Analisis Kinerja Retrieval Augmented Generation (RAG) Klasik Dalam Chatbot Akademik berbasis Multimodal

Kartika Wulandari[1], Yuni Yamasari[2]

[1,2] Teknik Informatika, Fakultas Teknik, Universitas Negeri Surabaya
[1]kartika.22145@mhs.unesa.ac.id
[2]yuniyamasari@unesa.ac.id

***Abstrak*— Akses informasi akademik di Universitas Negeri Surabaya (UNESA) melalui website panduan saat ini dinilai belum optimal karena kurangnya interaktivitas, sehingga mahasiswa seringkali harus menghubungi pihak administrasi secara langsung. Penelitian ini bertujuan untuk melakukan evaluasi kinerja Retrieval-Augmented Generation (RAG), yaitu RAG Klasik guna menguji kinerja sistem dalam menghasilkan respon yang relevan, akurat, dan kontekstual pada chatbot akademik. Lingkup penelitian ini difokuskan pada pengujian evaluasi kinerja dan efisiensi metode tersebut dalam mengolah data multimodal yang bersumber dari portal SSO UNESA, mencakup format teks, gambar, dan dokumen PDF.**

Metodologi yang digunakan meliputi tahap pengumpulan data, ekstraksi teks menggunakan Optical Character Recognition (OCR), preprocessing, serta penerapan dua strategi pelabelan, yakni manual dan otomatis. Penelitian ini menganalisis kinerja RAG Klasik yang menggunakan algoritma BM25 dalam mengolah data multimodal.

RAG Klasik diimplementasikan menggunakan algoritma BM25 berbasis kata kunci. Evaluasi dilakukan secara sistematis menggunakan metrik efektivitas (Precision, Recall, F1-Score, Exact Match, dan Cosine Similarity) serta metrik efisiensi (waktu respons).

Hasil penelitian menunjukkan bahwa RAG Klasik memiliki efisiensi luar biasa dengan waktu respon rata-rata dibawah 0,01 detik. Namun, dari sisi kinerja, metode ini sangat bergantung pada kualitas pelabelan, di mana skenario label otomatis mencapai F1-Score 0.52 dan Csoine Similarity 0.51, mengungguli scenario label manual. Penelitian menyimpulkan bahwa RAG Klasik sangat ideal untuk informasi procedural yang membutuhkan respon instan meskipun memiliki keterbatsan dalam pemahaman semantic yang mendalam.

***Kata Kunci*— *Retrieval Augmented Generation* (RAG), RAG Klasik, BM25, Multimodal, UNESA.**

I. PENDAHULUAN

Perkembangan teknologi informasi yang pesat dalam era digital telah membawa perubahan besar di berbagai bidang, termasuk dalam sektor pendidikan tinggi [24]. Perguruan tinggi kini semakin bergantung pada sistem digital untuk mendukung kegiatan akademik dan administratif, seperti portal akademik, sistem e-learning, dan website panduan informasi kampus. Sistem-sistem ini diharapkan dapat membantu sivitas akademika memperoleh informasi dengan cepat dan mandiri tanpa harus selalu menghubungi pihak administrasi.

Namun, kenyataannya di Universitas Negeri Surabaya (UNESA), Pemanfaatan website panduan sebagai sumber informasi utama masih belum berjalan secara optimal. Banyak mahasiswa lebih memilih untuk bertanya langsung kepada staf administrasi dibandingkan mencari informasi melalui website. Kondisi ini menunjukkan bahwa sistem informasi akademik yang tersedia masih memiliki keterbatasan, baik dari sisi kemudahan penggunaan maupun kemampuan sistem dalam menjawab kebutuhan pengguna secara kontekstual.

Sebagai alternatif Solusi, chatbot berbasis kecerdasan buatan menjadi teknologi yang semakin relevan untuk meningkatkan kualitas layanan informasi akademik. Chatbot memungkinkan interaksi menggunakan bahasa alami sehingga pengguna dapat memperoleh informasi secara cepat, peronal, dan interaktif [12]. Penerapan chatbot juga terbukti mampu mengurangi beban kerja staf administratif serta meningkatkan efisiensi layanan informasi di berbagai institusi pendidikan.

