

PENERAPAN ALGORITMA COSINE SIMILARITY PADA CHATBOT PELAYANAN MASYARAKAT BERBASIS RETRIEVAL AUGMENTED GENERATION

Gilang Wahyu Pradana, Moh. Muhtarom, Aprilisa Arum Sari

Teknik Informatika, Universitas Duta Bangsa Surakarta

Jl. Bhayangkara No. 55-57, Tipes, Serengan, Kota Surakarta, Jawa Tengah 57154

220103125@mhs.udb.ac.id

ABSTRAK

Pelayanan publik di tingkat kecamatan seringkali menghadapi kendala keterlambatan respon informasi administrasi kependudukan. Permasalahan ini menyebabkan masyarakat harus datang langsung ke kantor hanya untuk menanyakan persyaratan dasar, sehingga memicu antrean fisik dan ketidakefisienan waktu. Penelitian ini bertujuan merancang chatbot cerdas untuk menyediakan akses informasi pelayanan publik 24 jam di Kecamatan Polokarto. Metode yang diterapkan adalah Retrieval-Augmented Generation (RAG) guna mengunci jawaban pada dokumen legal, yang dipadukan dengan algoritma Cosine Similarity untuk menghitung kedekatan semantik antara pertanyaan kasual kuesioner masyarakat dengan dokumen SOP resmi. Sistem dibangun menggunakan framework Laravel, Node.js, dan vector database Supabase. Hasil pengujian akurasi terhadap 60 skenario pertanyaan riil dari kuesioner menunjukkan bahwa sistem memperoleh tingkat keberhasilan sebesar 95%, dengan 57 pertanyaan terjawab secara valid. Selain itu, implementasi filter otomatis failure rate 80% terbukti handal dalam menolak pertanyaan di luar konteks guna menghindari risiko informasi tidak valid.

Kata kunci: Chatbot, Cosine Similarity, Laravel, RAG, Supabase.

1. PENDAHULUAN

Transformasi digital dalam sektor pelayanan publik menjadi agenda utama pemerintah Indonesia untuk meningkatkan efisiensi dan transparansi [1]. Di tingkat pemerintahan daerah, khususnya Kantor Kecamatan Polokarto, penyampaian informasi mengenai prosedur administrasi kependudukan seperti pembuatan Kartu Tanda Penduduk (KTP), Kartu Keluarga (KK), akta kelahiran, surat pindah, dispensasi nikah, hingga keterangan ahli waris masih sering dilakukan secara konvensional. Masyarakat cenderung harus datang langsung ke kantor kecamatan atau menghubungi petugas secara manual hanya untuk menanyakan persyaratan dasar [2]. Kondisi tersebut berdampak pada terjadinya antrean fisik yang menumpuk, keterbatasan waktu operasional bagi warga, serta tingginya beban kerja administratif yang bersifat repetitif bagi petugas pelayanan di lapangan [3].

Perkembangan teknologi Artificial Intelligence (AI) melalui Large Language Models (LLM) sebenarnya menawarkan solusi berupa chatbot yang mampu berinteraksi secara natural selama 24 jam penuh [4],[5]. Namun, penggunaan AI generatif murni di sektor publik seringkali menghadapi kendala fatal berupa risiko kekeliruan informasi atau jawaban tidak valid, di mana sistem memberikan jawaban yang terdengar meyakinkan tetapi tidak akurat secara fakta hukum dan dokumen legal [4]. Masalah akurasi ini menjadi sangat krusial mengingat informasi administrasi kependudukan berkaitan erat dengan kepastian hukum dan legalitas dokumen masyarakat [1].

Penelitian ini bertujuan untuk merancang dan membangun sistem chatbot pelayanan masyarakat yang cerdas, responsif, dan memiliki tingkat akurasi

tinggi di Kecamatan Polokarto. Melalui sistem ini, masyarakat diharapkan dapat memperoleh informasi persyaratan administrasi secara mandiri tanpa harus terikat jam kerja kantor formal. Selain itu, penelitian ini bertujuan mengimplementasikan kontrol keamanan ketat berupa filter otomatis untuk menolak memberikan jawaban spekulatif apabila pertanyaan yang diajukan berada di luar konteks dokumen pelayanan publik.

Rencana penyelesaian masalah dilakukan dengan menerapkan metode Retrieval-Augmented Generation (RAG) yang dikombinasikan dengan algoritma Cosine Similarity [6]. Melalui arsitektur RAG, chatbot tidak akan mengandalkan pengetahuan internal model AI generatif murni, melainkan secara aktif mencari dokumen rujukan dari basis pengetahuan resmi sebelum memproduksi teks jawaban [7]. Algoritma Cosine Similarity diterapkan sebagai metrik utama pencarian semantik untuk mengukur sudut kedekatan teks berdasarkan orientasi vektor [6].

