

ISSN (ONLINE) 2598-9936



INDONESIAN JOURNAL OF INNOVATION STUDIES
PUBLISHED BY
UNIVERSITAS MUHAMMADIYAH SIDOARJO

Table Of Contents

Journal Cover	1
Author[s] Statement	3
Editorial Team	4
Article information	5
Check this article update (crossmark)	5
Check this article impact	5
Cite this article.....	5
Title page	6
Article Title	6
Author information	6
Abstract	6
Article content	7

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Copyright © Author(s). This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licences/by/4.0/legalcode>

Indonesian Journal of Innovation Studies

Vol. 27 No. 1 (2026): January
DOI: 10.21070/ijins.v27i1.1836

EDITORIAL TEAM

Editor in Chief

Dr. Hindarto, Universitas Muhammadiyah Sidoarjo, Indonesia

Managing Editor

Mochammad Tanzil Multazam, Universitas Muhammadiyah Sidoarjo, Indonesia

Editors

Fika Megawati, Universitas Muhammadiyah Sidoarjo, Indonesia

Mahardika Darmawan Kusuma Wardana, Universitas Muhammadiyah Sidoarjo, Indonesia

Wiwit Wahyu Wijayanti, Universitas Muhammadiyah Sidoarjo, Indonesia

Farkhod Abdurakhmonov, Silk Road International Tourism University, Uzbekistan

Bobur Sobirov, Samarkand Institute of Economics and Service, Uzbekistan

Evi Rinata, Universitas Muhammadiyah Sidoarjo, Indonesia

M Faisal Amir, Universitas Muhammadiyah Sidoarjo, Indonesia

Dr. Hana Catur Wahyuni, Universitas Muhammadiyah Sidoarjo, Indonesia

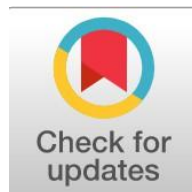
Complete list of editorial team ([link](#))

Complete list of indexing services for this journal ([link](#))

How to submit to this journal ([link](#))

Article information

Check this article update (crossmark)



Check this article impact (*)



Save this article to Mendeley



(*) Time for indexing process is various, depends on indexing database platform

Comparative Performance of Retrieval Augmented Generation Tourism Chatbots

Kinerja Komparatif Retrieval Augmented Generation pada Chatbot Pariwisata

Amar Al Farizi, izzyamaralfa@gmail.com, (0)

Program Studi Informatika, Universitas Amikom Purwokerto, Indonesia

Primandani Arsi, ukhti.prima@amikompurwokerto.ac.id, (1)

Program Studi Informatika, Universitas Amikom Purwokerto, Indonesia

Pungkas Subarkah, subarkah@amikompurwokerto.ac.id, (0)

Program Studi Informatika, Universitas Amikom Purwokerto, Indonesia

⁽¹⁾ Corresponding author

Abstract

General Background: The rapid adoption of artificial intelligence in smart tourism has increased the use of contextual chatbots to deliver destination information efficiently. **Specific Background:** However, tourism chatbots based on Large Language Models frequently encounter information hallucination, reducing reliability when handling dynamic and local tourism data. **Knowledge Gap:** Existing studies mainly focus on rule-based or single-model chatbot implementations and provide limited comparative evaluation of Retrieval Augmented Generation configurations combining embedding models and Large Language Models. **Aims:** This study aims to comparatively evaluate multiple Retrieval Augmented Generation configurations to identify the most suitable combination for contextual tourism chatbots and to analyze differences between large multilingual and small monolingual embedding models using a local tourism dataset. **Results:** Experimental evaluation using data from 49 tourist destinations in Banyumas Regency shows that the Multilingual-E5-Large embedding model consistently achieves perfect Precision, Recall, and F1-Score across all tested Large Language Models. The combination of Multilingual-E5-Large and GPT-4.1-Mini demonstrates the most balanced performance, achieving a BERTScore F1 of 0.7515 with an average response time of 1.555 seconds. **Novelty:** This research provides a systematic comparative assessment of embedding capacity and Large Language Model selection within a unified Retrieval Augmented Generation framework for tourism chatbots. **Implications:** The findings offer practical guidance for selecting model configurations that ensure accurate retrieval, high-quality responses, and efficient system performance in contextual tourism information services.

Highlights

- Multilingual embedding models deliver consistently higher retrieval accuracy across all tested configurations
- GPT-4.1-Mini produces the most balanced generative quality and response latency
- Embedding model selection plays a more decisive role than language model variation

Keywords

Retrieval Augmented Generation; Tourism Chatbot; Large Language Model; Embedding Model; Comparative Evaluation

Published date: 2026-01-07

I. Pendahuluan

Digitalisasi pariwisata semakin meluas seiring dengan perkembangan teknologi informasi yang menjadi bagian tak terpisahkan dari tren modern. Penerapan teknologi dalam sektor ini terbukti meningkatkan efisiensi pengelolaan, memperkuat daya saing, serta berkontribusi terhadap pertumbuhan ekonomi nasional [1]. Strategi seperti digital marketing dan experiential marketing turut berperan penting dalam membangun kepuasan serta loyalitas pengguna, menjadikannya elemen kunci dalam penguatan daya saing industri pariwisata [2].