Meskipun demikian, sebagian besar terdapat chatbot konvensional masih bergantung pada pendekatan pencocokan pola atau *string matching* yang bersifat kaku. Pendekatan tersebut sering kali gagal memahami konteks, maupun variasi bahasa alami dalam pertanyaan pengguna. Selain itu, banyak sistem chatbot yang dirancang dengan data unimodal, yaitu hanya memanfaatkan data teks, sehingga kurang mampu menangani informasi yang berasal dari dokumen dan gambar. Keterbatasan ini menyebabkan jawaban yang dihasilkan menjadi tidak lengkap atau kurang relevan.

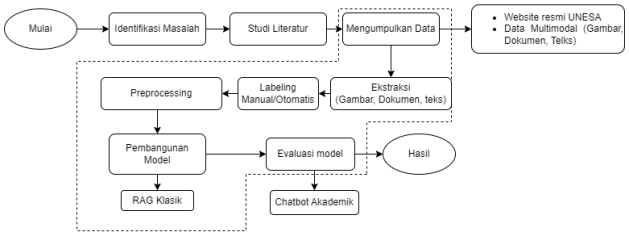
Untuk mengatasi permasalahan tersebut, data multimodal menjadi penting karena memungkinkan sistem chatbot memanfaatkan berbagai jenis data. Data multimodal digunakan sebagai basis pengetahuan karena informasi resmi umumnya tersedia dalam berbagai format seperti teks, dokumen, maupaun gambar.

Salah satu pendekatan yang banyak digunakan dalam pengembangan chatbot adalah Retrieval Augmented generation (RAG). RAG mengombinasikan proses pencarian informasi (retrieval) dari kumpulan dokumen dengan proses pembangkitan jawaban (generation). Dalam implementasinya, RAG klasik, retrieval umimnya dilakukan menggunakan metode berbasis kata kunci seperti TF-IDF dan cosine similarity, yang kemudian digunakan dasar dalam menghasilkan jawaban. Pendekatan ini relative sederhana dan efisien, namun masih memiliki keterbatasan dalam memahami konteks semantic serta variasi bahsa pengguna

Beberapa penelitian sebelumnya telah menunjukkan keunggulan RAG dalam meningkatkan kualitas chatbot. Hidayat dkk. (2025) membuktikan efektivitas RAG berbasis LLM pada integrasi Telegram, Danuarta dkk. (2024) meningkatkan presisi dan relevansi chatbot edukasi dengan RAG, dan Pratama dkk. (2024) mengembangkan chatbot administratif berbasis RAG yang lebih responsif di lingkungan nyata. Namun, sebagian besar penelitian tersebut masih terbatas pada data *unimodal* serta belum secara khusus mengevaluasi kinerja RAG klasik pada chatbot informasi akademik. Selain itu, kajian mendalam terkait peforma RAG klasik dalam menjawab berbagai jenis pertanyaan akademik masih relative terbatas.

Berdasarkan kondisi tersebut, penelitian ini berfokus pada evaluasi kinerja Retrieval Augmented Generation (RAG) klasik dalam chatbot informasi akademik di lingkungan Universitas Negeri Surabaya (UNESA).

II. METODE PENELITIAN



Gbr 1. Flowchart Alur Penelitian

.

A. *Identifikasi Masalah*

Berdasarkan hasil eksplorasi yang telah dilakukan, dirumuskan pertanyaan penelitian yang relevan dengan kebutuhan pengembangan sistem informasi akademik di era digital. Perumusan masalah ini berangkat dari pemahaman mendalam terhadap keterbatasan chatbot konvensional yang masih berbasis pencocokan kata kunci (keyword-based), yang cenderung kurang mampu memahami konteks serta variasi bahasa alami dalam pertanyaan pengguna. Kondisi tersebut mendorong perlunya pendekatan yang lebih adaptif dalam penyajian informasi akademik.

