Journal of Computer and Information System (J-CIS)
Vol.8, No.2, October 2025, Page 85-96
ISSN (print): 2622-5859, ISSN (online): 2622-0881
DOI: https://doi.org/10.31605/jcis.v8i2

Sistem Tanya Jawab Layanan Administrasi Kependudukan dengan *Retrieval Augmented Generation* Komodo-7B

Adelia Azizatul Haq[1], I Gede Susrama Mas Diyasa*[2], Sugiarto[3]
[1,2,3]Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jawa Timur
E-mail: [1]adeliaazizatul56@gmail.com, *[2]igsusrama.if@upnjatim.ac.id,
[3]sugiarto.if@upnjatim.ac.id

Abstrak

Pesatnya transformasi digital di Indonesia mendorong pemerintah untuk terus berinovasi dalam meningkatkan kualitas layanan publik, termasuk di bidang administrasi kependudukan. Dinas Kependudukan dan Pencatatan Sipil (Disdukcapil) Kota Surabaya sebagai salah satu penyedia layanan kependudukan, menghadapi tantangan dalam menjawab kebutuhan masyarakat akan layanan yang cepat, responsif, dan tersedia di luar jam operasional. Dalam rangka meningkatkan kualitas layanan, penelitian ini mengembangkan sistem Question Answering (QA) berbasis Large Language Model (LLM) untuk menjawab pertanyaan seputar layanan Kartu Tanda Penduduk (KTP) dan Kartu Keluarga (KK). Sistem dirancang dengan memanfaatkan model LLM Komodo-7B yang telah disesuaikan menggunakan teknik fine-tuning Quantized Low-Rank Adaptation (QloRA) dan pendekatan Retrieval Augmented Generation (RAG) guna meningkatkan akurasi dan relevansi jawaban. Data pelatihan mencakup dataset pengaduan-pertanyaan Disdukcapil Kota Surabaya dan empat data open source. Proses RAG menggunakan vektorisasi kalimat dengan sentence transformer dan pemanggilan konteks berbasis cosine similarity. Evaluasi dilakukan menggunakan metrik ROUGE-1, ROUGE-L, dan METEOR. Hasil evaluasi RAG fine-tuning Komodo-7B didapatkan F1-Score ROUGE-1 mencapai 0.299, ROUGE-L 0.251, dan skor METEOR 0.275. Meskipun hasil ini menunjukkan peningkatan dibandingkan model dasar, performa yang dicapai masih tergolong rendah.

Kata kunci*— Question Answering, Retrieval Augmented Generation, Large Language Models, Komodo-7B, Q-LoRA*

Abstract

The rapid digital transformation in Indonesia encourages the government to continue to innovate in improving the quality of public services, including in the field of population administration. The Population and Civil Registration Office (Disdukcapil) of Surabaya City as one of the population service providers, faces challenges in responding to the community's need for services that are fast, responsive, and available outside of operational hours. In order to improve the quality of service, this research develops a Question Answering (QA) system based on the Large Language Model (LLM) to answer questions about the Identity Card (KTP) and Family Card (KK) services. The system is designed by utilizing the Komodo-7B LLM model that has been customized using QLoRA fine-tuning techniques and the Retrieval Augmented Generation (RAG) approach to improve the accuracy and relevance of answers. The training data includes the Surabaya City Disdukcapil question-complaint dataset and four open source data. The RAG process uses sentence vectorization with a sentence transformer and cosine similarity-based context calling. Evaluation was conducted using ROUGE-1, ROUGE-L, and

METEOR metrics. Komodo-7B fine-tuning RAG evaluation results obtained F1-Score ROUGE-1 reached 0.299, ROUGE-L 0.251, and METEOR score 0.275. Although these results show an improvement over the base model, the performance achieved is still relatively low.

Keywords—*Question Answering, Retrieval Augmented Generation, Large Language Models, Komodo-7B, Q-LoRA*

1. PENDAHULUAN

Pesatnya transformasi digital di Indonesia mendorong pemerintah untuk berinovasi dalam meningkatkan layanan publik [1]. Salah satu instansi yang turut beradaptasi dengan perubahan ini adalah Dinas Kependudukan dan Pencatatan Sipil (Disdukcapil) Kota Surabaya adalah organisasi perangkat daerah (OPD) di Bidang Administrasi Kependudukan dan Pencatatan Sipil [2]. Sebagai penyelenggara layanan publik, Disdukcapil menyediakan berbagai layanan kepada masyarakat, termasuk layanan pengaduan. Dalam 6 bulan terakhir, terdapat keluhan dari salah satu warga pada penilaian google review dengan menyertakan bintang satu, sebagaimana ditampilkan pada Gambar 1 berikut.

Gambar 1 Aspirasi masyarakat melalui *google review*

Keluhan pada Gambar 1 disebabkan karena warga yang bersangkutkan menghubungi layanan *call center* tetapi panggilan sedang sibuk. Situasi panggilan yang sibuk ini disebabkan oleh warga yang menelepon di luar jam operasional atau selama hari libur. Keluhan yang dikemukakan menunjukkan aspirasi warga untuk mendapatkan layanan dengan cepat dan responsif, bahkan di luar jam operasional. Mempertimbangkan aspirasi tersebut, maka dibutuhkan *Question Answering* (QA) *System* sebagai salah satu solusi untuk meningkatkan pelayanan. QA termasuk kedalam *Natural Language Preprocessing* (NLP), yang merupakan salah satu teknologi kecerdasan buatan berkaitan dengan pemrosesan bahasa alami manusia (memahami, menganalisis, dan memanipulasi) oleh mesin atau komputer [3][4]. Perkembangan teknologi *Large Language Models* (LLM) turut mendorong transformasi sistem QA. LLM mampu memahami konteks, sintaksis, dan semantik bahasa secara lebih mendalam dibandingkan pendekatan konvensional berbasis aturan atau statistik, karena dilatih pada korpus teks dalam jumlah besar menggunakan arsitektur Transformer [5].