Sektor smart tourism didorong oleh kebutuhan wisatawan modern yang semakin terhubung dan menuntut. Mereka tidak hanya mencari pengalaman, tetapi juga layanan yang sangat personal, responsif, dan nyaman [3]. Pemanfaatan teknologi modern menjadi keharusan agar sektor ini terus tumbuh secara kuat dan adaptif terhadap dinamika global [4]. Perkembangan chatbot dalam industri pariwisata sangat pesat dalam beberapa tahun terakhir, khususnya dengan adopsi teknologi AI generatif dan integrasi chatbot dalam berbagai layanan destinasi wisata, agen perjalanan, dan platform digital pariwisata [5]. Chatbot merupakan program perangkat lunak yang memanfaatkan teknologi natural language processing dan machine learning [6].

Perkembangan teknologi kecerdasan buatan mendorong lahirnya pendekatan Retrieval Augmented Generation, yaitu arsitektur yang menggabungkan proses pencarian informasi dan generasi teks dalam satu sistem. Penelitian ini mengadopsi pendekatan Retrieval Augmented Generation untuk mengatasi keterbatasan model bahasa konvensional yang bergantung sepenuhnya pada data pelatihan. Berbeda dengan model standar, RAG memungkinkan chatbot mengakses basis pengetahuan eksternal untuk menelusuri informasi terkini sebelum menghasilkan respons. Hasil pengujian menunjukkan bahwa integrasi ini secara signifikan meningkatkan relevansi dan akurasi informasi yang disajikan. Oleh karena itu, pendekatan Retrieval Augmented Generation dinilai lebih sesuai untuk pengelolaan informasi yang bersifat dinamis, khususnya pada konteks panduan wisata. [7].

Pendekatan ini juga menunjukkan efektivitas dalam mengurangi fenomena hallucination atau kesalahan informasi yang umum terjadi pada model generatif murni, karena Retrieval Augmented Generation mengacu langsung pada sumber data valid sebelum menghasilkan jawaban [8].

Integrasi Retrieval Augmented Generation dengan ChatGPT efektif dalam mengurangi potensi halusinasi model. Sistem menyediakan konteks yang relevan melalui pemanfaatan data FAQ serta penerapan instruksi pembatas, seperti “jawablah hanya berdasarkan informasi yang tersedia”. Selain itu, penggunaan konteks hasil pencarian dan riwayat percakapan membantu menjaga konsistensi respons chatbot agar tetap sesuai dengan sumber informasi. Strategi tersebut menjadikan Retrieval Augmented Generation dan ChatGPT lebih andal dalam menghasilkan jawaban yang kredibel dan kontekstual, terutama pada layanan chatbot interaktif yang membutuhkan akurasi tinggi dalam penyajian informasi [9].

Beberapa penelitian mulai mengarah pada pengembangan layanan berbasis chatbot untuk sektor pariwisata. Chatbot informasi wisata interaktif di Tangerang Selatan menggunakan framework Rasa, yang menunjukkan efektivitas membantu wisatawan memperoleh informasi destinasi secara cepat dan efisien. Meskipun Penelitian tersebut berhasil menghadirkan chatbot wisata berbasis Rasa, pendekatan ini masih terbatas pada pola percakapan tertentu [10]. Untuk menghasilkan jawaban yang lebih dinamis dan relevan, diperlukan pendekatan yang lebih canggih, dan dapat memberikan jawaban yang kontekstual dan selalu diperbarui [7]. Retrieval Augmented Generation sudah terbukti bermanfaat [11]. Sistem Retrieval Augmented Generation terdiri atas dua komponen utama, yaitu model embedding yang berfungsi mengonversi teks menjadi representasi vektor untuk proses pencarian konteks, dan Large Language Model yang berperan menghasilkan respons berdasarkan konteks yang ditemukan [12], [13].

Pemilihan komponen pada arsitektur Retrieval Augmented Generation merupakan aspek penting yang menentukan performa keseluruhan sistem. Arsitektur ini terdiri atas dua elemen utama, yakni model embedding dan Large Language Model, di mana masing-masing varian memiliki karakteristik dan keunggulan tersendiri. Model embedding berbeda dalam hal kapasitas dan orientasi bahasanya, beberapa dikembangkan untuk efisiensi ukuran kecil, penggunaan umum, maupun kinerja tinggi, serta ada pula yang dioptimalkan untuk penggunaan multibahasa atau bahasa tertentu seperti bahasa Inggris. Sementara itu, setiap Large Language Model menunjukkan variasi kemampuan dalam penalaran, retensi konteks, serta kelancaran respons [14], [15].

Penelitian lain mengungkap bahwa meskipun model bahasa besar dengan miliaran parameter mampu mencapai performa unggul pada sejumlah tugas dan subset bahasa, model publik dengan hasil terbaik justru adalah multilingual-e5-large-instruct yang hanya berjumlah 560 juta parameter. Untuk meningkatkan efisiensi komputasi dan aksesibilitas, dikembangkan metode downsampling berbasis korelasi antar-tugas yang menjaga keberagaman data tanpa mengubah peringkat performa antar-model. Dengan dilakukan optimasi pada tugas retrieval melalui penambahan hard negatives sehingga diperoleh dataset yang lebih ringkas namun tetap efektif. Pendekatan ini menghasilkan tolok ukur baru yang lebih hemat sumber daya komputasi [16].