Salah satu pendekatan yang banyak diterapkan untuk meningkatkan kualitas respons chatbot adalah Retrieval Augmented Generation (RAG). Pada pendekatan RAG klasik, proses pencarian informasi dilakukan menggunakan metode *retrieval* berbasis kata kunci, seperti BM25, yang kemudian dimanfaatkan sebagai dasar dalam menghasilkan jawaban. Meskipun pendekatan ini relative sederhana dan efisien, kinerjanya dalam konteks chatbot informasi akademik, khususnya dalam menghasilkan jawaban yang relevan dan akurat, masih perlu di evaluasi secara menyeluruh.

Oleh karena itu, penelitian ini difokuskan untuk menjawab pertanyaan bagaimana kinerja pendekatan Retrieval Augmented Generation (RAG) klasik berbasis BM25 dalam menghasilkan respons yang relevan, akurat, dan kontekstual pada chatbot informasi akademik penerapan berbasis multimodal.

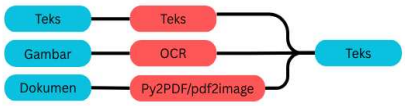
B. *Studi Literatur*

Pada tahap ini, studi literatur dilakukan sebagai pendukung dan acuan penelitian dengan menelaah berbagai sumber yang membahas pengembangan chatbot, information Retrieval (IR), Natural Languange Processing (NLP), serta Retrieval Augmented Generation (RAG) klasik berbasis metode retrieval kata kunci, khususnya BM25. Literatur dikumpulkan dari buku, artikel ilmiah, dan prosiding konferensi untuk mempelajari konsep dasar, arsitektur, dan mekanisme kerja RAG kalsik dalam chatbot informasi akademik, sekaligus mengindentifikasi keterbatasan penelitian sebelumnya terkait evaluasi kinerja RAG klasik.

C. *Pengumpulan Data*

Data dalam penelitian ini diperoleh dari website akademik resmi Universitas Negeri Surabaya (UNESA) melalui portal Single Sign-On (SSO UNESA). Sumber data mencakup berbagai sistem dan layanan akademik yang umum digunakan mahasiswa untuk mengakses informasi perkuliahan, administrasi, dan panduan akademik. Data tersebut digunakan sebagai basis pengetahuan (knowledge base) bagi system chatbot akademik yang dikembangkan.Secara keseluruhan, data yang digunakan terdiri dari tiga jenis format (multimodal), yaitu teks (.txt), gambar (.png), dan dokumen (.pdf).

D. *Ekstraksi Data*



Gbr 2. Ekstraksi Data

Data multimodal yang telah dikumpulkan, selanjutnya dilakukan ekstraksi data multimodal yaitu mengubah jenis format data teks, gambar, dan dokumen menjadi satu format teks digital seragam. Proses ini bertujuan untuk menghasilkan dataset teks yang nantinya akan digunakan sebagai basis pengetahuan (knowledge base) pada tahap penelitian selanjutnya. Secara teknis, ekstraksi dilakukan langsung pada file teks, menggunakan OCR (Optical Character Recognition) untuk file gambar, serta pustaka PyPDF2 atau konversi gambar untuk dokumen PDF.

E. Pelabelan

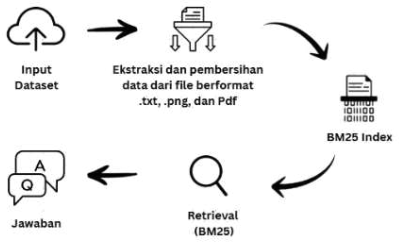
Data yang sudah berhasil di ekstraksi akan dilakukan pelabelan untuk menghasilkan pasangan data yang akan digunakan sebagai basis pengetahuan (knowledge base) bagi sistem chatbot akademik. Proses ini menerapkan dua strategi utama yaitu label manual yang menghasilkan pasangan tanya-jawab (QA) dan label otomatis menggunakan teknik lexicon

matching. Tahap ini bertujuan untuk memperkaya dataset serta menyediakan data acuan (gorund truth) untuk proses evalausi model.