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian yang ditulis oleh Muhammad Roihan Rachman, Muhammad Rosidin, Wicaksono Yuli Sulistyo pada tahun 2025 dengan judul “Implementasi Metode Retrieval Augmented Generation Pada Chatbot Untuk Otomatisasi Layanan Pelanggan Kontrakan” membuktikan bahwa penerapan metode RAG pada chatbot secara signifikan meningkatkan relevansi dan ketepatan jawaban. Hasil penelitian mereka menunjukkan bahwa chatbot berbasis RAG mampu merespons lebih cepat daripada layanan manual dan sangat efektif dalam mengurangi beban kerja operasional dalam pemberian informasi yang berulang [4].

Penelitian yang ditulis oleh Bagas Wahyu Herdiansyah, Mukhammad Andri Setiawan pada tahun 2025 dengan judul “Implementasi RAG Multimodal Untuk Rekomendasi Teks Dan Gambar Di E-Commerce” menjelaskan keunggulan RAG dalam meminimalisir risiko “halusinasi” pada model bahasa dengan cara menyediakan konteks yang relevan dari basis data eksternal. Pendekatan ini selaras dengan kebutuhan penulis untuk menyajikan informasi prosedur administrasi kependudukan yang harus memiliki tingkat akurasi tinggi [7].

Penelitian yang ditulis oleh Tarisa Ramadhani, Noora Qotrun Nada, Nugroho Dwi S pada tahun 2025 dengan judul “Penerapan Metode Retrieval-Augmented Generation (RAG) Pada Chatbot E-Commerce Berbasis Gemini AI” mempertegas fleksibilitas arsitektur RAG yang menggunakan model bahasa besar LLM untuk menghasilkan jawaban yang natural. Hal ini mendukung rencana penulis dalam membangun sistem yang adaptif terhadap perubahan dokumen regulasi tanpa perlu melakukan pelatihan ulang model [5].

Penelitian yang ditulis oleh Sarima Lumbansiantar, Saruni Dwiasnati, Nenden Siti Fatonah pada tahun 2023 dengan judul “Penerapan Metode Cosine Similarity dalam Mendeteksi Plagiarisme pada Jurnal” menjelaskan bahwa algoritma Cosine Similarity bekerja dengan menghitung sudut antara dua objek dalam bentuk vektor untuk menentukan tingkat kemiripan teks. Metode ini menjadi pilihan Peneliti karena kemampuannya dalam mendeteksi relevansi dokumen meskipun terdapat perbedaan diksi antara pertanyaan masyarakat dengan dokumen resmi [3].

Berdasarkan beberapa penelitian sebelumnya yang telah dipaparkan, penulis membuat sebuah pembaruan sistem yang mengintegrasikan berbagai kelebihan dari masing-masing penelitian tersebut ke dalam satu arsitektur yang utuh. Sistem yang dikembangkan dalam penelitian ini menggabungkan keunggulan efisiensi otomatisasi layanan dari penelitian Rachman dkk, kemampuan mitigasi halusinasi berbasis data eksternal dari penelitian Herdiansyah & Setiawan, serta fleksibilitas respons bahasa alami yang adaptif dari penelitian Ramadhani dkk. Keunggulan-keunggulan tersebut disempurnakan dengan penerapan algoritma Cosine Similarity seperti pada penelitian Lumbansiantar dkk untuk menjamin akurasi pencarian dokumen SOP tingkat kecamatan yang kaku dan tekstual. Berbeda dengan penelitian-penelitian terdahulu yang umumnya berfokus pada sektor komersial (*e-commerce* dan layanan pelanggan swasta), sistem *Chatbot* ini dirancang khusus untuk sektor pelayanan publik pemerintahan tingkat kecamatan dengan menggabungkan kerangka kerja Laravel sebagai backend utama, Node.js untuk pemrosesan kecerdasan buatan via *LangChain*, serta Supabase sebagai *vector database*. Dengan cara ini, diharapkan sistem tidak hanya mampu menyajikan informasi prosedur administrasi kependudukan

(seperti KTP, KK, dan akta kelahiran) secara *real-time* dan responsif, tetapi juga memiliki tingkat akurasi yang sangat tinggi, konsisten, serta terhindar dari risiko misinformasi atau jawaban palsu yang dapat merugikan masyarakat luas.