Salah satu penelitian yang relevan terkait *Large Language Models* (LLM) dalam konteks bahasa Indonesia dilakukan oleh Hakim, dkk. Dalam penelitian tersebut, dikaji berbagai pengembangan model LLaMa2-7B yang telah diadaptasi untuk teks berbahasa Indonesia. Beberapa model yang dikembangkan antara lain model Komodo-7B dengan Bahasa Indonesia dan 11 bahasa daerah, Cendol-7B dengan Bahasa Indonesia umum, dan Sealion-7B dengan bahasa-bahasa di Asia Tenggara, termasuk Bahasa Indonesia. Penelitian ini bertujuan untuk meningkatkan kemampuan LLM berbahasa Indonesia dengan membandingkan model meliputi Komodo-7B, Cendol-7B, Sealion-7B, dan Bactrian-X-7B, yang diuji dalam tiga skenario: *zero-shot prompting*,*five-shot prompting*, dan*fine-tuning* menggunakan metode *Low-Rank Adaptation*

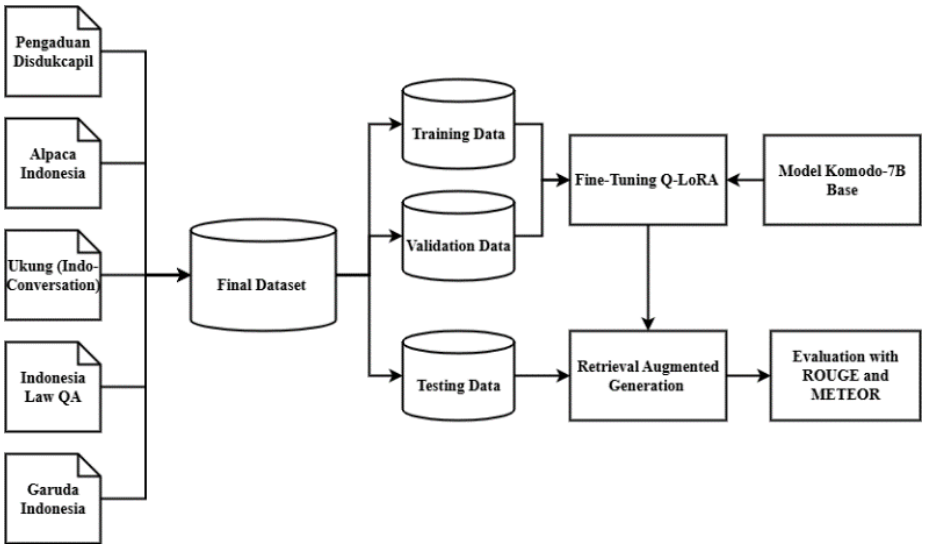
(LoRA). Hasil eksperimen menunjukkan bahwa model Komodo-7B memberikan hasil terbaik dengan *fine-tuning* LoRA, dengan skor ROUGE-L sebesar 35.29 [6].

Penelitian oleh Akhila Abdulnazar, dkk., menunjukkan bahwa metode Retrieval Augmented Generation dengan LLM lebih baik dibandingkan metode bi-encoder dan dictionary matching dengan hasil F1 score sebesar 0.607 pada data korpus teks klinis bahasa jerman [7]. Penelitian lain terkait LLM adalah penelitian oleh Harshit Kumar Chaubey, dkk., yang membandingkan metode RAG, *Fine-Tuning*, dan *Prompt Engineering* untuk penerapan LLM dengan set data "openassistant-guanaco" di Hugging Face. Hasil evaluasi metode *fine-tuning* merupakan yang paling baik dengan akurasi 85,7%, skor BLEU 0,81, dan HES 8,9, dan *perplexity score*-nya adalah 10,3. Penelitian ini menyarankan untuk menggabungkan metode *fine-tuning* dengan RAG dalam penerapan LLM [8].

Penelitian-penelitian tersebut telah membahas performa *Large Language Models* (LLM) dan berbagai pendekatan peningkatannya. Hakim, dkk. menunjukkan bahwa model Komodo-7B memberikan performa terbaik dalam skenario berbahasa Indonesia, khususnya setelah melalui proses *fine-tuning* dengan LoRA. Sementara itu, studi oleh Akhila Abdulnazar, dkk. dan Harshit Kumar Chaubey, dkk. masing-masing mendapatkan hasil evaluasi terbaik pada metode *Retrieval-Augmented Generation* (RAG) dan *fine-tuning*, meskipun belum secara khusus diarahkan pada konteks layanan publik. Penelitian Harshit Kumar Chaubey, dkk. bahkan merekomendasikan integrasi antara RAG dan *fine-tuning* sebagai strategi potensial untuk menghasilkan model yang lebih akurat dan relevan. Berdasarkan hal tersebut, penelitian ini memberikan kebaruan berupa implementasi kombinasi *fine-tuning* menggunakan *Quantized Low-Rank Adaptation* (QLoRA) dan pendekatan RAG pada model Komodo-7B dalam sistem *Question Answering* berbahasa Indonesia. Pendekatan ini difokuskan secara khusus untuk menjawab pertanyaan dalam domain layanan administrasi kependudukan (ADMINDUK), seperti KTP dan KK.

2. METODE

Penelitian ini diawali dengan persiapan data, mencakup penggabungan data, augmentasi, dan tahapan *preprocessing*. Selanjutnya, data digunakan untuk pelatihan model Komodo-7B dengan *fine-tuning* Q-LoRA. Kemudian, diintegrasikan dengan pendekatan *Retrieval Augmented Generation* dan dievaluasi menggunakan metrik ROUGE-1, ROUGE-L, dan METEOR. Gambaran umum proses penelitian terdapat pada Gambar 2.