F. *Preprocessing*

Teks dibersihkan, ditokenisasi (NLTK), dan dipecah menjadi *chunk* sebesar 150 kata dengan *overlap* 30 kata untuk menjaga kesinambungan konteks.

G. *Pembangunan Model*



Gbr 3. Alur kerja model BM25

Dokumen diindeks menggunakan BM25Okapi yang menghitung skor relevansi berdasaran frekuensi kemunculan kata kunci kueri terhadap dokumen.

III. HASIL DAN ANALISIS

Bagian ini membahas hasil implementasi model dan evaluasi sistem chatbot akademik berbasis Retrieval Augmented Generation (RAG) yang mencakup tahapan ekstraksi data multimodal, pelabelan data, preprocessing, serta analisis hasil pengujian model. Pembahasan difokuskan pada evaluasi kinerja pendekatan Retrieval Augmented Generation (RAG) klasik dalam menghasilkan respons yang relevan dan akurat untuk memenuhi kebutuhan informasi akademik di Universitas Negeri Surabaya.

A. *Hasil Evaluasi Kinerja*

Berdasarkan pengujian sistem, kinerja RAG Klasik dirangkum berikut:

TABLE I
HASIL EVALUASI KINERJA

Metrik	Label Manual	Label Otomatis
Precision	0.47	0.53
Recall	0.47	0.52
F1-Score	0.45	0.52
Cosine Similarity	0.41	0.51
Exact Match	0.21	0.28

Hasil Evaluasi menunjukkan bahwa pendekaran RAG Klasik memiliki pefroma yang lebih optimal pada skenario label otomatis dibandingkan label manual diseluruh metrik pengujian. Skenario label otomatis mencapai nilai F1-Score sebesar 0.52 dan Cosine Similarity sebesar 0.51, lebih unggul dari label manual yang hanya mencatat skor masing-masing 0.45 dan 0.41.

Penggunaan strategi pelabelan otomatis berbasis lexicon matching terbukti efektif dalam memperkaya variasi data basis pengetahuan, sehingga meningkatkan kemampuan sistem dalam menemukan dokumen yang relevan. Namun, secara keseluruhan, rendahnya skor Cosine Similarity dan Exact Match mengonfirmasik keterbatasan algoritma BM25 yang hanya mengandalkan pencocokan kata kunci kaku. Sistem ini cenderung gagal memberikan jawaban akurat ketika pengguna menggunakan variasi bahasa alami atau parafrase yang tidak identik dengan dokumen sumber.

B. *Efisiensi Waktu Respon*

TABLE II
HASIL EFISIENSI WAKTU RESPON

Skenario	Respon Time (S)
Label Manual	0.007
Label Otomatis	0.0017

RAG Klasik menunjukkan keunggulan pada aspek kecepatan sistem. Berdasarkan data tersebut, skenario label manual mencatat waktu respon rata-rata sebesar 0.007 detik, sedangkan skenario label otomatis mencatat waktu yang jauh lebih cepat, yakni 0.0017 detik.

Kecepatan respon yang sangat singkat ini dimungkinkan karena mekanisme RAG Klasik hanya melibatkan proses pencarian dokumen (retrieval) berbasis algoritma BM25 tanpa melalui tahapan model genaratif.

C. *Analisis Kesalahan*

TABLE III
ANALISIS KESALAHAN

Skenario	Query	Ground Truth	Prediction	Jenis Kesalahan
Label Manual	gimana cara masuk perpustakaan unesa pakai sidia?	Akses melalui kartu perpustakaan elektronik (QR Code) pada SIDIA	Menjelaskan pengguna an SIDIA secara umum	Tidak sesuai konteks
Label Otomatis	Bagaimana prosedur pembayaran UKT?	Tahapan dan ketentuan pembayaran UKT	Cara mengubah foto profil mahasiswa	Halusinasi / salah dokumen

RAG Klasik ditemukan memiliki kelemahan pada variasi bahasa karena berbasis kata kunci kaku, system sering gagal jika kueri pengguna menggunakan paraphrase yang tidak ada di dokumen dan kelemahan pada kesalahan konteks, terkadang system mengalami “salah dokumen” (halusinasi pengambilan) jika terdapat kata kunci yang mirip pada kategori yang berbeda.