Penelitian ini menggunakan dua model embedding yang berbeda, yaitu model multibahasa berukuran besar (Intfloat/Multilingual-E5-Large) dan model monolingual Indonesia berukuran kecil (LazarusNLP/All-Indo-E5-Small-V4), untuk membandingkan pengaruh kapasitas dan orientasi bahasa terhadap performa sistem. Penelitian ini juga menerapkan 3 Large Language Model seperti Claude Haiku 3.5, GPT-4o-mini, dan GPT-4.1-mini yang memiliki perbedaan arsitektur dan kemampuan dalam menghasilkan jawaban kontekstual. Perbandingan berbagai kombinasi model tersebut dilakukan untuk menilai bagaimana variasi embedding dan Large Language Model memengaruhi akurasi retrieval, relevansi jawaban

generatif, dan efisiensi waktu respons pada chatbot pariwisata berbasis Retrieval Augmented Generation. Pada penelitian sebelumnya hanya berfokus pada pengembangan chatbot wisata berbasis rule-based atau framework tertentu seperti Rasa, tanpa mengeksplorasi kombinasi antara Large Language Model dan Retrieval Augmented Generation. Penelitian ini dilakukan untuk mengevaluasi secara komparatif beberapa model bahasa besar yang diintegrasikan dengan dua jenis model embedding dalam kerangka Retrieval Augmented Generation. Proses evaluasi difokuskan pada analisis kinerja tiap konfigurasi model dalam menghasilkan respons chatbot panduan wisata yang sesuai dengan konteks pertanyaan dan kebutuhan pengguna. Hasil penelitian ini memberikan gambaran mengenai konfigurasi model yang paling sesuai untuk diterapkan pada pengembangan chatbot wisata di Indonesia.

II. Metode

Penelitian ini menerapkan pendekatan eksperimen komparatif untuk mengevaluasi kinerja sistem Retrieval Augmented Generation dalam pengembangan chatbot pariwisata kontekstual. Dengan menggunakan model embedding yang sebelumnya dijelaskan dan menggunakan tiga model bahasa besar yang bertujuan untuk mengidentifikasi perbedaan performa antar konfigurasi model. Proses evaluasi dilakukan menggunakan dataset yang memuat wisata di Banyumas dalam format JSON. Tahapan penelitian meliputi empat proses dapat dilihat pada Gambar 1.

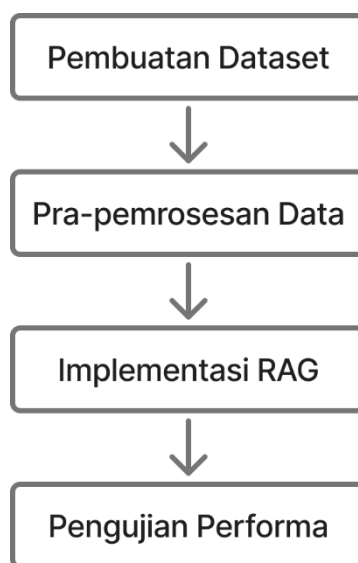


Figure 1. Tahapan Penelitian

1. Pembuatan Dataset

Dataset dalam penelitian ini disusun dari sejumlah sumber daring, termasuk situs resmi Pemerintah Kabupaten Banyumas dan laman informasi pariwisata terkait. Pengumpulan data difokuskan pada atribut utama destinasi, seperti lokasi, harga tiket, jam operasional, serta ragam aktivitas. Seluruh informasi tersebut selanjutnya dikonversi menjadi format JSON. Format terstruktur ini dipilih agar data dapat dibaca oleh sistem Retrieval Augmented Generation saat melakukan pencarian konteks jawaban.

2. Pra-pemrosesan Data

Pada tahap pra-pemrosesan, data JSON yang telah dibuat disesuaikan strukturnya agar dapat digunakan dalam sistem Retrieval Augmented Generation. Setiap data destinasi wisata diperlakukan sebagai satu dokumen dan direpresentasikan dalam bentuk objek Document pada framework LangChain. Informasi yang dimasukkan ke dalam setiap dokumen difokuskan pada atribut utama destinasi wisata, seperti lokasi, deskripsi singkat, harga tiket, serta jam operasional dan aktivitas yang tersedia. Tahapan ini dilakukan sebelum data dikonversi ke dalam bentuk vektor untuk kebutuhan proses embedding.

3. Implementasi Retrieval Augmented Generation

Sistem retrieval augmented generation dibangun dengan memanfaatkan framework LangChain sebagai orkestrator dan Pinecone sebagai basis data vektor. Kumpulan data yang telah melalui pra-pemrosesan ditransformasi ke dalam representasi vektor melalui dua model embedding yang berbeda untuk analisis komparatif Intfloat/Multilingual-E5-Large dan LazarusNLP/All-Indo-E5-Small-V4. Representasi vektor ini kemudian diindeksasi ke dalam basis data Pinecone, memfasilitasi retrieval konteks berlatensi rendah berdasarkan similaritas semantik. Pipeline retrieval augmented generation mengintegrasikan komponen retriever ini dengan tiga Large Language Models yang dievaluasi sebagai generator, antara lain Claude Haiku 3.5, GPT-4o-mini, dan GPT-4.1-mini. Hal ini dapat dilihat dari workflow pada gambar 2.

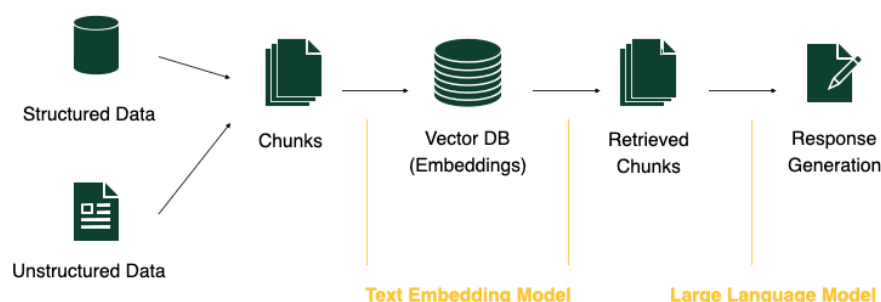


Figure 2. *Workflow Retrieval Augmented Generation* [17]

Gambar 2 menunjukkan bahwa proses dimulai ketika sistem menerima query atau pertanyaan dari pengguna terkait destinasi wisata. Selanjutnya, sistem melakukan tahap retrieval untuk menelusuri dan mengambil data pariwisata terstruktur dalam format JSON dari basis data vektor Pinecone.