Gambar 2 Gambaran umum penelitian

2. 1 Dataset Penelitian

Penelitian ini menggunakan lima sumber data, satu data berasal dari instansi dan empat lainnya merupakan dataset *open-source* dari Hugging Face: (1) Data pengaduan dari Dinas Kependudukan dan Pencatatan Sipil (Disdukcapil) Kota Surabaya; (2) Alpaca Indonesia; (3) Ukung (Indo-Conversation); (4) Indonesia Law QA; (5) Garuda Indonesia. Data dari Dispendukcapil merupakan data primer karena diperoleh langsung dari sumber internal dan telah divalidasi secara manual oleh petugas dengan proses wawancara dan memeriksa satu-persatu untuk memastikan tidak ada informasi sensitif atau keluhan yang tidak relevan. Selanjutnya dilakukan proses *preprocessing* teks, yang mencakup mengubah seluruh teks menjadi huruf kecil dan penghapusan yang tidak diperlukan seperti simbol agar data menjadi berkualitas dan siap untuk pelatihan model [9][10]. Setelah tahap *preprocessing*, dilakukan proses augmentasi data dari yang semula 570 menjadi 9690 secara manual dan metode GenAug (*generative augmentation*).

Sebanyak 570 data hasil augmentasi yang merepresentasikan tiap konteks digunakan sebagai data uji untuk sistem *Retrieval Augmented Generation* (RAG). Sementara itu, sisa data hasil augmentasi digabungkan dengan empat dataset dari Hugging Face. Semua data diselaraskan formatnya ke dalam struktur *prompt* dan *response* untuk keperluan *fine-tuning* model Komodo-7B menggunakan metode QLoRA. Kemudian hasil *fine-tuning* model diintegrasikan ke dalam sistem RAG dan diuji menggunakan 570 data uji. Rincian masing-masing dataset dapat dilihat pada Tabel 1.

Tabel 1 Rincian dataset penelitian

No	Dataset	Deskripsi	Jumlah Data
1	Pengaduan Disdukcapil	Dataset internal yang dikembangkan berdasarkan pengaduan dan pertanyaan masyarakat kepada layanan Disdukcapil Kota Surabaya, khususnya yang berkaitan dengan layanan KTP dan KK.	9690
2	Alpaca Indonesia	Dataset ini merupakan adaptasi dari instruksi Alpaca yang telah diterjemahkan ke dalam Bahasa Indonesia, terdiri dari pasangan instruksi dan jawaban yang telah dibersihkan serta disesuaikan dengan konteks lokal.	45631
3	Ukung (Indo-*Conversation*)	Dataset percakapan Bahasa Indonesia berbasis QA yang mencakup interaksi dialog sehari-hari serta beberapa topik umum seperti pendidikan, pekerjaan, dan sosial.	59022
4	Indonesia *Law* QA	Dataset QA bertema hukum Indonesia disusun dari sumber hukum seperti peraturan perundang-undangan untuk mendukung pengembangan model dengan pemahaman semantik terhadap teks hukum di Indonesia.	7172
5	Garuda Indonesia	Dataset skala besar berbahasa Indonesia berisi kombinasi data hasil pretraining dari berbagai domain seperti Wikipedia, berita, percakapan daring, dan forum.	500000
Total keseluruhan			620945

2. 2 Model LLM Komodo-7B

Model Komodo-7B dikembangkan oleh peneliti AI Louise Owen, Vishesh Tripathi, Abhay Kumar, dan Bidwan Ahmed dari perusahaan teknologi layanan pelanggan Yellow AI. Model ini merupakan hasil pengembangan dari LLM Llama-2 dan memiliki tujuh miliar parameter, yang mencerminkan namanya, "7B." [11]. Menariknya, meskipun Llama-2 dianggap tidak cocok untuk penggunaan non-Inggris oleh Meta pada tahun 2023, Komodo-7B dirancang

untuk mendukung layanan penerjemahan bahasa dan berkontribusi dalam mengatasi kesenjangan pendidikan di Indonesia dengan menyediakan terjemahan langsung dari bahasa Inggris ke bahasa Indonesia dan 11 bahasa daerah [12]. Salah satu keunggulan dari Komodo-7B-Base adalah perluasan kosakata tokenizer-nya, yakni dengan menambahkan sekitar 2.000 kata penting dari Bahasa Indonesia dan 1.000 kata dari bahasa daerah yang sebelumnya belum tersedia dalam Llama-2. Proses ini dilakukan tanpa membangun tokenizer baru dari awal, melainkan memperluas tokenizer yang telah ada agar lebih peka terhadap konteks lokal. Secara teknis, model ini memiliki 32 lapisan, dimensi sebesar 4096, *head attention* sebanyak 32, panjang maksimum token 4096, dan jumlah kosakata sebanyak 35.008 [13].