IV. PEMBAHASAN

Berdasarkan hasil pengujian pada Tabel I, scenario pelabelan otomatis secara konsisten mengunguli pelabelan manual diseluruh metrik pengujian, dengan pencapaian F1-Score sebesar 0.52 dan Cosine Similarity 0.51. keunggulan ini mengindikasikan bahwa penggunaan teknik *lexicon matching* sangat efektif dalam membantu algoritma BM25 untuk mengelompokkan data ke dalam kategori akademik yang tepat, sehingga meminimalisir kesalahan dalam pencarian dokumen referensi. Sebaliknya, pelabelan manual yang bersifat lebih spesifik justru menghasilkan skor yang lebih rendah karena keterbatasan mendasar algoritma BM25 yang hanya mengandalkan pencocokan kata kunci secara kaku tanpa kemampuan pemahaman semantic mendalam.

Dari aspek efisiensi, penelitian ini menemukan bahwa pendekatan RAG Klasik memiliki kecepatan respon yang luar biasa dengan rata-rata waktu dibawah 0.01 detik. Data pada Tabel II menunjukkan bahwa scenario label otomatis bahkan mampu memberikan jawaban dalam waktu secepat 0.0017 detik. Kecepatan respon yang sangat singkat ini dimungkinkan karena mekanisme kerja RAG Klasik hanya melibatkan proses pencarian dokumen (retrieval) berbasis indeks frekuensi kata kunci tanpa melalui tahapan model generatif yang memrlukan beban komputasi tinggi. Hal ini menjadikan RAG Klasik sebagai Solusi yang sangat unggul untuk kebutuhan layanan informasi akademik yang menuntut respon instan dan efisisensi sumber daya.

Meskipun menunjukkan efisisensi tinggi, RAG Klasik memiliki kelemahan mendasar dalam menangani variasi Bahasa alami dan paraphrase dari kueri pengguna. Analisis kesalahan pada Tabel III mengungkapkan bahwa system cenderung mengalami halusinasi apabila pertanyaan pengguna mengandung kata kunci yang mirip namun berada pada kategori yang berbeda. Rendahnya nilai Exact Match yang hanya berada pada rentang 0.21 hingga 0.28 mengonfirmasikan bahwa algoritma berbasis kata kunci kaku ini sulit memberikan jawaban yang identik secara tekstual dengan referensi jika kueri pengguna bersifat ambigu. Keterbatasan ini menegaskan bahwa RAG Klasik lebih ideal untuk informasi prosedura statis dibandingkan untuk kebutuhan interaksi Bahasa alami yang kompleks.

V. KESIMPULAN

Penelitian ini berhasil mengimplementasikan dan mengevaluasi kinerja RAG Klasik untuk chatbot akademik berbasis data multimodal. Berdasarkan hasil analisis, dapat disimpulkan bahwa:

1. RAG Klasik menawarkan efisiensi waktu yang sangat tinggi, dengan waktu respon rata-rata antara 0.0017 hingga 0.007 detik, menjadikannya sangat ideal untuk informasi procedural yang membutuhkan jawaban instan.
2. Kinerja sistem sangat beruntung pada strategi pelabelan, di mana pelabelan otomatis terbukti lebih optimal (F1-Score 0.52) dalam membantu proses pemetaan informasi dibandingkan pelabelan manual.
3. Keterbatsan utama metode ini terletak pada pemahaman semantic yang rendah, di mana algoritma BM25 gagal memberikan jawaban yang relevan jika pengguna menggunakan variasi bahsa alami yang berbeda dari dokumen referensi.