Alur kerja RAG standar, atau yang dikenal sebagai query based retrieval augmented generation, terbagi menjadi dua tahap utama, yaitu indexing dan retrieval generation. Tahap indexing dilakukan secara offline, di mana data sumber baik terstruktur maupun tidak terstruktur dibagi menjadi potongan-potongan informasi yang logis. Setiap chunk kemudian diproses menggunakan text embedding model sebagai dense retriever untuk dikonversi menjadi representasi vektor yang padat. Seluruh vektor hasil embedding disimpan dalam vector database yang telah terindeks guna mendukung proses pencarian yang efisien. Tahap kedua, retrieval-generation, berjalan secara online ketika sistem menerima kueri pengguna. Pada tahap ini, retriever mencari dan mengambil konteks yang paling relevan dari vector database berdasarkan makna semantik kueri tersebut. Hasil retrieval kemudian digabungkan dengan pertanyaan asli pengguna untuk membentuk composite input, yang selanjutnya diteruskan ke Large Language Model. Model ini kemudian menghasilkan respons akhir berbasis konteks yang diperoleh [18].

4. Pengujian Performa

Tahap ini menilai efektivitas konfigurasi model dalam sistem Retrieval Augmented Generation. Pengujian dilakukan pada kombinasi dua model dengan tiga model bahasa. Evaluasi sistem dilakukan dengan mempertimbangkan tiga aspek utama. Aspek pertama adalah kinerja retrieval yang diukur menggunakan metrik precision, recall, dan F1-score. Aspek kedua mencakup kualitas generasi respons yang dievaluasi menggunakan BERTScore dan BLEU. Aspek terakhir berkaitan dengan efisiensi respons, yang dinilai berdasarkan rata-rata kecepatan sistem dalam menghasilkan jawaban. Hasil pengujian ini digunakan untuk menganalisis pengaruh variasi model terhadap akurasi pencarian, kualitas jawaban, dan efisiensi sistem chatbot pariwisata berbasis Retrieval Augmented Generation. Pada tahap ini juga dilakukan analisis untuk menentukan model yang paling efektif dan sesuai dalam sistem. Setiap hasil pengujian dibandingkan secara komparatif guna mengidentifikasi konfigurasi model dengan performa paling optimal dalam menjawab pertanyaan terkait destinasi wisata. Selain itu, dilakukan analisis deskriptif terhadap pola kesalahan, tingkat relevansi konteks, dan konsistensi jawaban antar model untuk memperoleh gambaran komprehensif mengenai keandalan sistem yang dikembangkan.

III. Hasil dan Pembahasan

A. Dataset

Dataset dalam penelitian ini mencakup 49 destinasi wisata di Kabupaten Banyumas yang disusun dalam format JSON terstruktur. Setiap entri berisi informasi utama meliputi id, nama objek wisata, lokasi, deskripsi, harga tiket (dewasa dan anak-anak), jam operasional, serta aktivitas utama yang dihimpun dari sumber resmi pemerintah daerah dan portal pariwisata terpercaya dengan total kata sebesar 4765 kata. Dataset pada penelitian ini berperan sebagai sumber informasi eksternal yang digunakan oleh sistem Retrieval Augmented Generation. Melalui dataset tersebut, komponen retriever melakukan pencarian dan pemilihan informasi yang relevan sebagai dasar pembentukan respons chatbot. Pada tahap pengujian, disusun sebanyak 215 pertanyaan yang merepresentasikan berbagai kebutuhan informasi wisata. Pertanyaan-pertanyaan ini digunakan untuk menilai performa sistem pada setiap model, khususnya terkait ketepatan hasil retrieval, kualitas teks yang dihasilkan, serta waktu respons.

```
{
  "id": "baturraden",
  "nama": "Lokawisata Baturraden",
  "lokasi": "Karangmangu, Baturraden",
  "deskripsi": "Lokawisata Baturraden merupakan destinasi utama di kaki Gunung Slamet yang menawarkan taman rekreasi keluarga,
  "harga_tiket_dewasa": "Rp20.000 ",
  "harga_tiket_anak": "Rp10.000",
  "harga_tiket_weekend": "Rp25.000 (dewasa), Rp12.500 (anak)",
  "jam_buka": "7 AM",
  "jam_tutup": "7 PM",
  "aktivitas_utama": "Rekreasi keluarga, Pemandian air panas, Bermain air, Taman bermain"
},
```