2. 3 Fine Tuning Q-LoRA

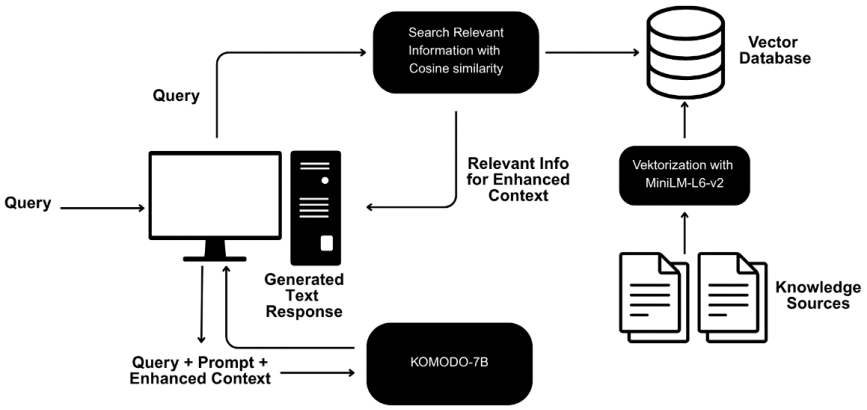
Proses menyesuaikan Komodo-7B dengan domain layanan publik, dilakukan proses *fine-tuning* menggunakan metode *Quantized Low-Rank Adaptation* (QLoRA). QLoRA memungkinkan pemanfaatan perangkat keras dengan kapasitas memori terbatas untuk melatih kembali model berskala besar tanpa menyebabkan penurunan akurasi yang berarti. Prinsip kerjanya adalah dengan mempertahankan bobot asli model dalam bentuk terkuantisasi 4-bit guna mengurangi konsumsi memori, kemudian menyisipkan adaptor *low-rank adapter* yang dilatih secara terpisah untuk menyesuaikan representasi model terhadap tugas baru [14]. Proses *fine-tuning* dilakukan pada lapisan-lapisan penting model, yaitu: q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, dan down_proj. Pelatihan dilakukan menggunakan parameter sebagai berikut: *batch_size* per perangkat sebesar 8, *gradient_accumulation_steps* sebanyak 1, total langkah (*max_steps*) sebanyak 1000, dan *learning_rate* sebesar 2e-4. Model menggunakan fp16 untuk kompatibilitas dengan *quantization* 4-bit, dan *optimizer* paged_adamw_8bit. Evaluasi dilakukan setiap 25 langkah dengan logging setiap 25 langkah. Spesifikasi lingkungan *fine-tuning* Q-LoRA terdapat pada Tabel 2.

Tabel 2 Spesifikasi Lingkungan

Spesifikasi	Keterangan
Platform	Kaggle Notebook
GPU	NVIDIA Tesla T4
Total Memori GPU	14.741 GB

2. 4 Retrieval Augmented Generation

Pendekatan *Retrieval Augmented Generation* (RAG) adalah teknik dalam pemrosesan bahasa alami (*Natural Language Processing*) yang menggabungkan kekuatan LLM dengan data eksternal untuk menghasilkan jawaban yang lebih akurat dan kontekstual [15]. Berikut adalah ilustrasi proses RAG pada Gambar 3.



Gambar 3 Ilustrasi proses *Retrieval Augmented Generation*

Proses diawali dengan pembuatan indeks vektor dokumen (vektorisasi) dari rekap pertanyaan dan pengaduan masyarakat oleh Dinas Kependudukan dan Pencatatan Sipil Kota Surabaya. Vektorisasi dilakukan menggunakan *sentence transformer* berbasis MiniLM-L6-v2, yaitu model transformer berukuran ringan yang dioptimalkan melalui *knowledge distillation* untuk menghasilkan *sentence embeddings* berkualitas tinggi secara efisien, sehingga tetap akurat meskipun menggunakan sumber daya komputasi yang terbatas [16]. Setelah pengguna memberikan input, sistem akan melakukan pencarian dua informasi yang paling relavan menggunakan *cosine similarity,* proses ini dinamakan *retrieval*. *Cosine similarity* merupakan suatu metode yang digunakan untuk mengukur tingkat kemiripan antar teks berdasarkan dua vektor. Metode pengukuran kemiripan teks ini termasuk yang paling populer diantara metode lainnya [17]. Ketiga informasi yang telah dikumpulkan kemudian digabungkan dengan input pengguna melalui rekayasa *prompt*, proses ini disebut *augmentation*. Selanjutnya proses *generation*, mengirim *prompt* ke model Komodo-7B yang telah di *fine-tuning* untuk menghasilkan jawaban yang kemudian dikirimkan kepada pengguna. Prompt terdapat pada Tabel 3 berikut.

Tabel 3 *Prompt* pada RAG *fine-tuning*

Prompt Augmentasi
" Kamu adalah Sistem *Question Answering* Layanan Administrasi Kependudukan Disdukcapil Surabaya. Jawab pertanyaan berdasarkan referensi dengan sopan dan jelas. Jangan menambahkan informasi lain dari luar referensi. Pertanyaan: {*query*} Referensi: {*reference*} Jawaban:"

2. 5 Metrik Evaluasi

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) merupakan metrik evaluasi yang mengukur kemiripan antara ringkasan mesin dan ringkasan acuan berdasarkan tumpang tindih n-gram, sehingga efektif untuk menilai kesamaan leksikal namun kurang mampu menangkap kesamaan semantik yang lebih dalam. Sebaliknya, METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) mengombinasikan kesesuaian n-gram dengan analisis sinonim dan bentuk kata dasar, sehingga lebih baik dalam mengakomodasi variasi bahasa dan makna dibandingkan ROUGE, meskipun tetap memiliki keterbatasan pada teks panjang. ROUGE unggul dalam kesederhanaan perhitungan dan interpretasi, sementara METEOR memberikan

penilaian yang lebih seimbang antara presisi dan recall dengan mempertimbangkan hubungan semantik antar kata [18] [19]. Berikut adalah perhitungan untuk tiap metrik evaluasi terdapat pada Tabel 4.

