta^2 P_{\mathrm{LCS}}} \tag{3}$$

Pada Persamaan (1)–(3), $|X|$ dan $|Y|$ adalah panjang (jumlah token) kandidat dan rujukan; LCS(X,Y) adalah panjang subsekuens bersama terpanjang antara keduanya; R_{LCS}, P_{LCS}, dan F_{LCS} berturut-turut adalah recall, precision, dan F-measure ROUGE-L; dan β adalah parameter yang mengatur bobot relatif recall terhadap precision ($\beta = 1$ memberi bobot seimbang).

METEOR melengkapi ROUGE-L dengan pencocokan unigram dan penalti fragmentasi [30]. Precision dan recall unigram dihitung pada Persamaan (4) dan (5), keduanya digabung menjadi rata-rata harmonik berbobot pada Persamaan (6), penalti fragmentasi dihitung pada Persamaan (7), dan skor METEOR akhir diperoleh pada Persamaan (8).

$$P = \frac{m}{|X|} \tag{4}$$

$$R = \frac{m}{|Y|} \tag{5}$$

$$F_{\mathrm{mean}} = \frac{10PR}{R + 9P} \tag{6}$$

$$\mathrm{Penalty} = 0{,}5\left(\frac{ch}{m}\right)^3 \tag{7}$$

$$\mathrm{METEOR} = (1 - \mathrm{Penalty}) \times F_{\mathrm{mean}} \tag{8}$$

Pada Persamaan (4)–(8), m adalah jumlah unigram kandidat yang cocok dengan rujukan; ch adalah jumlah chunk cocok yang berurutan (contiguous); P dan R adalah precision dan recall unigram; F_{mean} adalah rata-rata harmonik yang memberi bobot recall sembilan kali bobot precision; dan Penalty adalah penalti fragmentasi yang membesar ketika kecocokan tersebar pada banyak chunk yang tidak berurutan, sehingga metrik ini dapat membedakan keselarasan yang koheren dari kumpulan kata yang cocok namun tidak berurutan.

BERTScore mengevaluasi kemiripan pada tingkat makna dengan mencocokkan contextual token embeddings antara kandidat x dan rujukan y [31]. Precision, recall, dan F1 didefinisikan berturut-turut pada Persamaan (9), (10), dan (11).

$$P_{\mathrm{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{y_j \in y} x_i^T y_j \tag{9}$$

$$R_{\mathrm{BERT}} = \frac{1}{|y|}\sum_{y_j \in y} \max_{x_i \in x} x_i^T y_j \tag{10}$$

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

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}}R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{11}$$

Pada Persamaan (9)–(11), x_i adalah embedding kontekstual token ke-i pada kandidat x, y_j adalah embedding token ke-j pada rujukan y, dan $|x|$ serta $|y|$ adalah jumlah token masing-masing; $x_i^{\mathrm{T}}y_j$ adalah hasil kali dalam kedua vektor (setara kemiripan kosinus karena embedding dinormalisasi), dan operator maksimum memilih token yang paling mirip secara semantik pada barisan lawannya. P_{BERT}, R_{BERT}, dan F_{BERT} adalah precision, recall, dan F1 BERTScore. Evaluasi semantik menggunakan representasi IndoBERT sehingga kemiripan tingkat token mencerminkan karakteristik linguistik bahasa Indonesia, bukan hanya kesamaan kata secara permukaan.

Makalah RAGAS [32] menjabarkan perhitungan faithfulness, answer relevancy, dan context recall berbasis rasio serta kemiripan kosinus, sedangkan context precision dihitung sebagaimana implementasi pustaka RAGAS. Faithfulness dihitung pada Persamaan (12), answer relevancy pada Persamaan (13), context recall pada Persamaan (14), dan context precision pada Persamaan (15) dan (16).

$$\text{Faithfulness} = \frac{|V|}{|S|} \tag{12}$$

$$\text{Answer relevancy} = \frac{1}{n}\sum_{i=1}^{n}\cos\big(E(q_i), E(q)\big) \tag{13}$$

$$\text{Context recall} = \frac{|G_{\text{ctx}}|}{|G|} \tag{14}$$

$$\text{Precision@k} = \frac{\text{TP@k}}{\text{TP@k} + \text{FP@k}} \tag{15}$$

$$\text{Context precision} = \frac{\sum_{k=1}^{K}\text{Precision@k} \times v_k}{\sum_{k=1}^{K} v_k} \tag{16}$$

Pada Persamaan (12)–(16), S adalah himpunan pernyataan (statement) yang diekstraksi dari jawaban dan V adalah himpunan bagian S yang dapat disimpulkan dari konteks; n adalah jumlah pertanyaan sintetis yang dibangkitkan dari jawaban, q_i adalah pertanyaan sintetis ke-i, q adalah pertanyaan asli, E adalah fungsi embedding, dan cos adalah kemiripan kosinus; G adalah himpunan pernyataan pada jawaban rujukan (ground truth) dan G_{ctx} adalah himpunan bagian G yang dapat diatribusikan pada konteks. Pada context precision, K adalah jumlah potongan konteks yang diperingkat, $v_k \in \{0,1\}$ adalah indikator relevansi potongan ke-k, serta TP@k dan FP@k adalah jumlah potongan relevan dan tidak relevan pada k teratas. Untuk SFT, konteks yang dinilai adalah konteks peraturan asal (Bagian 3.G).

Skor agregat S_{agr} pada Persamaan (17) adalah rata-rata tujuh skor metrik, dengan s_j adalah rata-rata skor metrik ke-j (answer relevancy, faithfulness, context recall, context precision, ROUGE-L, METEOR, dan BERTScore F1) pada seluruh pasangan tanya-jawab dalam satu konfigurasi eksperimen.

$$S_{\text{agr}} = \frac{1}{7}\sum_{j=1}^{7} s_j \tag{17}$$

Seluruh nilai metrik pertama kali dihitung untuk setiap pasangan tanya-jawab, kemudian dirata-ratakan dalam setiap konfigurasi eksperimen.