Figure 3. *Sampel Dataset*

B. Hasil Preprocessing Data

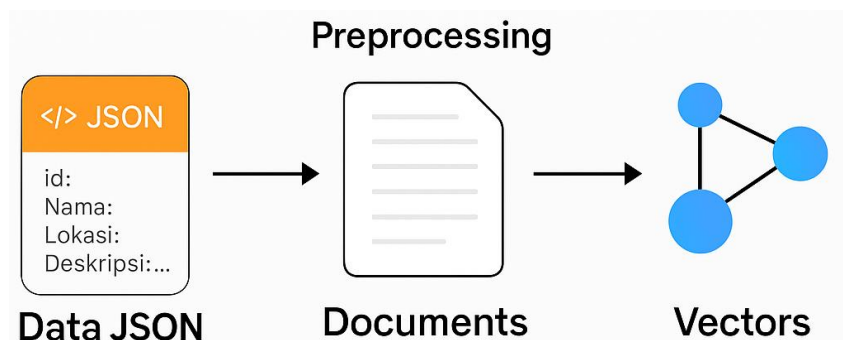


Figure 4. *Diagram Alur Preprocessing Data JSON menjadi Representasi Vektor*

Tahap pra-pemrosesan data dilakukan untuk mengubah kumpulan data wisata dalam format JSON menjadi representasi dokumen yang dapat diproses oleh framework LangChain dan diindeks ke dalam basis data vektor Pinecone. Setiap entri destinasi wisata yang terdiri atas atribut seperti id, nama, lokasi, deskripsi, aktivitas utama, harga tiket, serta jam operasional, diubah menjadi objek Document yang berisi `page_content` dan metadata. Proses ini memastikan bahwa seluruh data memiliki format yang seragam serta dapat dikonversi menjadi representasi vektor pada tahap embedding. Hasil pra-pemrosesan menghasilkan 49 dokumen siap proses dengan struktur teks yang telah distandardisasi dan siap diintegrasikan ke pipeline Retrieval Augmented Generation untuk tahap embedding dan retrieval.

C. Hasil Implementasi Retrieval Augmented Generation

Pada proses ini mencakup proses embedding, penyimpanan vektor, serta integrasi dengan model bahasa besar. Pada tahap embedding, seluruh dokumen hasil pra-pemrosesan diubah menjadi representasi vektor berdimensi tinggi menggunakan dua model embedding, yaitu Intfloat/Multilingual-E5-Large dan LazarusNLP/All-Indo-E5-Small-V4. Kedua model tersebut dipilih untuk membandingkan pengaruh kapasitas dan orientasi bahasa terhadap akurasi pencarian konteks. Hasil embedding selanjutnya diindeks ke dalam basis data vektor Pinecone, yang berfungsi sebagai vector store untuk mendukung pencarian semantik (semantic similarity search) dengan latensi rendah. Proses ini memungkinkan sistem untuk menemukan konteks yang paling relevan dari data wisata Banyumas berdasarkan pertanyaan pengguna sebelum tahap generatif dilakukan oleh LLM. Adapun untuk model embedding Intfloat/Multilingual-E5-Large menggunakan dimension 1024 sedangkan untuk model embedding LazarusNLP/All-Indo-E5-Small-V4 menggunakan dimension 384 bisa dilihat pada gambar 5.

chatbot-wisata-e5large					
METRIC	DIMENSIONS	HOST			
cosine	1024	https://chatbot-wisata-e5large-9nmvwp.svc.aped-4627-b74a.pinecone.io			
CLOUD	REGION	TYPE	CAPACITY MODE		RECORD COUNT
AWS AWS	us-east-1	Dense	Serverless		49

all-indo-e5-small-v4					
METRIC	DIMENSIONS	HOST			
cosine	384	https://all-indo-e5-small-v4-9nmvwp.svc.aped-4627-b74a.pinecone.io			
CLOUD	REGION	TYPE	CAPACITY MODE		RECORD COUNT
AWS AWS	us-east-1	Dense	Serverless		49

Figure 5. *Tampilan Antarmuka Vector Database*

Model embedding Intfloat/Multilingual-E5-Large memiliki dimension 1024, sedangkan LazarusNLP/All-Indo-E5-Small-V4 menggunakan dimension 384. Perbedaan dimensi ini merepresentasikan variasi kapasitas pemahaman semantik antar-model. Dimensi tinggi memungkinkan penangkapan konteks yang lebih kompleks namun memerlukan sumber daya komputasi lebih besar, sedangkan dimensi rendah lebih efisien meski berpotensi menurunkan akurasi pada konteks rumit. Penggunaan kedua model ini bertujuan menilai pengaruh ukuran dimensi dan orientasi bahasa terhadap akurasi retrieval, kualitas jawaban generatif, serta efisiensi waktu respons dalam sistem retrieval augmented generation chatbot pariwisata kontekstual [19].

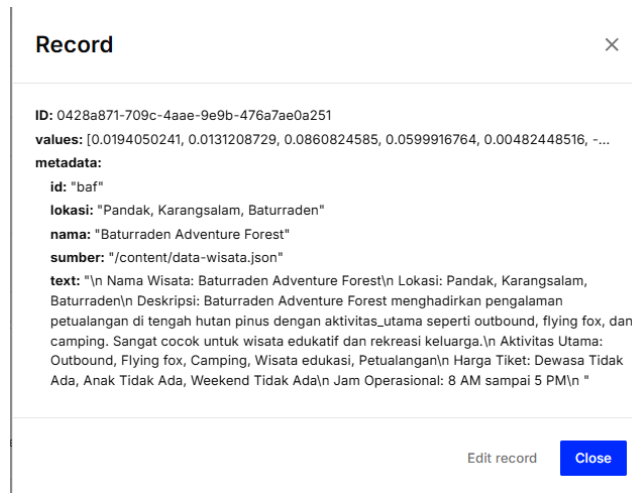


Figure 6. Sample data pada Pinecone Vector Database

Setelah proses embedding dan penyimpanan vektor, sistem diintegrasikan dengan tiga Large Language Models yang berfungsi sebagai generator dalam pipeline Retrieval Augmented Generation. Integrasi dilakukan menggunakan framework LangChain sebagai orchestrator antara retriever, model embedding, dan Large Language Models. Proses pencarian konteks dilakukan melalui perhitungan cosine similarity pada basis data vektor Pinecone, di mana sistem mengambil satu dokumen paling relevan untuk setiap pertanyaan pengguna. Hasil retrieval tersebut kemudian digunakan oleh Large Language Models untuk menghasilkan jawaban berbasis konteks. Tampilan antarmuka vector database pada Gambar 6 memperlihatkan hasil indexing dari salah satu data wisata Banyumas yang telah dikonversi ke dalam bentuk vektor.