Tabel 4 Perhitungan metrik evaluasi

No	Metrik Evaluasi	Perhitungan	
1	ROUGE-N	$Precision\ ROUGE-N = \frac{Jumlah\ n-gram\ yang\ cocok}{Jumlah\ n-gram\ dalam\ hasil}$	(1)
		$Recall\ ROUGE-N = \frac{Jumlah\ n-gram\ yang\ cocok}{Jumlah\ n-gram\ dalam\ refrensi}$	(2)
		$f1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$	(3)
2	ROUGE-L	$Precision\ ROUGE-L = \frac{LCS(hasil,refrensi)}{panjang\ hasil}$	(4)
		$Recall\ ROUGE-L = \frac{LCS(hasil,refrensi)}{panjang\ refrensi}$	(5)
		$f1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$	(6)
3	METEOR	$P = \frac{Jumlah\ token\ yang\ cocok}{Jumlah\ token\ dalam\ hasil}$	(7)
		$R = \frac{Jumlah\ token\ yang\ cocok}{Jumlah\ token\ dalam\ refrensi}$	(8)
		$fmean = \frac{P \times R}{a \times P+(1-a) \times R}$	(9)
		$Penalty = \gamma \times \left(\frac{fragmen}{jumlah\ token\ cocok}\right)^{\beta}$	(10)
		$METEOR = fmean \times (1-penalty)$	(11)

3. HASIL DAN PEMBAHASAN

Pada bagian ini menjelaskan hasil penelitian sistem *Question Answering* dengan *Retrieval Augmented Generation* model *fine-tuning* Komodo-7B untuk studi kasus layanan ADMINDUK KTP dan KK. Berikut adalah hasil *generate* jawaban dari Komodo-7B Base yang merupakan model dasar dan belum dilatih dengan model yang sudah melalui *fine-tuning* Q-LoRA dengan pendekatan RAG pada Tabel 5.

Tabel 5 Hasil jawaban dari Komodo-7B *Base* dan RAG *fine-tuning* Komodo-7B

No	Pertanyaan	Hasil yang diharapkan	Komodo-7B *Base*	RAG *fine-tuning* Komodo-7B
1	Bisa nggak mengajukan perubahan data pekerjaan di KK secara online?	Bisa dilakukan secara online melalui menu perubahan biodata di Klampid.	Nggak bisa , hanya untuk pengajuan dokumen kependudukannya saja.	Ya, perubahan data pekerjaan di KK dapat diajukan secara online melalui Klampid.
2	Status perkawinan pada KTP bisa diubah kan?	Bisa, tetapi data dalam KTP bisa diubah setelah data dalam KK diperbarui. Pengajuan perubahan atau cetak ulang KTP melalui kelurahan atau kecamatan setempat dengan	*1) Yes, you can change the status of marriage on your ID card by going to this link https://dispendukp3sby 2019-surabaya.id/index.php/formulir_pergantianst*	Ya, data pada KK diperbarui terlebih dahulu .

		membawa KTP lama, KK, dan dokumen pendukung	*atuskewarganyataandi SurabayakhususnyaKelurahanku*	
3	Apakah cetak ulang KK yang hilang harus dilakukan di Siola?	Tidak perlu ke Siola, pengajuan dapat dilakukan secara online melalui Klampid atau langsung di kelurahan	*No. , you don't have to do it at the SIOL office because there is a service called "SIM-TIK" which can be accessed through https://simtik.sby.go id/ or by using QR code below :*	Tidak perlu , bisa lewat Klampid atau kelurahan.

Pada model Komodo-7B *Base*, ditemukan tingkat halusinasi yang tinggi dalam menjawab pertanyaan-pertanyaan terkait layanan administrasi kependudukan seperti Kartu Keluarga (KK) dan Kartu Tanda Penduduk (KTP). Berdasarkan analisis terhadap tiga pertanyaan yang ditampilkan dalam Tabel 5, model Komodo-7B *Base* mampu memahami konteks pertanyaan dan memberikan jawaban yang berkaitan dengan topik, namun jawaban yang dihasilkan tidak sesuai dengan prosedur yang benar. Pada pertanyaan pertama, model menyatakan bahwa permohonan tidak bisa dilakukan karena hanya untuk pengajuan dokumen kependudukan saja. Padahal, perubahan data pekerjaan pada KK jelas termasuk dalam kategori dokumen kependudukan. Selanjutnya, pada pertanyaan kedua dan ketiga, meskipun secara garis besar benar ("bisa" atau "tidak"), namun model tidak konsisten dalam penggunaan bahasa dengan menggunakan bahasa Inggris. Selain itu, hasil generasi jawaban tersebut juga mengandung informasi yang salah, seperti menyertakan tautan situs yang tidak relavan, tidak terkait dengan Disdukcapil Surabaya, dan bahkan tidak benar-benar ada. Kondisi ini menunjukkan bahwa model Komodo-7B *Base* tidak hanya mengalami inkonsistensi bahasa, tetapi juga menghasilkan informasi yang menyesatkan, sehingga mengurangi tingkat kepercayaan dan memperkuat indikasi adanya halusinasi yang tinggi.