4. Hasil dan Pembahasan

A. Hasil Pemrosesan Dokumen

Proses klasifikasi dokumen mengidentifikasi lima berkas PDF yang dibuat secara digital sejak awal dan 50 berkas PDF hasil pemindaian. Seluruh dokumen berhasil diproses menggunakan metode ekstraksi masing-masing, menghasilkan lima keluaran melalui ekstraksi teks langsung dan 50 keluaran menggunakan optical character recognition (OCR). Hasil ini menunjukkan bahwa sebagian besar korpus bergantung pada OCR sebelum dapat digunakan dalam alur kerja model bahasa. Akibatnya, sistem tanya-jawab hukum yang dirancang untuk Peraturan Desa harus mampu menangani

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

dokumen hasil pemindaian, bukan mengasumsikan input sudah tersedia dalam bentuk teks digital yang bersih. Hal ini menegaskan pentingnya kualitas OCR, pelestarian hierarki dokumen hukum, dan penyimpanan metadata, karena faktor-faktor tersebut secara langsung memengaruhi akurasi ketentuan hukum, penomoran pasal, nama lembaga, nilai keuangan, dan rujukan sumber yang digunakan selama pelatihan dan ekstraksi data.

Meskipun seluruh dokumen berhasil melalui proses ekstraksi, keberhasilan ekstraksi tidak selalu menjamin transkripsi yang bebas kesalahan. OCR tetap dapat menghasilkan kesalahan pengenalan meskipun pemrosesan selesai tanpa gangguan. Dalam ranah hukum, kesalahan kecil pada nomor pasal, nilai nominal uang, kata negasi, atau nama instansi pemerintah dapat secara signifikan mengubah interpretasi bukti hukum. Oleh karena itu, pembersihan tambahan dan validasi struktural diterapkan setelah ekstraksi untuk meminimalkan perambatan kesalahan terkait OCR ke dalam dataset tanya-jawab yang dihasilkan, serta memastikan teks yang tidak akurat tidak diperlakukan sebagai informasi ground-truth yang andal.

B. Kinerja Leksikal dan Semantik

Tabel 3 menyajikan skor rata-rata leksikal, semantik, dan agregat untuk keempat konfigurasi model. Secara keseluruhan, RAFT menunjukkan kinerja yang lebih baik dibandingkan SFT pada skor agregat dan sebagian besar metrik individual; pengecualiannya adalah BERTScore F1 pada RAFT-Full (0,7651) yang lebih rendah daripada kedua model SFT. RAFT-Full mencapai skor ROUGE-L dan METEOR tertinggi, menunjukkan keselarasan yang lebih baik dengan jawaban rujukan dari segi pilihan kata dan urutan kalimat. Sementara itu, RAFT-LoRA mencapai skor BERTScore F1 tertinggi, menunjukkan kemiripan semantik yang lebih baik pada tingkat embedding. Kedua model RAFT juga mencapai skor agregat di atas 0,856, sedangkan model SFT tetap berada di kisaran 0,70.

Hasil ini menunjukkan bahwa pendekatan pelatihan RAFT memberikan pengaruh yang lebih besar terhadap kinerja dibandingkan perbedaan antara LoRA fine-tuning dan full-parameter training.

Temuan ini menunjukkan pentingnya mengevaluasi model dengan berbagai metrik. RAFT-Full mencapai skor ROUGE-L dan METEOR yang lebih tinggi, sedangkan RAFT-LoRA memperoleh BERTScore F1 tertinggi. Perbedaan ini menegaskan bahwa setiap metrik mengevaluasi aspek kualitas jawaban yang berbeda. Sebagian model menghasilkan jawaban yang sangat mendekati susunan kata jawaban rujukan, sementara model lain tetap mempertahankan makna hukum yang sama menggunakan ungkapan yang berbeda. Oleh karena itu, kinerja keseluruhan sebaiknya diinterpretasikan menggunakan skor agregat sekaligus metrik evaluasi individual.

Tabel 3 Rata-Rata kinerja per konfigurasi adaptasi

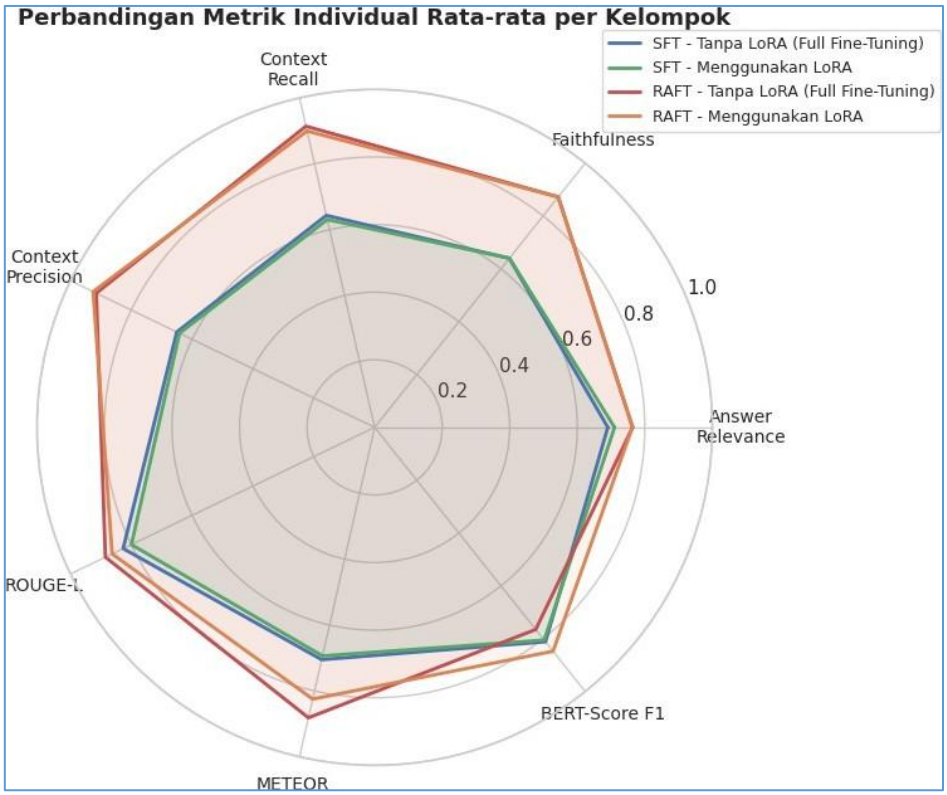
Konfigurasi	ROUGE-L	METEOR	BERTScore F1	Agregat
SFT-LoRA	0,7995	0,6937	0,8052	0,7030
SFT-Full	0,8267	0,7050	0,8115	0,7096
RAFT-LoRA	0,8628	0,8247	0,8474	0,8564
RAFT-Full	0,8850	0,8814	0,7651	0,8566