2.2. Landasan Teori

Chatbot adalah program komputer yang berfungsi untuk mensimulasikan percakapan layaknya manusia melalui perintah suara, teks, ataupun gambar. *Chatbot* dikenal juga sebagai asisten virtual, *chatbot* mampu memahami permintaan pengguna dan merespon dengan jawaban yang tepat. *Chatbot* bekerja untuk menggantikan peranan manusia dalam melayani pembicaraan melalui website maupun aplikasi untuk membantu komunikasi dengan para pengunjung website ataupun aplikasi tersebut [8].

Large Language Model (LLM) merupakan model kecerdasan buatan yang dikembangkan untuk memahami dan menghasilkan teks melalui kemampuan pemrosesan bahasa alami yang tinggi, serta dilatih menggunakan data berskala besar. Dalam implementasinya, LLM diintegrasikan ke dalam sistem melalui sebuah *microservice* yang dihubungkan dengan *Application Programming Interface* (API) [9].

Natural Language Processing (NLP) adalah teknologi yang memungkinkan komputer untuk memahami, menafsirkan, dan menganalisis bahasa manusia. Dalam arsitektur *chatbot*, NLP berfungsi sebagai jembatan komunikasi yang mengubah input teks tidak terstruktur dari pengguna menjadi format yang dapat dipahami oleh sistem komputer [10].

Retrieval-Augmented Generation (RAG) adalah pendekatan inovatif yang menggabungkan kekuatan LLM dengan kemampuan pencarian informasi eksternal [11]. Secara teknis, alur kerja RAG dalam penelitian ini dibagi menjadi dua fase utama yaitu

2.3. Fase Ingesti (Persiapan Data):

- Dokumen: Dokumen Pelayanan Umum (SOP KK, KTP, dll) dikumpulkan.
- Chunking*: Dokumen dipecah menjadi bagian-bagian kecil agar mudah dikelola.
- Embedding*: Setiap potongan teks diubah menjadi vektor numerik.
- Vector Database*: Vektor disimpan di Supabase sebagai basis pengetahuan.

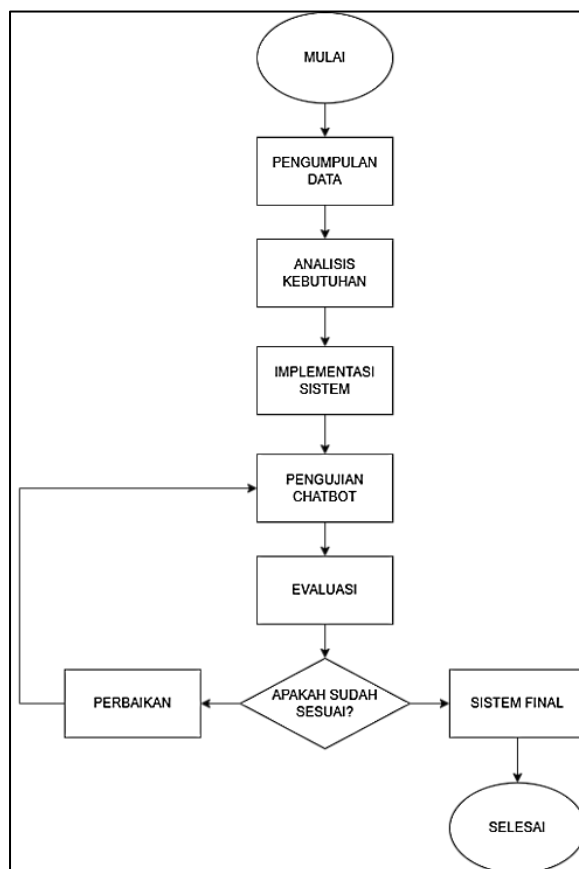
2.4. Fase Inferensi (Saat Pengguna Bertanya):

- Retrieval*: Saat masyarakat bertanya, sistem mencari potongan dokumen yang paling mirip menggunakan *Cosine Similarity*.
- Augmentation*: Potongan dokumen yang ditemukan digabungkan dengan pertanyaan asli sebagai konteks tambahan.
- Generation*: LLM menerima pertanyaan beserta konteks tersebut untuk menghasilkan jawaban yang akurat dan berbasis data [12].

Cosine similarity merupakan salah satu metrik yang banyak digunakan untuk mengukur kesamaan antar vektor, terutama pada ruang berdimensi tinggi seperti data teks, citra, maupun data pengambilan keputusan. Secara definisi, *cosine similarity* mengukur nilai kosinus dari sudut antara dua vektor tak nol dengan menitikberatkan pada orientasi, bukan pada besarannya. Sifat ini menjadikannya efektif untuk data yang bersifat sparse dan berdimensi tinggi, seperti dokumen teks, panjang vektor sering kali tidak penting arah dari representasi vektor tersebut [12][6].