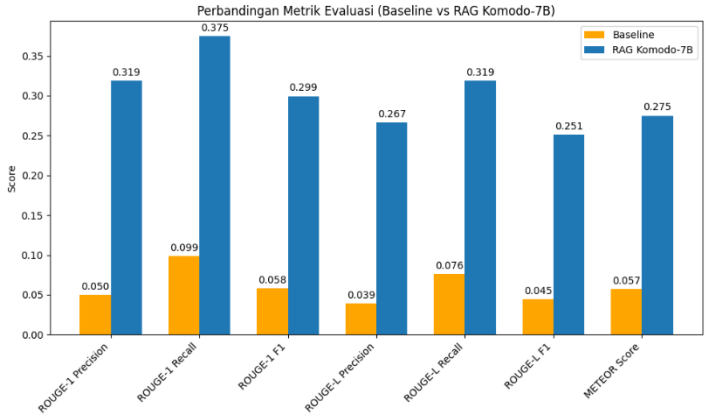
Sebaliknya, RAG *fine-tuning* Komodo-7B menunjukkan peningkatan kualitas jawaban yang signifikan dibandingkan dengan versi *base*. Model ini mampu memahami maksud pertanyaan, menyampaikan prosedur yang sesuai regulasi, juga menjaga konsistensi bahasa Indonesia yang formal tanpa campuran bahasa asing. Pada pertanyaan pertama, model memberikan jawaban yang sesuai alur administratif, yaitu mewajibkan pembaruan data KK terlebih dahulu sebelum perubahan status di KTP. Pada pertanyaan kedua, generasi jawaban yang dihasilkan menjelaskan alternatif proses melalui Klampid atau langsung di kelurahan tanpa mengarahkan ke tautan eksternal yang tidak relevan. Pada pertanyaan ketiga, model memberikan instruksi tepat bahwa perubahan data pekerjaan di KK dapat dilakukan secara online melalui Klampid. Peningkatan ini mengindikasikan bahwa integrasi pendekatan *Retrieval Augmented Generation* dan proses *fine-tuning* berhasil memperbaiki *grounding* informasi dan menghilangkan halusinasi. Namun, model cenderung memberikan generasi jawaban yang relatif singkat dan langsung ke inti jawaban. Hal ini dapat memudahkan pengguna memperoleh informasi yang jelas dan cepat. Akan tetapi, dapat juga menjadi kekurangan jika pengguna membutuhkan uraian yang lebih detail, sehingga berpotensi membuat jawaban kurang komprehensif pada konteks tertentu. Adapun durasi yang dihabiskan selama proses *fine-tuning* model, serta waktu yang diperlukan untuk menghasilkan jawaban dari 570 data uji pada model *base* dan RAG *fine-tuning*, dapat dilihat pada Tabel 6.

Tabel 6 Durasi pemrosesan

No	Keterangan	Durasi
1	*Fine-tuning* model Komodo-7B	4 jam 19 menit 8 detik

2	Proses generasi jawaban (Komodo-7B *base*)	1 jam 19 menit 58 detik
3	Proses generasi jawaban (RAG *fine-tuning* Komodo-7B)	2 jam 23 menit 1

Berdasarkan Tabel 6, menunjukkan bahwa proses *fine-tuning* memerlukan waktu yang cukup signifikan, yaitu lebih dari empat jam. Sementara itu, waktu generasi jawaban pada model *base* relatif lebih singkat dibandingkan dengan RAG *fine-tuning*. Perbedaan ini dapat disebabkan oleh kompleksitas tambahan pada RAG *fine-tuning*, seperti proses pengambilan informasi (*retrieval*) sebelum generasi jawaban, yang menambah beban komputasi. Meskipun demikian, waktu eksekusi yang lebih lama pada RAG *fine-tuning* berpotensi menghasilkan keluaran dengan kualitas yang lebih relevan terhadap konteks pertanyaan. Mendukung hal tersebut, pada Gambar 4 menyajikan representasi visual hasil evaluasi metrik ROUGE-1, ROUGE-L, dan METEOR terhadap hasil generasi model.



Gambar 4 Perbandingan Hasil Metrik Evaluasi

Berdasarkan hasil evaluasi, seluruh metrik menunjukkan bahwa RAG *fine-tuning* Komodo-7B secara konsisten melampaui performa Komodo-7B *Base*. Pada ROUGE-1, model RAG *fine-tuning* memperoleh *precision* 0,319, *recall* 0,375, dan F1-*score* 0,299, meningkat signifikan dibandingkan *base* yang hanya 0,050, 0,099, dan 0,058. Pola serupa terlihat pada ROUGE-L, di mana *precision*, *recall*, dan F1-*score* RAG *fine-tuning* masing-masing 0,267, 0,319, dan 0,251, sedangkan *base* hanya 0,039, 0,076, dan 0,045. METEOR juga naik dari 0,057 menjadi 0,275. Meskipun terdapat peningkatan, mempertimbangkan standar, hasil ini jelas belum dapat dikategorikan "baik". Salah satu faktor yang menyebabkan rendahnya nilai evaluasi metrik adalah adanya sejumlah pertanyaan yang tidak dapat dijawab secara tepat, sebagaimana ditunjukkan pada Tabel 7.

Tabel 7 Hasil generasi jawaban yang tidak sesuai

No	Pertanyaan	Refrensi	Hasil RAG *fine-tuning* Komodo-7B
1	kk yang asli sekarang tu bagaimana ya	KK saat ini memang hitam putih pak, jadi memang tidak ada bedanya dengan fotocopyan, kalau mau mengecek data KK sudah valid bisa menscan barcode yang ada di KK nya	KK yang asli sekarang tu bagaimana ya Jawa b pertanyaan : kk yang asli sekarang tu bagaimana ya Jawa b jawaban : KK yang asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang

			asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang asli sekarang tu bagaimana ya KK yang
2	Perubahan data pendidikan di KK bisa dilakukan di Siola?	Tidak perlu ke Siola, cukup melalui Klampid atau kelurahan.	Perlu dokumen pendukung berupa ijazah.

Pada Tabel 7, hasil jawaban pertanyaan pertama menunjukkan bahwa model hanya melakukan pengulangan pertanyaan dan tidak menghasilkan jawaban. Lalu hasil jawaban kedua tidak sesuai dengan pertanyaan. Kemudian, *recall* yang lebih tinggi dibanding *precision* pada RAG menunjukkan bahwa model menangkap lebih banyak kata atau frasa yang berkaitan dengan referensi, namun sebagian juga berisi tambahan atau reformulasi kalimat yang tidak diakui oleh metrik sebagai relevan. Akibatnya, *precision* menjadi lebih rendah. Hal tersebut tidak selalu menandakan bahwa jawaban kurang tepat, sering kali sebenarnya model menjawab dengan tepat, namun metrik evaluasi tidak mengenalinya. Selain itu, terkadang juga pemisahan token seperti spasi atau tanda baca yang tidak tepat sehingga menurukan metrik evaluasi. Beberapa contoh kasus yang seusai terdapat pada Tabel 8.