Pada model SFT, full-parameter fine-tuning mencapai skor yang sedikit lebih tinggi dibandingkan LoRA. Peningkatan tersebut sebesar 0,0272 untuk ROUGE-L, 0,0113 untuk METEOR, dan 0,0063 untuk BERTScore. Namun, skor agregat hanya meningkat sebesar 0,0066. Hal ini menunjukkan bahwa LoRA mempertahankan sebagian besar kinerja SFT dengan jauh lebih sedikit parameter yang dapat dilatih.

Pada RAFT, full fine-tuning meningkatkan skor ROUGE-L sebesar 0,0222 dan skor METEOR sebesar 0,0567 dibandingkan LoRA, tetapi skor BERTScore F1-nya 0,0823 lebih rendah. Selisih ini cukup besar, meskipun uji signifikansi statistik belum dilakukan dalam penelitian ini. Tingginya kesamaan leksikal menunjukkan bahwa RAFT-Full mereproduksi kata dan urutan rujukan dengan lebih akurat, sedangkan tingginya BERTScore pada RAFT-LoRA menunjukkan bahwa keluaran yang diparafrase tetap selaras secara semantik meskipun berbeda dalam struktur kalimat. Akibatnya, tidak ada satu metrik otomatis pun yang cukup untuk QA hukum: kemiripan leksikal, kemiripan semantik, dan dasar bukti harus diinterpretasikan secara bersama-sama.

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

Untuk kedua strategi pembaruan, skor agregat rata-rata adalah 0,8565 untuk RAFT dan 0,7063 untuk SFT, dengan selisih 0,1502. Peningkatan ini lebih besar dibandingkan perbedaan antara LoRA dan penyempurnaan penuh dalam masing-masing paradigma. Pola ini menunjukkan bahwa struktur contoh pelatihan pemetaan jawaban langsung pada SFT dibandingkan diskriminasi oracle dan distractor pada RAFT memiliki pengaruh yang lebih besar dibandingkan jumlah parameter yang diperbarui.

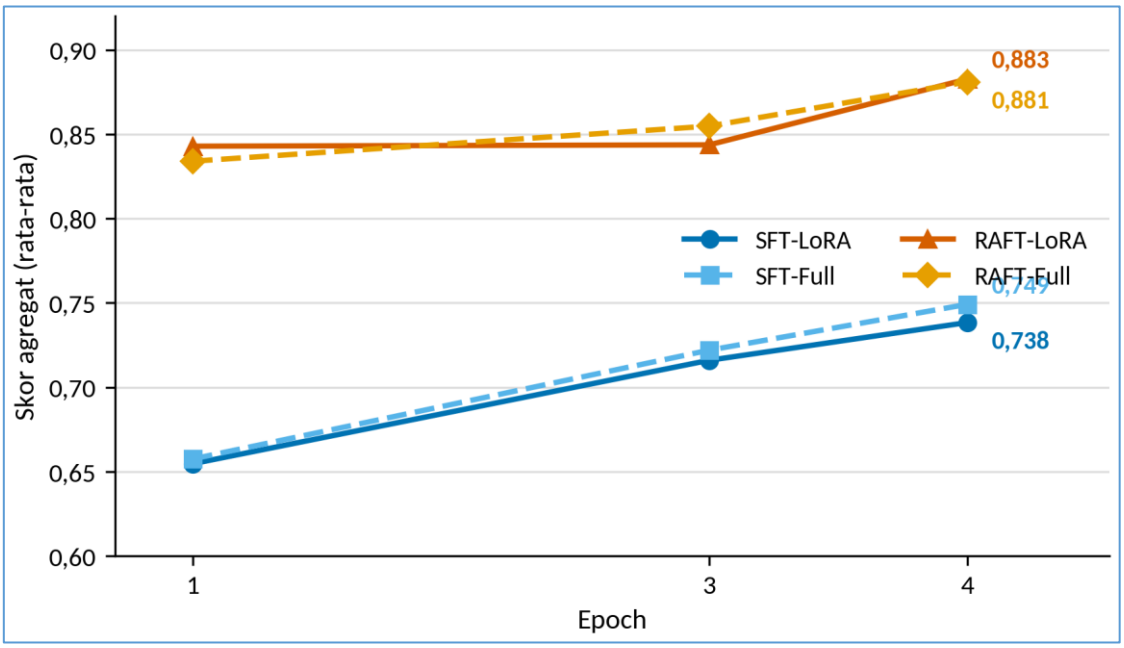


Gambar 2 Profil radar rata-rata tujuh metrik untuk empat konfigurasi.

Setiap sumbu adalah satu metrik berskala 0 - 1 (lingkaran dalam = 0,2; lingkaran luar = 1,0); semakin jauh titik dari pusat, semakin tinggi skornya. Warna biru menunjukkan SFT dan oranye menunjukkan RAFT; garis putus-putus menunjukkan full fine-tuning.

Profil radar pada Gambar 2 menunjukkan bahwa kedua konfigurasi RAFT mengungguli konfigurasi SFT pada hampir seluruh sumbu. Perbedaan paling menonjol terlihat pada context recall, context precision, faithfulness, dan METEOR, sedangkan selisih pada relevansi jawaban lebih kecil. Pola ini juga menegaskan bahwa keunggulan utama RAFT terletak pada pemanfaatan bukti kontekstual dan grounding, bukan sekadar menghasilkan jawaban yang relevan secara topik.