3. METODE PENELITIAN

Sistem dikembangkan menggunakan metode yang terstruktur untuk memastikan bahwa seluruh tahapan pembangunan chatbot berjalan sesuai dengan tujuan penelitian. Pengembangan sistem *Chatbot* dilakukan secara bertahap mengikuti siklus hidup perancangan sistem cerdas berbasis pencarian dokumen. Prosedur perancangan digambarkan melalui diagram alir (*flowchart*) berikut:



Gambar 1. Flowchart perancangan sistem chatbot

3.1. Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini dikategorikan menjadi dua jenis, yaitu data primer dan data sekunder, dengan rincian sebagai berikut:

3.2. Data Primer

Data primer merupakan data yang diperoleh penulis secara langsung melalui observasi di lokasi penelitian untuk kebutuhan pengujian sistem. Data ini berupa daftar pertanyaan riil yang merepresentasikan kebutuhan informasi masyarakat terkait Bagian Pelayanan Umum di Kecamatan Polokarto. Data ini mencakup pertanyaan mengenai persyaratan pembuatan Kartu Keluarga (KK), KTP, akta kelahiran, surat pindah, hingga layanan khusus seperti dispensasi nikah dan keterangan ahli waris. Data ini diperoleh melalui observasi langsung terhadap alur konsultasi dan identifikasi pertanyaan yang sering muncul (*Frequently Asked Questions*) pada media sosial instansi.

3.3. Data Sekunder

Data sekunder adalah data pendukung yang diperoleh dari dokumen internal yang sudah ada untuk digunakan sebagai basis pengetahuan (*knowledge base*). Data sekunder dalam penelitian ini berupa dokumen resmi yang memuat prosedur dan persyaratan layanan publik di Kantor Kecamatan Polokarto, mencakup pembuatan Kartu Keluarga (KK), KTP, akta kelahiran, surat pindah, hingga layanan khusus seperti dispensasi nikah dan keterangan ahli waris. Dokumen-dokumen ini merupakan data krusial yang akan diproses menjadi *vector embedding* agar sistem RAG memiliki referensi yang valid dan terhindar dari informasi palsu.

3.1. Metode Pengumpulan Data

Metode pengumpulan data dalam penelitian ini dilakukan melalui studi pustaka dan dokumentasi. Studi pustaka dilakukan dengan mempelajari literatur ilmiah seperti jurnal, buku, dan artikel penelitian yang berkaitan dengan *chatbot*, *Retrieval-Augmented Generation* (RAG), algoritma *Cosine Similarity*, serta metode evaluasi sistem berbasis *Large Language Model* (LLM).

Metode dokumentasi dilakukan dengan mengumpulkan dokumen resmi pelayanan publik yang digunakan sebagai sumber data sekunder. Dokumen tersebut kemudian dianalisis dan diproses untuk membentuk basis data vektor yang akan digunakan dalam proses *retrieval*.

3.2. Analisis Kebutuhan Sistem

Tahap ini dilakukan untuk memetakan seluruh spesifikasi komponen yang diperlukan agar sistem *chatbot* dapat berjalan optimal dan responsif.

3.3. Kebutuhan Fungsional

- Sistem harus mampu menerima pesan berbasis bahasa alami dari masyarakat pengguna.
- Sistem harus dapat melakukan pencarian semantik terhadap dokumen SOP yang relevan dalam hitungan milidetik.
- Sistem harus memiliki kemampuan filter otomatis (80% kegagalan) untuk menolak

memberikan jawaban apabila pertanyaan berada di luar konteks dokumen legal pelayanan publik.

- d. Sistem harus menyediakan dashboard pemantauan untuk admin guna mengamati total interaksi, rata-rata skor kemiripan, grafik tren akurasi, dan log riwayat pesan.

3.4. Kebutuhan Non-Fungsional

- a. Konsistensi jawaban AI yang harus selalu merujuk pada regulasi resmi.
- b. Kecepatan pemrosesan data (*response time*) dan stabilitas koneksi microservice.

3.5. Kebutuhan Perangkat Lunak

- a. *Framework* Laravel sebagai *backend* utama dan penyaji antarmuka pengguna.
- b. Node.js dengan *framework* LangChain sebagai *microservice* pemrosesan logika AI dan RAG.
- c. Supabase sebagai *vector database* untuk menyimpan hasil *embedding* teks SOP.
- d. *Large Language Model* (LLM) sebagai mesin generator teks respons bahasa alami.
- e. Visual Studio Code sebagai teks editor utama untuk menulis, mengedit, dan melakukan debugging kode program.
- f. *Web Browser* sebagai media untuk menguji antarmuka *chatbot* lokal serta mengakses halaman *dashboard* monitoring admin.