Untuk pengembangan penelitian selanjutnya, disarankan menggunakan Integrasi model modern yaitu menggabungkan arsitektur RAG dengan model generative seperti *Sequence to Sequence* (Seq2seq) atau *embedding* seperti Word2Vec untuk meningkatkan pemahaman kontekstual dan mengatasi kelemahan pada variasi bahasa.

REFERENSI

[1] S. Alqaidi, W. S. Alharbi, and O. Almatrafi, “A support system for college students: A case study of a chatbot application,” in Proc. 2021 19th Int. Conf. Inf. Technol. Based Higher Educ. Training (ITHET), 2021, pp. 1–5, doi: 10.1109/ithet50392.2021.9759796.

[2] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” Mach. Learn. Appl., vol. 2, p. 100006, Dec. 2020, doi: 10.1016/j.mlwa.2020.100006.

[3] S. Mendoza, P. Hernández-León, P. J. Muñoz-Merino, C. Delgado Kloos, and O. C. Santos, “Supporting student–teacher interaction through a chatbot,” in Learning and Collaboration Technologies. Designing, Developing and Deploying Learning Experiences, P. Zaphiris and A. Ioannou, Eds. Cham: Springer, 2020, pp. 93–107, doi: 10.1007/978-3-030-50506-6_8.

[4] Al-farish Alfarizi et al., “Penggunaan Python sebagai bahasa pemrograman untuk machine learning dan deep learning,” Karya Ilmiah Mahasiswa Bertauhid (KARIMAH TAUHID), vol. 2, no. 1, pp. 1–6, 2023.

[5] Amazon Web Services (AWS), “Apa itu Pemrosesan Bahasa Alami (NLP)?,” AWS, 2025. Accessed: Aug. 21, 2025. [Online]. Available: https://aws.amazon.com/id/what-is/nlp/

[6] M. A. Nasution et al., “Implementasi NLP Dalam Pembuatan Chatbot Customer Service Publisher Jurnal Studi Kasus LARISMA,” SAINTEK J. Sains Teknol. Komput., vol. 1, no. 1, pp. 13–17, 2024.

[7] L. Stuhlmann, M. A. Saxer and J. Fürst, "Efficient and Reproducible Biomedical Question Answering Using Retrieval Augmented Generation," *2025 IEEE Swiss Conference on Data Science (SDS)*, Zürich, Switzerland, 2025, pp. 154-157, doi: 10.1109/SDS66131.2025.00029.

[8] M. S. Oghli and M. M. Almustafa, “Comparison of basic Information Retrieval Models,” Int. J. Eng. Res. Technol. (IJERT), vol. 10, no. 09, Sep. 2021.

[9] W. A. G. Kodri, M. Haris, and R. Fitriadi, “Fine-Hybrid: Integration of BM25 and Finetuned SBERT to Enhance Search Relevance,” *Teknika*, vol. 14, no. 2, pp. 213–222, Jul. 2025, doi: 10.34148/teknika.v14i2.1229.

[10] H. Tohir, N. Merlina, and M. Haris, “Utilizing retrieval-augmented generation in large language models to enhance Indonesian language NLP,” J. Inform. dan Teknol. Komput. (JITK), vol. 10, no. 2, pp. 352–360, 2024, doi: 10.33480/jitk.v10i2.5916.

[11] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint arXiv:2312.10997, 2023, doi: 10.48550/arXiv.2312.10997.

[12] Y. Setiawan, M. H. Z. Al Faroby, M. N. P. Ma’ady, I. M. W. A. Sanjaya, and C. V. C. Ramadhani, “Modality-based modeling with data balancing and dimensionality reduction for early stunting detection,” Jurnal Online Informatika (JOIN), vol. 10, no. 1, pp. 53–65, Jun. 2025, doi: 10.15575/join.v10i1.1495.