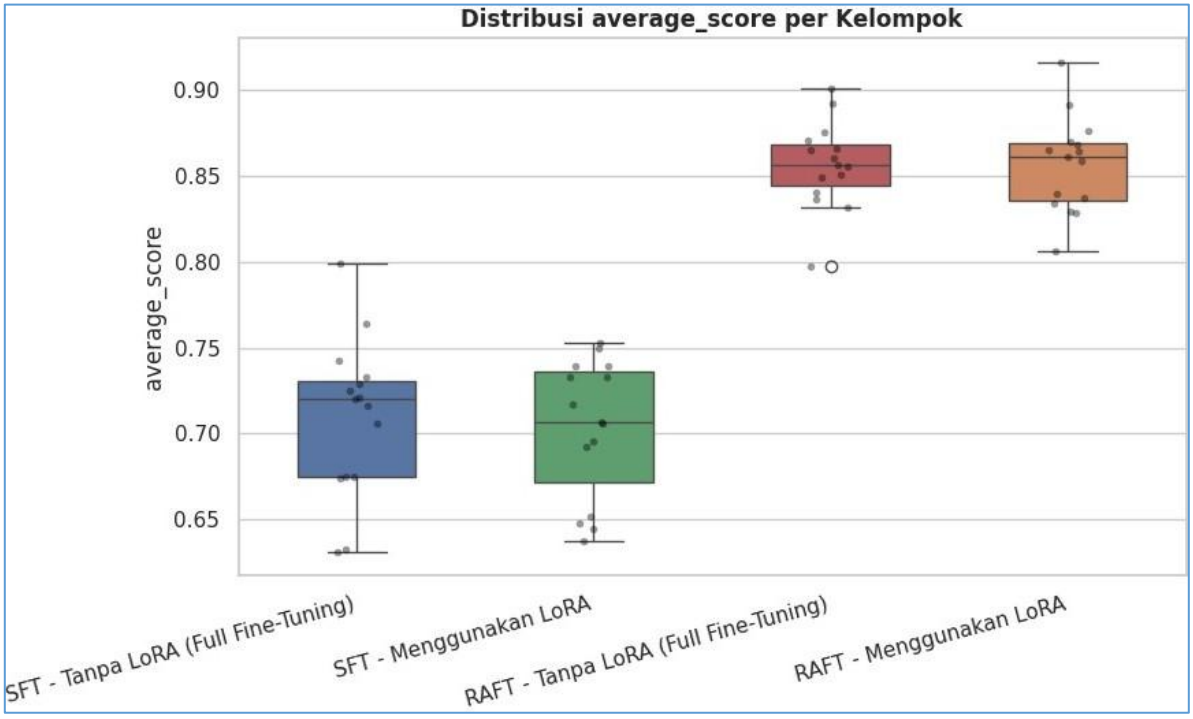
 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 3 Skor agregat per epoch (1, 3, dan 4) untuk empat konfigurasi.

Setiap titik adalah rata-rata skor agregat pada lima ukuran himpunan uji (10–50 pasangan tanya-jawab); angka pada epoch 4 adalah nilai akhir. Sumbu vertikal dimulai dari 0,60.

Sebagaimana diilustrasikan pada Gambar 3, SFT-Full sedikit melampaui SFT-LoRA pada setiap epoch yang diamati, tetapi selisihnya tetap kecil. Garis RAFT bahkan lebih berdekatan: LoRA sedikit lebih tinggi pada epoch 1 dan 4, sedangkan full fine-tuning lebih tinggi pada epoch 3. Selisih yang terbatas antar strategi pembaruan, dibandingkan dengan selisih yang jauh lebih besar antara SFT dan RAFT, menunjukkan bahwa struktur contoh pelatihan memiliki pengaruh yang lebih kuat dibandingkan jumlah parameter yang diperbarui.

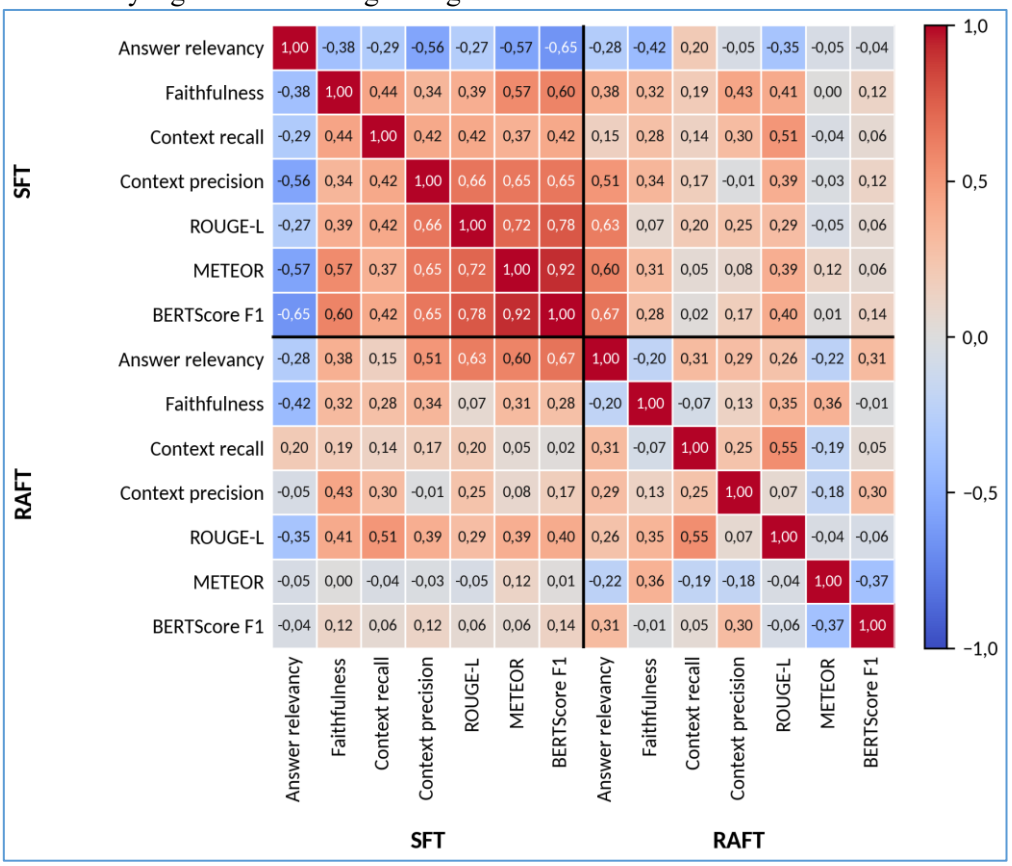


Gambar 4 Distribusi skor agregat pada 15 observasi per konfigurasi (3 epoch × 5 ukuran himpunan uji).