Id	Input	Output
pancuran_pitu	aktivitas apa yang bisa dilakukan di pancuran pitu?	"Aktivitas utama di Pancuran Pitu adalah pemandian air panas dan relaksasi. Pengunjung bisa merendam tubuh sambil menikmati udara pegunungan yang sejuk."
pagubugan_melung	pagubugan melung itu apa? dan berapa harga tiket masuknya ?	"Pagubugan Melung adalah sebuah desa wisata yang berlokasi di Desa Melung, Kecamatan Kedungbanteng, Kabupaten Banyumas. Terletak sekitar 13 km dari pusat Kabupaten Banyumas, lokasi ini dapat ditempuh dalam waktu 30 menit menggunakan kendaraan bermotor. Wisatawan dapat mengunjungi tempat ini untuk rekreasi umum dengan tiket masuk dewasa seharga Rp10.000."
lembu_benggolo_farm	harga tiket Lembu Benggolo farm Resort berapa? dan jam bukanya?	"Harga tiket Lembu Benggolo farm dan Resort adalah Rp 5.000 untuk dewasa dan anak, termasuk pada akhir pekan. Jam operasionalnya adalah dari 8 AM sampai 10 PM."
Data Referensi		
Pancuran Pitu dikenal sebagai pemandian air panas dengan tujuh pancuran alami. Pengunjung bisa merendam tubuh sambil menikmati udara pegunungan yang sejuk. Pagubugan Melung berada di Desa Wisata Melung, Kecamatan Kedungbanteng, Kabupaten Banyumas. lokasinya kurang lebih berjarak 13 km dari pusat Kabupaten Banyumas. Anda dapat menempuh perjalanan selama 30 menit dengan kendaraan bermotor untuk mencapai lokasi wisata Pagubugan Melung. Tiket Dewasa Rp 10.000. Lembu Benggolo merupakan di tempat wisata yang memiliki daya tarik mini zoo, yaitu terdapat banyak hewan- hewan seperti sapi, monyet, burung, kelinci, kelelawar, dan masih banyak lagi, yang dapat digunakan untuk belajar anak-anak. Selain mini zoo, ada juga resto yang mengusung konsep		

saung-saung kecil dan joglo yang dapat digunakan untuk tempat istirahat dan pesan makanan. Untuk bisa masuk ke area wisata, setiap pengunjung hanya perlu membayar tiket masuk sebesar Rp 5.000 saja. Namun jika ingin menikmati wahana seperti rainbow slide ataupun ATV tentunya dikenakan biaya tersendiri.

Table 1. Hasil Uji Respon Chatbot terhadap Pertanyaan Wisata

Setelah sistem mampu menghasilkan respons yang kontekstual sebagaimana ditunjukkan pada Tabel 1, langkah selanjutnya merupakan pengujian performa guna menilai efektivitas konfigurasi antara model embedding dan model bahasa besar dalam arsitektur Retrieval Augmented Generation. Tujuan pengujian ini adalah untuk mengukur sejauh mana kombinasi model tersebut dapat menghasilkan jawaban yang akurat, relevan, serta efisien dalam merespons pertanyaan terkait destinasi wisata.

D. Pengujian Performa

Pengujian ini dilakukan untuk membandingkan hasil teks generasi dari tiga model bahasa besar dan dua varian model embedding, nantinya setiap pertanyaan akan diujikan pada setiap model agar dapat melihat perbedaan hasil respon dan ketepatan hasil kalimat dengan referensi datanya, waktu dan juga kualitas hasil teks generasinya.

Embedding Model	LLM	Precision	Recall	F1-Score	BERTScore F1	BLEU-2	Avg Latency (s)
Multilingual-E5-Large	GPT-4o-Mini	1.0000	1.0000	1.0000	0.7512	0.1581	1.583
Multilingual-E5-Large	GPT-4.1-Mini	1.0000	1.0000	1.0000	0.7515	0.1594	1.555
Multilingual-E5-Large	Claude Haiku 3.5	1.0000	1.0000	1.0000	0.7428	0.0929	2.173
All-Indo-E5-Small-V4	GPT-4o-Mini	0.9628	0.9628	0.9628	0.7492	0.1570	1.438
All-Indo-E5-Small-V4	GPT-4.1-Mini	0.9628	0.9628	0.9628	0.7477	0.1466	1.279
All-Indo-E5-Small-V4	Claude Haiku 3.5	0.9628	0.9628	0.9628	0.7405	0.0926	1.928

Table 2. Perbandingan Kinerja Model Embedding dan LLM pada Sistem Retrieval Augmented Generation

Berdasarkan Tabel 2, hasil pengujian menunjukkan perbandingan kinerja berbagai kombinasi model embedding dan LLM pada sistem Retrieval Augmented Generation. Secara keseluruhan, dapat diamati bahwa kombinasi model embedding Multilingual-E5-Large dengan berbagai LLM menunjukkan performa yang superior dibandingkan dengan All-Indo-E5-Small-V4, terutama pada metrik Precision, Recall, dan F1-Score yang mencapai nilai akurasi maksimal sebesar 1.0000.