Tabel 8 Contoh kasus hasil generasi jawaban

No	Pertanyaan	Refrensi	Hasil RAG *fine-tuning* Komodo-7B
1	Apa saja berkas yang perlu saya bawa untuk foto ulang KTP?	Anda perlu membawa KK dan KTP lama.	1. KK dan KTP lama 2. Mengajukan ke kecamatan atau siola untuk perekaman ulang , dengan membawa KK dan KTP lama , lalu datang sesuai undangan untuk perekaman ulang , selanjutnya dapat menunggu KTP fisik selesai .
2	Jika ingin mengurus cetak ulang KK yang rusak, menu apa yang digunakan?	Silakan gunakan menu "Cetak Ulang Kartu Keluarga" di Klampid untuk pengajuan cetak ulang KK	Jika ingin men guru s ce tak ulang KK yang rusak , menu yang digunakan adalah "Ce tak Ulang Kartu Keluarga ".

Selanjutnya, pada Tabel 8, jawaban pertanyaan pertama masih berada dalam konteks yang relevan, namun model menghasilkan jawaban dengan struktur dan pilihan kata yang berbeda dari referensi. Pada pertanyaan kedua, hasil *generate* jawaban menunjukkan kekurangan pada token yang dihasilkan, sehingga kata yang terbentuk menjadi tidak lengkap, misalnya “men guru s ce tak”. Kedua hal tersebut yang juga mempengaruhi *precision* yang lebih rendah dari *recall*. Jika dibandingkan dengan penelitian sebelumnya oleh Hakim dkk., hasil eksperimen menunjukkan bahwa Komodo-7B yang di-*fine-tune* dengan LoRA mencapai kinerja terbaik dengan skor ROUGE-L sebesar 0.3529, mengungguli model lainnya. Jika dibandingkan, RAG *fine-tuning* Komodo-7B pada penelitian ini yangmenghasilkan nilai ROUGE-L 0.251 memang berada jauh di bawah capaian LoRA pada Hakim dkk. Namun, hal tersebut tidak selalu berarti model dalam penelitian ini lebih buruk, melainkan menunjukkan bahwa model cenderung tidak mempertahankan urutan kata atau struktur panjang seperti di referensi. Model lebih memilih memformulasi ulang sehingga membuat evaluasi metrik yang mengandalkan l*ongest common subsequence* ini menjadi lebih rendah.

.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa *fine-tuning* model Komodo-7B dengan pendekatan *Retrieval-Augmented Generation* (RAG) mampu meningkatkan performa dibandingkan versi *base* pada seluruh metrik evaluasi (ROUGE-1, ROUGE-L, dan METEOR), dengan skor ROUGE-1 *precision* 0,319, *recall* 0,375, dan F1-*score* 0,299; ROUGE-L *precision* 0,267, *recall* 0,319, dan F1-*score* 0,251; serta METEOR 0,275. Meskipun peningkatan terhadap *base* signifikan, nilai tersebut masih tergolong rendah menurut standar. Kecenderungan *recall* yang lebih tinggi dibanding *precision* mengindikasikan *coverage* yang baik namun disertai keluaran yang mengandung reformulasi, padanan kata, atau penambahan konteks yang tidak terdeteksi sebagai relevan oleh metrik, sehingga *precision* dan F1 menurun meskipun secara semantik jawaban bisa jadi sesuai. Selain itu, performa turut dipengaruhi kesalahan pemisahan token dan adanya pertanyaan yang tidak dijawab dengan tepat. Dibandingkan penelitian terdahulu (Hakim dkk.), skor ROUGE-L lebih rendah (25,1 vs 35,29) yang tidak serta-merta menunjukkan penurunan kualitas, melainkan menggambarkan kecenderungan model untuk memformulasi ulang susunan kata daripada mempertahankan urutan kalimat panjang seperti referensi, sehingga mendapat penalti lebih besar pada evaluasi berbasis *longest common subsequence*. Untuk penelitian lanjutan, disarankan integrasi teknik *answer filtering* untuk meningkatkan *precision* tanpa mengorbankan *recall*, pembatasan keluaran (*output constraining*) untuk mengurangi redundansi, *fine-tuning* dengan format data beranotasi struktur konten agar model mempertahankan urutan dan alur kalimat sesuai referensi. Selain itu juga dapat ditambahkan penerapan *Reinforcement Learning from Human Feedback* (RLHF) untuk menyelaraskan keluaran dengan ekspektasi pengguna sekaligus menjaga kepatuhan terhadap kebijakan publik.

REFERENSI

[1] S. S. M. Wara, A. F. Adziima, M. Nasrudin, dan A. R. Pratama, "Predictive Analysis of Government Application Comment on Playstore with Clustered Support Vector Machine," dalam *2024 IEEE 10th Information Technology International Seminar (ITIS)*, IEEE, Nov 2024, hlm. 84–88. doi: 10.1109/ITIS64716.2024.10845453.

[2] Pemerintah Kota Surabaya, "Peraturan Walikota Surabaya Nomor 80 Tahun 2021," 2021. Diakses: 7 Januari 2025. [Daring]. Tersedia pada: https://jdih.surabaya.go.id/peraturan/3966

[3] A. Muhaimin, I. A. Taufik, dan D. D. Daniswara, "Pendeteksian Spam pada E-mail menggunakan Pendekatan Natural Language Processing," *PROSIDING SEMINAR NASIONAL SAINS DATA*, vol. 3, no. 1, hlm. 116–121, Nov 2023, doi: 10.33005/senada.v3i1.90.