Kotak menunjukkan kuartil pertama hingga ketiga, garis tebal adalah median, whisker menunjukkan rentang data non-pencilan, dan lingkaran adalah pencilan.

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

Distribusi skor pada Gambar 4 memberikan perspektif berorientasi stabilitas terhadap perbandingan ini. Kedua kelompok RAFT terkonsentrasi pada rentang skor yang lebih tinggi dibandingkan kelompok SFT, dengan tumpang tindih yang terbatas antar paradigma. Variasi dalam masing-masing kelompok RAFT lebih kecil dibandingkan perbedaan antar paradigma, mendukung kesimpulan bahwa peningkatan yang dilaporkan bersifat sistematis di seluruh konfigurasi, bukan bergantung pada satu percobaan tertentu yang kebetulan menguntungkan.



Gambar 5 Matriks korelasi antarmetrik leksikal, semantik, dan RAGAS

Merah menunjukkan korelasi positif dan biru negatif; angka di dalam sel adalah koefisien korelasi. Garis tebal memisahkan metrik SFT dan RAFT.

Gambar 5 menunjukkan bahwa metrik-metrik yang digunakan menangkap sifat yang sebagian tumpang tindih namun tetap berbeda. Beberapa ukuran leksikal dan semantik SFT menunjukkan asosiasi positif yang kuat (misalnya 0,92 antara METEOR dan BERTScore F1), sedangkan hubungan pada RAFT lebih heterogen (seluruh koefisien antarmetrik RAFT berada di bawah 0,60 dalam nilai absolut) karena kualitas retrieval dan proses penghasilan jawaban memunculkan sumber variasi yang terpisah. Hasil ini menjustifikasi tetap dipertahankannya ROUGE-L, METEOR, BERTScore, dan dimensi RAGAS, alih-alih menggantinya dengan satu indikator leksikal atau semantik tunggal.

C. Hasil RAGAS dan Terkait Halusinasi

Hasil RAGAS pada Tabel 4 menunjukkan bahwa RAFT secara konsisten mengungguli SFT pada keempat metrik RAGAS. Peningkatan terbesar terlihat pada context precision dan context recall, diikuti oleh faithfulness. Answer relevancy juga meningkat dari 0,701 menjadi 0,764, menunjukkan peningkatan sedang dalam kemampuan model menghasilkan jawaban yang selaras dengan pertanyaan.

Namun, relevansi saja tidak menjamin jawaban hukum yang andal. Sebuah jawaban dapat membahas topik yang diminta sambil mengutip peraturan yang salah atau informasi yang tidak didukung. Oleh karena itu, metrik berbasis konteks memberikan ukuran keandalan model yang lebih bermakna untuk tugas tanya-jawab hukum.

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

Tabel 4 Rata-Rata Skor RAGAS untuk SFT dan RAFT

Metrik	SFT	RAFT	Selisih
Answer relevancy	0,701	0,764	+0,063
Faithfulness	0,640	0,872	+0,232
Context recall	0,637	0,907	+0,270
Context precision	0,646	0,920	+0,274

Faithfulness meningkat dari 0,640 pada model SFT menjadi 0,872 pada RAFT, atau peningkatan sebesar 0,232. Skor yang lebih tinggi ini menunjukkan bahwa sebagian besar pernyataan model dapat ditemukan secara langsung dalam konteks hukum yang relevan. Temuan ini mendukung hipotesis utama penelitian: penalaran kontekstual mengungguli ketergantungan pada pengetahuan parametrik ketika menghasilkan jawaban hukum. Selama pelatihan RAFT, komponen oracle memberikan bukti pendukung yang valid, sedangkan komponen distractor memperkenalkan informasi serupa namun tidak relevan. Paparan berulang terhadap kedua jenis input ini melatih model untuk memisahkan detail yang relevan secara hukum dari konten yang hanya terkait secara semantik.

Context recall meningkat dari 0,637 pada SFT menjadi 0,907 pada RAFT, sedangkan context precision meningkat dari 0,646 menjadi 0,920. Secara keseluruhan, hasil ini menunjukkan bahwa RAFT tidak hanya memproses lebih banyak teks; model juga belajar menggunakan himpunan bukti yang lebih komprehensif dan selektif.

Context recall mencerminkan apakah informasi yang dibutuhkan oleh jawaban rujukan tersedia, sedangkan context precision mencerminkan apakah konteks lebih menekankan materi yang relevan dibandingkan elemen yang mengecoh. Peningkatan simultan pada kedua metrik ini sangat penting dalam konteks Peraturan Desa karena banyak bagian teks menggunakan kosakata administratif, struktur pasal, dan terminologi anggaran yang serupa. Model yang memperlakukan pasal-pasal yang secara semantik mirip namun secara hukum berbeda dapat menghasilkan jawaban yang lancar secara bahasa namun keliru untuk desa, instrumen, atau tahun yang diminta.

Untuk kondisi SFT, metrik konteks ini harus dimaknai secara hati-hati karena SFT tidak menggunakan mekanisme retrieval yang sama seperti pada inferensi standar. Metrik ini menunjukkan seberapa baik konteks evaluasi mencakup dan mendukung jawaban, bukan kinerja independen dari retriever SFT. Meskipun ada keterbatasan ini, selisih yang besar tetap informatif: tujuan pelatihan RAFT secara eksplisit diselaraskan dengan konteks yang memuat bukti dan penolakan distractor, sedangkan tujuan pelatihan SFT tidak demikian.