[13] K. Palasundram et al., "Enhancements to the sequence-to-sequence based natural answer generation models," IEEE Access, vol. 8, pp. 41103–41113, 2020, doi: 10.1109/access.2020.2978551.
[14] S. Li, "Understanding Word2Vec embedding in practice," Medium, 2019. Accessed: Sep. 05, 2025. [Online]. Available: https://medium.com/datascience/understanding-word2vec-embedding-in-practice-3e9b8985953
[15] Q. Maulida and D. B. Santoso, "Aplikasi layanan dan informasi akademik berbasis chatbot Telegram menggunakan natural language processing," J. Teknol. Inf. dan Komun. (JTIK), vol. 8, no. 2, 2024, doi: 10.35870/jtik.v8i2.1887.
[16] A. Nurdin, B. Aji, et al., "Perbandingan kinerja word embedding Word2Vec, GloVe, dan FastText pada klasifikasi teks," J. Tekno Kompak, vol. 14, no. 2, pp. 74–79, 2020, doi: 10.33365/jtk.v14i2.732.
[17] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," Multimed. Tools Appl., vol. 83, pp. 37979–38007, 2024, doi: 10.1007/s11042-023-17007-z.
[18] H. Pranata Tarigan, "Integrasi chatbot berbasis NLP pada sistem layanan akademik," J. Komput., vol. 3, no. 1, pp. 13–18, 2024, doi: 10.70963/jk.v3i1.110.
[19] M. S. Salim and S. H. Imran, "LLM-based QA chatbot builder: A generative AI-based chatbot builder for question answering," SoftwareX, vol. 27, p. 102029, 2024, doi: 10.1016/j.softx.2024.102029.
[20] H. Touvron et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023, doi: 10.48550/arXiv.2302.13971.
[21] A. Vinayan Kozhipuram, S. Shailendra, and R. Kadel, "Retrieval-Augmented Generation vs. Baseline LLMs: A multi-metric evaluation for knowledge-intensive content," Information, vol. 16, no. 9, p. 766, 2025, doi: 10.3390/info16090766.
[22] A. Zubaidi et al., "Integrasi sistem informasi akademik dan bot Telegram sebagai media pengaksesan informasi di Universitas Mataram," J. Teknol. Inf. dan Komput. (JTIKA), vol. 3, no. 2, pp. 253–268, 2021.
[23] L. R. Hidayat, I. G. P. S. Wijaya, and R. Dwiyansaputra, "Optimalisasi layanan sistem informasi mahasiswa dengan integrasi Telegram: Chatbot retrieval-augmented-generation berbasis large language model," J. Teknol. Inf. Komput. dan Apl. (JTIKA), vol. 7, no. 1, 2025, doi: 10.29303/jtika.v7i1.459.
[24] M. Al-Amin et al., "History of generative Artificial Intelligence (AI) chatbots: past, present, and future development," arXiv preprint arXiv:2402.05122, 2024, doi: 10.48550/arXiv.2402.05122.
[25] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Adv. Neural Inf. Process. Syst. 33 (NeurIPS 2020), 2020, pp. 9459–9474.
[26] S. Gupta et al., "Integrating Knowledge Retrieval with Generation: A Comprehensive Survey of RAG Models in NLP," Preprints.org, 2025, doi: 10.20944/preprints202504.0351.v1.
[27] I. Pratama and B. Sisephaputra, "Pengembangan sistem helpdesk menggunakan chatbot dengan metode retrieval-augmented generation (RAG)," J. Inform. Comput. Sci. (JINACS), vol. 6, no. 3, pp. 696–710, 2024, doi: 10.26740/jinacs.v6n03.p696-710.
[28] L. Danuarta, V. C. Mawardi, and V. Lee, "Retrieval-Augmented Generation (RAG) large language model for educational chatbot," in Proc. 2024 Ninth Int. Conf. Informatics Comput. (ICIC), 2024, doi: 10.1109/icic64337.2024.10957676.
[29] N. B. Korade, M. B. Salunke, et al., "Strengthening Sentence Similarity Identification Through OpenAI Embeddings and Deep Learning," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 4, pp. 821–829, 2024, doi: 10.14569/IJACSA.2024.0150485.