3.6. Implementasi Sistem

Tahap implementasi dilakukan dengan mengintegrasikan *framework* Laravel, Node.js (*LangChain*), dan *database* vektor Supabase. Proses pencarian semantik dokumen menggunakan perhitungan sudut kedekatan teks berdasarkan algoritma *Cosine Similarity* dengan formula matematika berikut:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Dimana:

$A \cdot B$ (hasil kali skalar) antara dua vektor A dan B
 $\|A\|$ adalah magnitude (panjang) dari vektor A
 $\|B\|$ adalah magnitude (panjang) dari vektor B

3.7. Perencanaan Pengujian dan Evaluasi

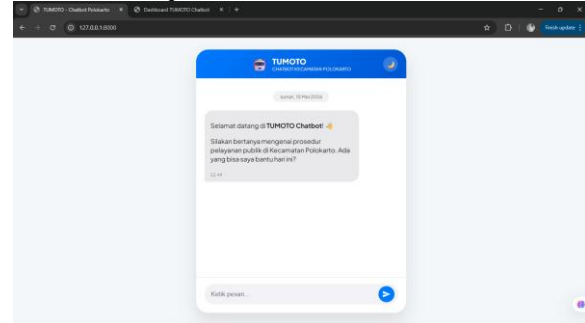
Pengujian dilakukan secara terukur menggunakan dua pendekatan utama:

- a. *Black Box Testing*: Digunakan untuk memvalidasi fungsionalitas murni seluruh fitur utama aplikasi, seperti fitur kirim pesan, fungsi retrieval *database*, ketepatan filter luar konteks, dan penyimpanan riwayat chat.
- b. *Accuracy Testing*: Pengujian kecerdasan dilakukan melalui simulasi mandiri menggunakan data primer berupa 60 variasi skenario pertanyaan masyarakat yang terbagi ke dalam 7 kategori layanan administrasi di Kecamatan Polokarto. Validasi respons jawaban

diukur secara otomatis menggunakan modul evaluasi AI Judge.

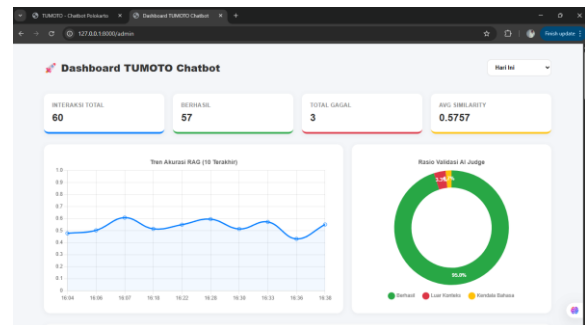
4. HASIL DAN PEMBAHASAN

4.1. Hasil Implementasi



Gambar 2. Antarmuka Chatbot

Gambar 2 menunjukkan antarmuka pengguna (User Interface) chatbot yang digunakan oleh masyarakat untuk mengajukan pertanyaan terkait layanan administratif. Desain mengadopsi konsep *rounded chat bubbles* bergaya aplikasi pesan instan guna meningkatkan aksesibilitas dan kenyamanan pengguna. Pada tampilan ini, warga dapat mengetikkan pertanyaan secara natural, dan sistem akan memberikan jawaban secara real-time berdasarkan rujukan dokumen SOP yang relevan.



Gambar 3. Dashboard Monitoring

Gambar 3 ini berfungsi sebagai pusat kendali administratif untuk memantau performa sistem secara keseluruhan. Tampilan ini menyajikan metrik kunci (*Key Performance Indicators*) seperti Statistik Interaksi yang menampilkan total interaksi, jumlah jawaban berhasil, dan jumlah jawaban gagal. Metrik *Similarity* yang menampilkan rata-rata skor *Cosine Similarity* untuk mengukur kedekatan antara pertanyaan dan konteks. Visualisasi grafik garis yang menunjukkan tren akurasi RAG untuk 10 interaksi terakhir serta *Pie Chart* yang menggambarkan rasio validasi oleh AI Judge. Sistem Filter yang dapat memfilter data performa berdasarkan rentang waktu tertentu, mulai dari hari ini, 7 hari terakhir, 30 hari terakhir, hingga keseluruhan waktu (*all time*) untuk keperluan evaluasi berkala.