Model Multilingual-E5-Large dengan Large Language Models GPT-4o-Mini, GPT-4.1-Mini, dan Claude Haiku 3.5 menunjukkan konsistensi yang sangat baik dengan nilai Precision, Recall, dan F1-Score dengan skor ketiganya sebesar 1.0000, mengindikasikan bahwa sistem mampu melakukan retrieval dan generasi jawaban dengan akurasi sempurna. Namun, terdapat perbedaan signifikan pada metrik evaluasi kualitas teks yang dihasilkan. GPT-4.1-Mini menunjukkan performa terbaik dengan BERTScore F1 sebesar 0.7515 dan BLEU-2 sebesar 0.1594, sedikit mengungguli GPT-4o-Mini BERTScore F1 dengan nilai 0.7512, sedangkan untuk BLEU-2 hasilnya 0.1581.

Pada hasil pengujian Claude Haiku 3.5 memperoleh skor BLEU-2 terendah sebesar 0.0929, sementara nilai BERTScore F1 yang dihasilkan berada pada angka 0.7428. Skor BLEU-2 yang relatif rendah menunjukkan bahwa kesamaan leksikal antara teks hasil generasi dan referensi tidak terlalu tinggi. Sementara itu, dari aspek waktu respons, GPT-4.1-Mini menunjukkan rata-rata latensi tercepat sebesar 1.555 detik, sedikit lebih cepat dibandingkan GPT-4o-Mini yang mencatat waktu respons sebesar 1.583 detik pada skenario pengujian yang sama. Sebaliknya, Claude Haiku 3.5 menunjukkan latensi paling tinggi dengan rata-rata waktu respons 2.173 detik.

Model All-Indo-E5-Small-V4 meskipun menggunakan model embedding yang lebih kecil dan spesifik untuk bahasa Indonesia, performa retrieval tetap menunjukkan hasil yang kompetitif dengan Precision, Recall, dan F1-Score sebesar 0.9628 untuk ketiga Large Language Models yang diuji. Pada model ini mengalami penurunan sekitar 0.0372 poin dibandingkan Multilingual-E5-Large menunjukkan trade off antara ukuran model dan akurasi retrieval. Pada aspek kualitas generasi, GPT-4o-Mini memperoleh nilai BERTScore F1 tertinggi sebesar 0.7492, disusul oleh GPT-4.1-Mini dengan nilai 0.7477 dan Claude Haiku 3.5 sebesar 0.7405. Pola yang sama juga terlihat pada metrik BLEU-2, di mana GPT-4o-Mini mencatat skor 0.1570 dan GPT-4.1-Mini sebesar 0.1466, sementara Claude Haiku 3.5 menghasilkan skor yang lebih rendah,

yaitu 0.0926.

Dari hasil pengamatan waktu respons, GPT-4.1-Mini menunjukkan latensi rata-rata paling rendah, yakni 1.279 detik, diikuti oleh GPT-4o-Mini sebesar 1.438 detik dan Claude Haiku 3.5 sebesar 1.928 detik. Secara komparatif, pemilihan model embedding memberikan dampak lebih signifikan terhadap akurasi retrieval daripada pemilihan LLM, sebagaimana ditunjukkan oleh nilai Precision, Recall, dan F1-Score yang konsisten dalam kelompok embedding yang sama.

Berdasarkan hasil pengujian, model Multilingual-E5-Large menunjukkan performa retrieval yang stabil pada seluruh konfigurasi LLM yang diuji. Temuan ini mengindikasikan bahwa kapasitas representasi semantik pada model embedding memiliki peran penting dalam menentukan ketepatan pencarian konteks.

Sementara itu, variasi model bahasa besar lebih memengaruhi karakteristik teks yang dihasilkan serta efisiensi komputasi. Pada pengujian ini, GPT-4.1-Mini mencatat nilai BERTScore dan BLEU-2 yang lebih tinggi disertai dengan waktu respons yang lebih rendah dibandingkan model lain dalam kelompok pengujian.

Pada model Claude Haiku 3.5, skor BLEU-2 yang diperoleh tergolong rendah. Namun, rendahnya skor tersebut tidak berarti bahwa jawaban yang dihasilkan tidak tepat, sebaliknya, jawaban yang diberikan sudah sesuai dengan referensi, meskipun urutan kalimatnya tidak sepenuhnya sama, hasilnya terdengar lebih alami dan tidak terkesan kaku. Hal ini terlihat dari BERTScore yang lebih tinggi, yang menunjukkan bahwa jawaban tersebut sangat mendekati referensi meskipun strukturnya berbeda. Dari sisi praktis GPT-4o-Mini juga dapat dipertimbangkan sebagai alternatif karena efisiensi biaya operasionalnya, meskipun aspek ini tidak dievaluasi secara kuantitatif dalam pengujian.

IV. Kesimpulan

Hasil penelitian menunjukkan bahwa pemilihan model embedding memberikan pengaruh yang paling signifikan terhadap tingkat akurasi proses retrieval dalam sistem RAG. Model Intfloat/Multilingual-E5-Large menunjukkan kinerja paling unggul dengan pencapaian nilai Precision, Recall, dan F1-Score sempurna 1.0000, melampaui model All-Indo-E5-Small-V4 yang memperoleh nilai 0.9628 pada ketiga metrik tersebut. Sementara itu, pemilihan Large Language Model berperan dominan dalam menentukan kualitas keluaran teks yang diukur melalui metrik BERTScore dan BLEU-2, serta efisiensi waktu respons sistem. Konfigurasi paling optimal diperoleh pada kombinasi Multilingual-E5-Large sebagai embedding dan GPT-4.1-Mini sebagai LLM, yang menghasilkan keseimbangan ideal antara akurasi sempurna, kualitas generatif tertinggi (BERTScore F1 = 0.7515 dan BLEU-2 = 0.1594), serta latensi terendah (1.555 detik). Dengan demikian, integrasi kedua model tersebut dapat direkomendasikan sebagai konfigurasi paling efektif untuk implementasi arsitektur retrieval augmented generation pada pengembangan chatbot pariwisata. Penelitian selanjutnya dapat menambah multiple question mengingat pada penelitian ini masih menggunakan single question dan melakukan pengujian ulang apakah dengan multi question akan menghasilkan hasil yang lebih optimal dan tinggi.