Strategi pembaruan parameter menghasilkan perbedaan RAGAS yang jauh lebih kecil. SFT-LoRA mencapai answer relevancy sebesar 0,7104, dibandingkan 0,6906 untuk SFT-Full, sedangkan RAFT-LoRA dan RAFT-Full hampir identik, yaitu sekitar 0,763 dan 0,764. Faithfulness tetap berada di sekitar 0,640 untuk kedua strategi SFT dan 0,872 untuk kedua strategi RAFT. Untuk context recall, full fine-tuning sedikit lebih tinggi dibandingkan LoRA baik pada SFT (0,643 berbanding 0,631) maupun RAFT (0,914 berbanding 0,899). Sebaliknya, RAFT-LoRA mencapai context precision sebesar 0,9254, melampaui nilai 0,9150 yang diperoleh RAFT-Full. Perbedaan ini relatif kecil dibandingkan selisih antara SFT dan RAFT, dan menegaskan bahwa tujuan pelatihan berorientasi bukti bukan pembaruan parameter penuh merupakan sumber utama peningkatan grounding.

Dari perspektif sumber daya, hasil ini lebih menguntungkan RAFT-LoRA. LoRA mengadaptasi sebagian kecil parameter sehingga membutuhkan memori optimizer dan penyimpanan yang lebih sedikit dibandingkan full fine-tuning [26], [27]. Ketika kinerja evaluasinya setara atau sedikit lebih tinggi dibandingkan RAFT-Full, biaya tambahan untuk memperbarui seluruh model sulit dijustifikasi untuk korpus yang diteliti. Temuan ini tidak membuktikan bahwa LoRA akan selalu mengungguli full fine-tuning, tetapi menunjukkan bahwa adaptasi parameter yang efisien dapat mencukupi ketika dataset secara eksplisit merepresentasikan perilaku pemilihan bukti yang dibutuhkan saat penerapan.

D. Pengaruh Epoch dan Ukuran Himpunan Evaluasi

Keempat kondisi model menunjukkan peningkatan antara epoch 1 dan epoch 4, sebagaimana ditunjukkan pada Tabel V. SFT mengalami peningkatan yang relatif lebih besar karena dimulai dari skor yang lebih rendah dan memerlukan paparan tambahan untuk menyesuaikan pemetaan prompt-

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

jawaban. RAFT sudah menunjukkan kinerja yang kuat pada epoch 1, menunjukkan bahwa sinyal oracle dan distractor memberikan tujuan pembelajaran yang lebih jelas sejak awal pelatihan. Hasil ini menunjukkan bahwa keunggulan RAFT bukan semata-mata berasal dari checkpoint pada tahap akhir, melainkan sudah terlihat sejak epoch pertama yang diamati dan tetap bertahan seiring berjalannya pelatihan.

Tabel 5 Skor agregat per epoch

Konfigurasi	Epoch 1	Epoch 3	Epoch 4
SFT-LoRA	0,6547	0,7160	0,7384
SFT-Full	0,6575	0,7220	0,7492
RAFT-LoRA	0,8429	0,8437	0,8827
RAFT-Full	0,8340	0,8548	0,8810

Berdasarkan Tabel 5, Epoch 4 menghasilkan skor rata-rata tertinggi di antara seluruh konfigurasi. RAFT-LoRA mencapai skor keseluruhan terbaik sebesar 0,8827, diikuti RAFT-Full dengan skor 0,8810. SFT-Full dan SFT-LoRA juga mencapai skor tertinggi pada Epoch 4, meskipun selisih kinerja antara kedua pendekatan tetap terlihat jelas.

Namun, hasil ini tidak berarti bahwa penggunaan lebih banyak epoch pelatihan akan selalu meningkatkan kinerja. Penelitian ini hanya menguji Epoch 1, 3, dan 4. Pelatihan dengan lebih banyak epoch berpotensi menyebabkan model menghafal pola-pola spesifik dalam dataset, sehingga terjadi overfitting dan menurunkan kemampuan model pada data baru.

Penambahan jumlah pasangan tanya-jawab evaluasi dari 10 menjadi 50 juga tidak secara konsisten meningkatkan kinerja. Beberapa konfigurasi menunjukkan skor yang lebih rendah ketika pertanyaan yang lebih beragam dan menantang diperkenalkan, menunjukkan bahwa hasil dipengaruhi oleh faktor seperti kompleksitas pertanyaan, topik hukum, dan konteks. Meskipun himpunan uji terpisah memberikan perbandingan yang lebih andal, evaluasi ini tetap dibatasi oleh ukuran himpunan uji yang relatif kecil, yaitu 50 pasangan tanya-jawab.

E. Implikasi bagi AI Hukum Indonesia

Hasil penelitian ini menunjukkan bahwa RAFT sesuai untuk menjawab pertanyaan mengenai Peraturan Desa di Indonesia karena pengetahuan hukum tetap tersimpan dalam dokumen eksternal, bukan dalam parameter model. Ketika peraturan direvisi atau diganti, indeks dokumen dapat diperbarui tanpa perlu melatih ulang LLM. Pendekatan ini sangat berguna dalam ranah regulasi hukum, di mana akurasi informasi bergantung pada peraturan, pasal, yurisdiksi, tahun anggaran, dan otoritas penerbit yang berlaku. Pendekatan ini juga mendukung keterverifikasian sistem dengan memungkinkan pengguna melihat peraturan dan bagian dokumen yang digunakan untuk menghasilkan setiap jawaban. Hasil eksperimen lebih lanjut menunjukkan bahwa RAFT-LoRA memberikan kinerja yang kompetitif sambil hanya memperbarui sebagian kecil parameter model. Meskipun demikian, evaluasi otomatis saja tidak cukup untuk penelitian hukum kami. Sebuah model dapat mencapai skor evaluasi yang tinggi tetapi tetap melewatkan detail hukum penting atau mengandalkan peraturan yang sudah tidak berlaku. Oleh karena itu, sistem yang praktis harus menyediakan sumber sitasi yang jelas, menyimpan bukti pendukung, dan dapat ditinjau oleh perangkat desa atau ahli hukum. Validasi oleh ahli tetap menjadi hal yang esensial sebelum sistem ini dapat diterapkan di dunia nyata.