Waktu	Pertanyaan	Similarity	Status
15/05/2025 16:30	apa apa persyaratan pindah antar kabupaten dan provinsi	0.5504	Berhasil
15/05/2025 16:30	apa perbedaan pindah antar kabupaten dan provinsi	0.4307	Gagal - Luar Konteks
15/05/2025 16:33	Berapa lama waktu yang dibutuhkan untuk mengurus surat pindah antar kabupaten dan provinsi?	0.5717	Berhasil
15/05/2025 16:30	Berapa ukuran dan jenis foto yang diperlukan untuk pindah antar kabupaten dan provinsi?	0.5128	Berhasil
15/05/2025 16:28	Formulir apa yang harus diisi untuk pindah penduduk antar kabupaten atau provinsi?	0.5949	Berhasil
15/05/2025 16:22	syarat surat kelahiran wajib	0.5489	Berhasil
15/05/2025 16:18	Surat Keterangan Waris biasanya digunakan untuk keperluan apa?	0.5158	Berhasil
15/05/2025 16:07	Siapa saja yang harus menandatangani surat Keterangan Waris di tingkat bawah?	0.6072	Berhasil
15/05/2025 16:00	Apakah Surat Keterangan persara diperlukan untuk membuat SKWT?	0.5816	Berhasil
15/05/2025 16:04	Dalam pengurusan Surat Keterangan Waris (SKW), dokumen apa yang membuktikan hubungan keluarga?	0.4772	Gagal - Luar Konteks
15/05/2025 16:06	Apakah surat yang dibawa pindah pemohon wajib untuk pindah kecamatan sesuai?	0.6155	Berhasil
15/05/2025 16:05	Berapa lama waktu yang dibutuhkan untuk memberi nomor agenda dan stamp pada surat pindah?	0.5884	Berhasil

Gambar 4. Dashboard Log Riwayat

Gambar 4 menampilkan halaman data log riwayat pada dashboard administratif yang menyajikan seluruh interaksi dalam sistem dalam bentuk tabel dinamis. Kolom pada tabel ini mencakup Waktu untuk mencatat stempel waktu (*timestamp*) setiap pesan masuk, Pertanyaan untuk menampilkan input teks mentah dari pengguna. *Similarity* untuk menampilkan

skor numerik hasil kalkulasi algoritma *Cosine Similarity* antara pertanyaan dengan dokumen rujukan. Status untuk menandai apakah jawaban berhasil divalidasi oleh sistem atau masuk dalam kategori gagal/di luar konteks.

4.2. Hasil Pengujian Aplikasi

Pengujian dilakukan untuk mengukur sejauh mana sistem mampu memberikan jawaban yang akurat berdasarkan dokumen yang ada. Pengujian dibagi menjadi dua tahap: *Black Box Testing* untuk fungsionalitas dan *Accuracy Testing* untuk validitas jawaban.

4.3. Hasil Pengujian Black Box

Hasil pengujian fungsional menunjukkan bahwa seluruh fitur utama berjalan dengan baik sesuai rancangan sistem. Hasil pengujian fungsional ini akan ditampilkan pada tabel 1 berikut ini.

Tabel 1. Hasil Pengujian Balck Box

No	Fitur	Skenario Uji	Hasil Diharapkan	Status
1	Kirim Pesan	Input pertanyaan SOP	Jawaban muncul di layar	Berhasil
2	Retrieval	Tanya syarat KK	Mengambil data dari database	Berhasil
3	Filter Konteks	Tanya diluar SOP	Muncul pesan "Maaf, data tidak ditemukan"	Berhasil
4	Riwayat Chat	Refresh halaman	Pesan sebelumnya tetap ada	Berhasil

4.4. Hasil Pengujian Akurasi

Analisis akurasi kecerdasan sistem dilakukan menggunakan instrumen uji berupa 60 variasi skenario pertanyaan masyarakat yang valid. Seluruh data uji tersebut diperoleh secara riil dari hasil penyebaran kuesioner digital (Google Form) kepada warga

Kecamatan Polokarto. Data uji ini dikelompokkan ke dalam 7 kategori layanan administrasi kependudukan guna mengukur ketepatan rujukan jawaban sistem chatbot. Hasil pengujian akurasi disajikan dalam Tabel 2 berikut ini.