References

1. S. Mammadova and L. Isayeva, "DIGITAL MARKETING OF TOURISM AS A TOOL FOR DESTINATION ECONOMIC DEVELOPMENT," PAHTEI-Proceedings of Azerbaijan High Technical Educational Institutions, vol. 49, no. 02, pp. 623–634, Feb. 2025, doi: <https://doi.org/10.5281/zenodo.14930999>.
2. April Vela, WA Armalia Reny, Alie Maria Septijantini, CN Yudhinanto, Hasbullah, and Oktaria Eka Travilta, "Pengaruh Digital Marketing Dan Experiential Marketing Terhadap Kepuasan Dan Loyalitas Pengguna Jasa Travel Pariwisata Rakata Tour Indonesia," COSTING, vol. 7, no. 5, pp. 5060–5071, Oct. 2024, doi: <https://doi.org/10.31539/costing.v7i6.12572>.
3. S. Febbi Handayani and D. Intan Af, "Assessment of Retrieval and Generative Chatbots in Tourism Information Service," 2025.
4. H. Sunarto, "Strategi Branding Pengembangan Industri Pariwisata 4.0 melalui Kompetitif Multimedia di Era Digital," Journal of Tourism and Creativity, vol. 4, no. 1, 2020.
5. G. M. Majid, I. Tussyadiah, and Y. R. Kim, "Exploring the Potential of Chatbots in Extending Tourists' Sustainable Travel Practices," J Travel Res, vol. 64, no. 6, pp. 1292–1317, Jul. 2025, doi: 10.1177/00472875241247316.
6. A Rahman, Mamun Abdullah Al, and Islam Alma, "Programming challenges of Chatbot: Current and Future Prospective," IEEE, 2017, doi: <https://doi.org/10.1109/R10-HTC.2017.8288910>.
7. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," p. 1, Apr. 2021, doi: <https://doi.org/10.48550/arXiv.2005.11401>.
8. J. L  la, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodrigues, and A. D. White, "PaperQA: Retrieval-Augmented Generative Agent for Scientific Research," Dec. 2023, [Online]. Available: <http://arxiv.org/abs/2312.07559>
9. M. Kulkarni, P. Tangarajan, K. Kim, and A. Trivedi, "Reinforcement Learning for Optimizing RAG for Domain Chatbots," Jan. 2024, [Online]. Available: <http://arxiv.org/abs/2401.06800>
10. A. D. Ferdian and S. N. Anwar, "Pengembangan Chatbot untuk Informasi Wisata Interaktif di Tangerang Selatan menggunakan Framework Rasa," Jurnal Teknologi Dan Sistem Informasi Bisnis, vol. 5, no. 4, pp. 476–483, Oct. 2023, doi: 10.47233/jteksis.v5i4.953.
11. Raya Rizky Dharma Andika and Cahyanti F Lia Dwi, "Perancangan Sistem Informasi Chatbot Retrieval Augmented Generation Berbasis Website Pada PT. Revolusi Cita Edukasi," Journal Computer Science, vol. 4, no. 1, pp. 15–21, 2025, doi: <https://doi.org/10.31294/m75d4782>.
12. Y. H. Ke et al., "Development and Testing of Retrieval Augmented Generation," 2024.
13. T. Wang, J. He, and C. Xiong, "RAGVIZ: Diagnose and Visualize Retrieval-Augmented Generation," Association for Computational Linguistics, 2024, pp. 320–327. [Online]. Available: <https://youtu.be/cTAbuTu6ur4>.
14. Z. Li, C. Li, M. Zhang, Q. Mei, and M. Bendersky, "Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach," Oct. 2024, [Online]. Available: <http://arxiv.org/abs/2407.16833>
15. H. Lin, S. Zhan, J. Su, H. Zheng, and H. Wang, "IRSC: A Zero-shot Evaluation Benchmark for Information Retrieval through Semantic Comprehension in Retrieval-Augmented Generation Scenarios," Sep. 2024, [Online]. Available: <http://arxiv.org/abs/2409.15763>

16. K. Enevoldsen et al., "MMTEB: Massive Multilingual Text Embedding Benchmark," Jun. 2025, [Online]. Available: <http://arxiv.org/abs/2502.13595>
17. Chaoyu Yang, "Building RAG Systems with Open-Source and Custom AI Models," <https://www.bentoml.com/blog/building-rag-with-open-source-and-custom-ai-models>.
18. P. Zhao et al., "Retrieval-Augmented Generation for AI-Generated Content: A Survey," Jun. 2024, [Online]. Available: <http://arxiv.org/abs/2402.19473>
19. A. Rao, H. Alipour, and N. Pendar, "Rethinking Hybrid Retrieval: When Small Embeddings and LLM Re-ranking Beat Bigger Models," May 2025, [Online]. Available: <http://arxiv.org/abs/2506.00049>