F. Keterbatasan

Penelitian ini memiliki beberapa keterbatasan. Pertama, cakupan 55 Peraturan Desa dari satu kabupaten membatasi generalisasi, karena variasi istilah dan struktur dokumen antarwilayah dapat memengaruhi kinerja model. Kedua, evaluasi yang hanya menggunakan 50 pasangan tanya-jawab pada epoch terpilih (1, 3, dan 4) memberikan gambaran yang terbatas mengenai perilaku pelatihan, distribusi skor, dan kinerja pada pertanyaan hukum yang kompleks.

Selain itu, hanya satu konfigurasi LoRA yang dievaluasi tanpa memvariasikan rank, alpha, dropout, atau target modules. Skor agregat juga memberikan bobot yang sama untuk seluruh metrik evaluasi,

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

meskipun berbagai jenis kesalahan dapat memiliki tingkat kepentingan yang berbeda dalam aplikasi hukum. Terakhir, evaluasi ini terutama mengandalkan metrik otomatis dan belum mencakup validasi hukum yang komprehensif oleh ahli. Penelitian selanjutnya sebaiknya melibatkan tinjauan ahli serta menelaah kualitas OCR, pengumpulan data, kinerja retrieval, dan akurasi sitasi untuk lebih memahami pengaruhnya terhadap sistem secara keseluruhan.

5. Kesimpulan

Penelitian ini membandingkan Supervised Fine-Tuning (SFT) dan Retrieval-Augmented Fine-Tuning (RAFT) menggunakan model dasar Llama-3.1-8B-Instruct yang sama untuk sistem tanya-jawab Peraturan Desa di Indonesia. Empat konfigurasi pelatihan dievaluasi: SFT-LoRA, SFT-Full, RAFT-LoRA, dan RAFT-Full. Secara keseluruhan, RAFT secara konsisten mengungguli SFT. Pada epoch 4, RAFT-LoRA mencapai skor agregat tertinggi (0,883), diikuti RAFT-Full (0,881), sedangkan model SFT terbaik mencapai 0,749. Temuan ini menunjukkan bahwa menggabungkan konteks oracle dan distractor memungkinkan model menghasilkan jawaban hukum yang lebih akurat dibandingkan hanya mengandalkan pengetahuan yang tersimpan pada parameter model.

Penelitian ini juga menunjukkan bahwa RAFT mencapai skor yang lebih tinggi pada relevansi jawaban, faithfulness, context recall, dan context precision, menunjukkan kemampuan yang lebih kuat dalam mengidentifikasi bukti hukum pendukung sekaligus menyaring informasi yang secara semantik mirip namun tidak relevan. Kemampuan ini sangat penting untuk menjawab pertanyaan hukum, di mana istilah yang serupa dapat mewakili makna hukum yang berbeda. Implikasinya, pengetahuan hukum desa dapat dijaga pada dokumen eksternal yang diperbarui tanpa pelatihan ulang, sementara model dilatih agar menjawab berdasarkan bukti.

Selain membandingkan kedua metode tersebut, penelitian kami juga menjabarkan tahapan penelitian pada Peraturan Desa di Indonesia, mencakup klasifikasi dokumen, ekstraksi teks dan OCR, pembersihan data, segmentasi struktur hukum, pembuatan pasangan tanya-jawab sintetis menggunakan LLM, konstruksi dataset, pelatihan model, dan evaluasi. Hasil kami juga menegaskan bahwa kualitas data dan praproses berperan penting dalam kinerja model bahasa hukum.

Terdapat beberapa keterbatasan, yaitu penelitian ini masih terbatas pada satu dataset desa, 50 pertanyaan evaluasi, jumlah epoch pelatihan yang sedikit, satu konfigurasi LoRA, serta evaluasi otomatis tanpa validasi ahli hukum. Kami berharap penelitian selanjutnya dapat memperluas dataset ke berbagai wilayah dan jenis peraturan, melibatkan ahli hukum dalam pembuatan dan evaluasi dataset, membandingkan konfigurasi LoRA tambahan dan penyempurnaan penuh, serta meningkatkan retrieval melalui hybrid search dan re-ranking. Evaluasi lebih lanjut juga sebaiknya menelaah akurasi sitasi, ketahanan OCR, kemampuan no-answer, stabilitas pelatihan, dan ketidakpastian statistik untuk menilai penerapan RAFT dalam sistem informasi hukum Indonesia yang lebih luas.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Institut Teknologi Del atas dukungan pengetahuan, teknis, dan komputasi yang diberikan selama penelitian ini, termasuk jaringan GPU B200 dan dokumentasi pendukung lainnya. Institusi juga menyediakan lingkungan penelitian yang mendukung sehingga memungkinkan penulis mengembangkan penelitian, menyiapkan dataset, melatih model, dan menyelesaikan proses evaluasi.

Selanjutnya, bimbingan akademik, diskusi yang bermakna, dan dukungan yang diterima selama proses penelitian memberikan kontribusi signifikan terhadap penyelesaian penelitian ini. Akhir kata, penulis menyampaikan apresiasi kepada seluruh pihak di Institut Teknologi Del.

Referensi

[1] A. Matarazzo and R. Torlone, "*A Survey on Large Language Models with Some Insights on Their Capabilities and Limitations*," *arXiv preprint arXiv:2501.04040*, 2025, DOI: 10.48550/arXiv.2501.04040.

[2] A. Vaswani et al., "*Attention is All You Need*," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vol. 30, 2017, pp. 5998–6008, DOI: 10.48550/arXiv.1706.03762.

[3] Z. Ji et al., "*Survey of Hallucination in Natural Language Generation*," *ACM Comput. Surv.*, Vol. 55, No. 12, pp. 1–38, Mar. 2023, DOI: 10.1145/3571730.