Tabel 2. Hasil Pengujian Akurasi Berdasarkan Kategori

No	Kategori Layanan	Jumlah Uji	Benar	Salah	Akurasi (%)
1	Kartu Tanda Penduduk (KTP)	10	10	0	100%
2	Kartu Keluarga (KK)	10	10	0	100%
3	Akta Kelahiran	10	10	0	100%
4	Dispensasi Menikah	10	9	1	90%
5	Pindah antar Kecamatan	10	10	0	100%
6	Pindah Antar Kabupaten dan Provinsi	5	4	1	80%
7	Surat Warisan	5	4	1	80%
Total		60	57	3	95%

Berdasarkan data pada Tabel 2, akumulasi tingkat akurasi sistem chatbot dalam merespons pertanyaan riil dari kuesioner masyarakat mencapai 95%. Secara keseluruhan, performa sistem menunjukkan hasil yang sangat optimal, khususnya pada kategori layanan dasar kependudukan seperti KTP, KK, Akta Kelahiran, dan Pindah antar Kecamatan yang berhasil mencapai nilai kesempurnaan (100%).

4.5. Pembahasan

Penggunaan *Cosine Similarity* pada penelitian ini terbukti efektif untuk menangani data kependudukan yang bersifat tekstual dan kaku. Dengan rata-rata akurasi 95%, sistem ini berhasil memvalidasi bahwa

pendekatan RAG mampu mengeliminasi potensi halusinasi pada *chatbot*. Hal ini didukung oleh tampilan UI yang responsif, di mana gelembung chat dan penanda waktu berfungsi dengan baik untuk menjaga alur komunikasi yang jelas. Integrasi *database* vektor Supabase memungkinkan pencarian dilakukan dalam hitungan milidetik, meskipun dataset terus bertambah.

Meskipun sistem mencatatkan hasil yang sangat optimal pada kategori utama seperti KTP dan KK (100% akurasi), terdapat 3 data gagal dari total 60 skenario yang diuji. Berdasarkan hasil analisis log riwayat pada dashboard administratif, kegagalan tersebut disebabkan oleh dua faktor utama, yaitu kendala bahasa dan penyaringan luar konteks.

Kategori Dispensasi Menikah mencatatkan 1 kali kesalahan respons dari sistem, sehingga tingkat akurasi turun menjadi 90%. Kegagalan ini teridentifikasi sebagai dampak dari "Kendala Bahasa" pada grafik evaluasi AI Judge. Hal ini terjadi karena pengguna mengajukan pertanyaan menggunakan terminologi atau bahasa daerah (non-formal) yang sangat awam, sehingga struktur kata tersebut tidak mampu dikenali secara tepat oleh model NLP saat proses kalkulasi *Cosine Similarity* berlangsung. Akibatnya, sistem gagal menarik dokumen rujukan yang relevan dari *database*.

Kategori Pindah Antar Kabupaten dan Provinsi beserta Surat Warisan masing-masing mencatatkan 1 kali kesalahan respons dengan tingkat akurasi sebesar 80%. Berdasarkan rekaman data log riwayat, kedua kesalahan ini masuk dalam kategori status "Gagal - Luar Konteks". Hal ini merupakan dampak langsung dari pembatasan filter otomatis sistem sebesar 80% yang dirancang untuk mencegah halusinasi. Ketika masyarakat mengajukan pertanyaan bercabang yang tidak tercantum atau menyimpang dari dokumen hukum *Standard Operating Procedure* (SOP) resmi Kecamatan Polokarto, sistem secara ketat memutus aliran data ke generator LLM dan langsung memunculkan respons penolakan standar.

Meskipun terdapat sedikit margin kegagalan tersebut, implementasi logika filter ini justru membuktikan kekuatan arsitektur RAG dalam menjaga aspek legalitas layanan publik. Sistem terbukti lebih memilih menolak memberikan jawaban gagal di luar konteks daripada memaksakan menghasilkan informasi spekulatif yang tidak memiliki dasar hukum formal

5. KESIMPULAN DAN SARAN

Berdasarkan hasil perancangan, implementasi, dan pengujian yang telah dipaparkan, dapat disimpulkan bahwa penerapan algoritma *Cosine Similarity* dalam arsitektur *Retrieval-Augmented Generation* (RAG) pada *chatbot* pelayanan masyarakat di Kecamatan Polokarto telah berhasil dilaksanakan dengan tingkat efektivitas yang sangat optimal. Integrasi *vector database* Supabase dan kerangka kerja Laravel mampu menghasilkan asisten virtual yang responsif dalam menyajikan informasi prosedur administratif kependudukan secara akurat. Hal ini dibuktikan melalui hasil pengujian terhadap 60 skenario pertanyaan yang menunjukkan tingkat akurasi sebesar 95%, di mana sistem berhasil memberikan 57 jawaban yang relevan dan valid sesuai dokumen SOP resmi. Penggunaan metode RAG terbukti secara signifikan dapat meminimalisir risiko halusinasi informasi pada model kecerdasan buatan dengan membatasi rujukan jawaban hanya pada basis data pengetahuan yang telah terverifikasi. Untuk pengembangan sistem di masa mendatang, terdapat beberapa saran yang dapat dipertimbangkan guna meningkatkan kualitas dan jangkauan layanan, antara lain Melakukan optimalisasi pada teknik *embedding*