[4] I. G. S. M. Diyasa, "Implementation Of Natural Language Processing for Spam Email Detection in Outcome Based Education (OBE) Application," *IJEBD (International Journal of Entrepreneurship and Business Development)*, vol. 6, no. 6, hlm. 1166–1171, Nov 2023, doi: 10.29138/ijebd.v6i6.2587.

[5] R. Aprilia, "Sistem tanya jawab ilmu keislaman dengan model Large Language Models," Universitas Islam Negeri Sultan Syarif Kasim Riau, 2024. Diakses: 13 April 2025. [Daring]. Tersedia pada: http://repository.uin-suska.ac.id/id/eprint/79122

[6] S. A. Hakim, R. S. Perdana, dan T. N. Fatyanosa, "Anak Baik: A Low-Cost Approach to Curate Indonesian Ethical and Unethical Instructions," dalam *Proceedings of the Second Workshop in South East Asian Language Processing*, D. Wijaya, A. F. Aji, C. Vania, G. I. Winata, dan A. Purwarianti, Ed., Online: Association for Computational Linguistics, Jan 2025, hlm. 52–62. [Daring]. Tersedia pada: https://aclanthology.org/2025.sealp-1.5/

[7] A. Abdulnazar, R. Roller, S. Schulz, dan M. Kreuzthaler, "Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization," *IEEE Access*, vol. 12, hlm. 147981–147990, Okt 2024, doi: 10.1109/ACCESS.2024.3472500.

[8] H. K. Chaubey, G. Tripathi, R. Ranjan, dan S. K. Gopalaiyengar, "Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development," dalam *ICFTSS 2024 - International Conference on Future Technologies for Smart Society*, Institute of Electrical and Electronics Engineers Inc., Agu 2024, hlm. 169–172. doi: 10.1109/ICFTSS61109.2024.10691338.

[9] P. A. Riyantoko dan A. Muhaimin, "A Simple Data Sentiment Analysis using Bjorka phenomenon on Twitter," dalam *Nusantara Science and Technology Proceedings*, Galaxy Science, Mei 2023. doi: 10.11594/nstp.2023.3353.

[10] P. Haryani, N. T. Putri, dan L. M. Jannah, "Bandung Sadayana: Partisipasi Digital Masyarakat Kota Bandung dalam Membangun Smart City," *VISA: Journal of Vision and Ideas*, vol. 4, no. 1, hlm. 102–121, Jan 2024, doi: 10.47467/visa.v4i1.5833.

[11] R. Jiwandono, "Yellow.ai meluncurkan Komodo-7B, LLM pertama di Indonesia yang dilatih 11 bahasa daerah," Techverse.asia. Diakses: 10 April 2025. [Daring]. Tersedia pada: https://www.techverse.asia/techno/6358/08032024/yellowai-meluncurkan-komodo-7b-llm-pertama-di-indonesia-yang-dilatih-11-bahasa-daerah

[12] R. Puspita Sari, "Komodo-7B: Model AI multibahasa terbaru untuk bahasa daerah," CloudComputing.id. Diakses: 16 April 2025. [Daring]. Tersedia pada: https://www.cloudcomputing.id/berita/komodo-7b-ai-multibahasa

[13] L. Owen, V. Tripathi, A. Kumar, dan B. Ahmed, "Komodo: A Linguistic Expedition into Indonesia's Regional Languages," *arXiv preprint arXiv:2403.09362*, Mar 2024, [Daring]. Tersedia pada: https://doi.org/10.48550/arXiv.2403.09362

[14] T. Dettmers, A. Pagnoni, A. Holtzman, dan L. Zettlemoyer, "QLORA: efficient finetuning of quantized LLMs," dalam *Proceedings of the 37th International Conference on Neural Information Processing Systems*, dalam NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023. [Daring]. Tersedia pada: https://dl.acm.org/doi/10.5555/3666122.3666563

[15] M. Yazan, S. Verberne, dan F. Situmeang, "The Impact of Quantization on Retrieval-Augmented Generation: An Analysis of Small LLMs," *arXiv preprint arXiv:2406.10251*, Jun 2024, [Daring]. Tersedia pada: https://doi.org/10.48550/arXiv.2406.10251

[16] C. Galli, N. Donos, dan E. Calciolari, "Performance of 4 Pre-Trained Sentence Transformer Models in the Semantic Query of a Systematic Review Dataset on Peri-Implantitis," *Information (Switzerland)*, vol. 15, no. 2, Feb 2024, doi: 10.3390/info15020068.

[17] T. M. Fahrudin, M. H. Hartanto, A. S. Paramita, A. Aulia, R. A. Maulana, dan I. R. Anniswa, "TEMU KEMBALI INFORMASI BERITA KEGIATAN PROGRAM STUDI MENGGUNAKAN ALGORITMA PEMBOBOTAN TF-IDF DAN COSINE SIMILARITY," *Prosiding Seminar Nasional Teknologi dan Sistem Informasi*, vol. 2, no. 1, hlm. 270–279, Sep 2022, doi: 10.33005/sitasi.v2i1.309.

[18] Junadhi, Agustin, L. Efrizoni, F. Okmayura, D. R. Habibie, dan Muslim, "Improving Evaluation Metrics for Text Summarization: A Comparative Study and Proposal of a Novel Metric," *Journal of Applied Data Sciences*, vol. 6, no. 2, hlm. 885–896, Mei 2025, doi: 10.47738/jads.v6i2.547.

[19] A. Al Foysal dan R. Böck, "Who Needs External References?—Text Summarization Evaluation Using Original Documents," *AI (Switzerland)*, vol. 4, no. 4, hlm. 970–995, Des 2023, doi: 10.3390/ai4040049.