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

[4] L. Huang et al., "*A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*," *ACM Trans. Inf. Syst.*, Vol. 43, No. 2, Jan. 2025, DOI: 10.1145/3703155.
[5] S. Cahyawijaya et al., "*Cendol: Open Instruction-Tuned Generative Large Language Models for Indonesian Languages*," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, Aug. 2024, pp. 14899–14914, DOI: 10.18653/v1/2024.acl-long.796.
[6] Republik Indonesia, *Undang-Undang Nomor 6 Tahun 2014 tentang Desa*, Lembaran Negara Republik Indonesia Tahun 2014 Nomor 7, 2014.
[7] Republik Indonesia, *Undang-Undang Nomor 12 Tahun 2011 tentang Pembentukan Peraturan Perundang-undangan*, 2011.
[8] Republik Indonesia, *Peraturan Presiden Nomor 33 Tahun 2012 tentang Jaringan Dokumentasi dan Informasi Hukum Nasional*, 2012.
[9] Z. Gekhman et al., "*Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?*," in *Proc. 2024 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Miami, FL, USA, Nov. 2024, pp. 7765–7784, DOI: 10.18653/v1/2024.emnlp-main.444.
[10] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "*Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs*," *arXiv preprint arXiv:2312.05934*, Dec. 2023, DOI: 10.48550/arXiv.2312.05934.
[11] T. Zhang et al., "*RAFT: Adapting Language Model to Domain Specific RAG*," *arXiv preprint arXiv:2403.10131*, 2024, DOI: 10.48550/arXiv.2403.10131.
[12] S. Gupta, V. Agrawal, Y. Negi, S. Karunanithi, and A. Balakrishnan, "*Reducing Hallucinations in Legal AI: A Retrieval Augmented Generation-based Model for Accurate Legal Guidance*," *IEEE Access*, Vol. 14, pp. 44451–44463, 2026, DOI: 10.1109/ACCESS.2026.3675624.
[13] P. Lewis et al., "*Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks*," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vol. 33, 2020, pp. 9459–9474, DOI: 10.48550/arXiv.2005.11401.
[14] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "*Retrieval Augmentation Reduces Hallucination in Conversation*," in *Findings Assoc. Comput. Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, Nov. 2021, pp. 3784–3803, DOI: 10.18653/v1/2021.findings-emnlp.320.
[15] A. Grattafiori et al., "*The Llama 3 Herd of Models*," *arXiv preprint arXiv:2407.21783*, 2024, DOI: 10.48550/arXiv.2407.21783.
[16] M. Turski, T. Stanisławek, K. Kaczmarek, P. Dyda, and F. Graliński, "*CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data*," in *Document Analysis and Recognition – ICDAR 2023*, Lecture Notes in Computer Science, Vol. 14189. Cham, Switzerland: Springer, 2023, pp. 348–365, DOI: 10.48550/arXiv.2304.14953.
[17] R. Sinha and R. B S, "*Digitization of Document and Information Extraction using OCR*," *arXiv preprint arXiv:2506.11156*, 2025, DOI: 10.48550/arXiv.2506.11156.
[18] R. Smith, "*An Overview of the Tesseract OCR Engine*," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, Vol. 2, Curitiba, Brazil, 2007, pp. 629–633, DOI: 10.1109/ICDAR.2007.4376991.
[19] A. Upadhye, "*A Comprehensive Survey of Text Data Cleaning Techniques: Challenges, Methods, and Best Practices*," Journal of Scientific and Engineering Research, Vol. 7, No. 8, pp. 205–210, 2020.
[20] F. Josi, C. Wartena, and U. Heid, "*Preparing Legal Documents for NLP Analysis: Improving the Classification of Text Elements by using Page Features*," in *Comput. SCI. Inf. Technol. (CS & IT)*, 2022, DOI: 10.5121/csit.2022.120102.
[21] C. Merola and J. Singh, "*Reconstructing Context: Evaluating Advanced Chunking Strategies for Retrieval-Augmented Generation*," in *Knowledge-Enhanced Information Retrieval: 2nd Int. Workshop, KEIR 2025, Revised Selected Papers*, Lucca, Italy, Apr. 2025. Berlin, Germany: Springer, 2025, pp. 3–18, DOI: 10.1007/978-3-032-02899-0_1.
[22] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "*SQuAD: 100,000+ Questions for Machine Comprehension of Text*," in *Proc. 2016 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Austin, TX, USA, Nov. 2016, pp. 2383–2392, DOI: 10.18653/v1/D16-1264.

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

[23] Y. Wan et al., "*SciQAG: A Framework for Auto-Generated Science Question Answering Dataset with Fine-Grained Evaluation*," *arXiv preprint arXiv:2405.09939*, May 2024, DOI: 10.48550/arXiv.2405.09939.
[24] S. Yuen, T. Su, Z. Wang, Y. Du, and A. Sobey, "*Automatic Dataset Generation for Knowledge Intensive Question Answering Tasks*," *arXiv preprint arXiv:2505.14212*, 2025, DOI: 10.48550/arXiv.2505.14212.
[25] X. Yu, Z. Zhang, F. Niu, X. Hu, X. Xia, and J. Grundy, "*What Makes a High-Quality Training Dataset For Large Language Models: A Practitioners' Perspective*," in *Proc. 39th IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE '24)*, New York, NY, USA: ACM, 2024, pp. 656–668, DOI: 10.1145/3691620.3695061.
[26] E. J. Hu et al., "*LoRA: Low-Rank Adaptation of Large Language Models*," *arXiv preprint arXiv:2106.09685*, 2021, DOI: 10.48550/arXiv.2106.09685.
[27] L. Wang et al., "*Parameter-Efficient Fine-Tuning in Large Models: A Survey of Methodologies*," *arXiv preprint arXiv:2410.19878*, Oct. 2024, DOI: 10.48550/arXiv.2410.19878.
[28] Y. Chang et al., "*A Survey on Evaluation of Large Language Models*," *ACM Trans. Intell. Syst. Technol.*, Vol. 15, No. 3, Mar. 2024, DOI: 10.1145/3641289.
[29] C.-Y. Lin, "*ROUGE: A Package for Automatic Evaluation of Summaries*," in *Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/
[30] S. Banerjee and A. Lavie, "*METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, Ann Arbor, MI, USA, Jun. 2005, pp. 65–72.
[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "*BERTScore: Evaluating Text Generation with BERT*," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, DOI: 10.48550/arXiv.1904.09675.
[32] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "*RAGAs: Automated Evaluation of Retrieval Augmented Generation*," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations*, St. Julian's, Malta, Mar. 2024, pp. 150–158, DOI: 10.18653/v1/2024.eacl-demo.16.