dan pemrosesan bahasa alami untuk meningkatkan pemahaman sistem terhadap penggunaan bahasa non-formal maupun terminologi hukum yang kompleks serta Menambahkan fitur keamanan data tingkat tinggi untuk menjamin kerahasiaan informasi kependudukan yang bersifat sensitif selama proses interaksi berlangsung.

DAFTAR PUSTAKA

- [1] E. Rahadian, T. Wijanarko, A. Putra, B. Hartono, A. A. Kuncoro, and E. Siswanto, "Pengembangan Aplikasi Chat Bot Whatsapp, Telegram, dan Website Laporan untuk Instansi Pemerintah Disperkim Pool Seroja Kota Semarang Berbasis Natural Language Processing," vol. 6, no. 1, 2026.
- [2] G. Rahardandi, M. Fauzan, and P. M. Wicaksana, "Analisis Penerapan Chatbot: Survei," vol. 4, no. 4, 2024.
- [3] S. Lumbansiantar, S. Dwiasnati, and N. S. Fatolah, "Penerapan Metode Cosine Similarity dalam Mendeteksi Plagiarisme pada Jurnal," *Jurnal Format*, vol. 12, no. 2, 2023.
- [4] M. R. Rachman, M. Rosidin, and W. Y. Sulisty, "Implementasi Metode Retrieval Augmented Generation Pada Chatbot Untuk Otomatisasi Layanan Pelanggan Kontrakan," vol. 11, no. 02, 2025.
- [5] T. Ramadhani, N. Q. Nada, and N. Dwi S, "Penerapan Metode Retrieval-Augmented Generation (RAG) Pada Chatbot E-Commerce Berbasis Gemini Ai," vol. 8, no. 2, 2025.
- [6] R. Putra, Sumema, and Z. Mustika, "Similarity Judul Tugas Akhir Menggunakan Metoda Cosine, Jaccard, Rabin-Karp Pada Jurusan TI," *JITSI: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 5, no. 4, 2024, doi: 10.62527/jitsi.5.4.291.
- [7] B. Wahyu Herdiansyah and A. Setiawan, "IMPLEMENTASI RAG MULTIMODAL UNTUK REKOMENDASI TEKS DAN GAMBAR DI E-COMMERCE," *Edusaintek: Jurnal Pendidikan, Sains dan Teknologi*, vol. 12, no. 3, pp. 2025–1767, 2025, doi: 10.47668/edusaintek.v12i3.1951.
- [8] S. Nova, N. Khotimah, and M. Y. Aryati Wahyuningrum, "PEMANFAATAN CHATBOT MENGGUNAKAN NATURAL LANGUAGE PROCESSING UNTUK PEMBELAJARAN DASAR-DASAR GUI TKINTER PADA BAHASA PEMROGRAMAN PYTHON," *JURNAL JUIT*, vol. 3, no. 1, pp. 58–65, 2024.
- [9] Y. Natadisastro, N. Safaat Harahap, and T. Darmizal, "Implementasi Large Language Models dalam Rancang Bangun Aplikasi Computer-Assisted Test Mobile Berbasis Android," vol. 6, no. 1, pp. 52–61, 2026, doi: 10.57152/malcom.v6i1.2402.

- [10] Rizal Chandra Rivaldi and T.D. Wismarini, "Analisis Sentimen Pada Ulasan Produk Dengan Metode Natural Language Processing (NLP)," *Elkom: Jurnal Elektronika dan Komputer*, vol. 17, no. 1, pp. 120–128, Jul. 2024, doi: 10.51903/elkom.v17i1.1680.
- [11] I. Fanani, "IMPLEMENTASI RETRIEVAL AUGMENTED GENERATION UNTUK EVALUASI PROPOSAL TUGAS AKHIR MAHASISWA," *Jurnal Teknologi Komputer dan Informatika* |, vol. 3, no. 2, 2025.
- [12] M. Yusuf, R. Setya Perdana, and P. P. Adikara, "Analisis Perbandingan Kinerja Metode Retrieval Cosine Similarity dan BM25 pada Chatbot berbasis Retrieval-Augmented-Generation," vol. 10, no. 2, pp. 2548–964, 2026