IMPLEMENTASI HYBRID RETRIEVAL DENGAN METODE RECIPROCAL RANK FUSION (RRF) PADA CHATBOT LAYANAN AKADEMIK PROGRAM STUDI SISTEM INFORMASI UNIVERSITAS MUHAMMADIYAH JEMBER

M. Vicky Mosafan[1], Deni Arifianto[2], Wiwik Suharso[3]
mvickymosafan@gmail.com[1], deniarifianto@unmuhjember.ac.id[2],
wiwiksuharso@unmuhjember.ac.id[3]
Universitas Muhammadiyah Jember

ABSTRAK

Layanan informasi akademik pada Program Studi Sistem Informasi Universitas Muhammadiyah Jember masih banyak bergantung pada layanan tatap muka dan komunikasi manual. Akibatnya, mahasiswa sering mengalami keterlambatan dalam memperoleh informasi terkait ketentuan pelaksanaan Kuliah Kerja Nyata (KKN), Kerja Praktik, Tugas Akhir, maupun perubahan kurikulum yang berlaku. Penelitian ini bertujuan menerapkan dan mengevaluasi chatbot layanan akademik berbasis Retrieval-Augmented Generation (RAG) yang menggabungkan semantic retrieval dan BM25 reranking melalui metode reciprocal rank fusion (RRF). Dalam implementasinya, metode RRF diterapkan tanpa skema pembobotan tambahan, dengan konstanta k = 60, sehingga semantic retrieval dan BM25 memberikan kontribusi yang setara dalam penggabungan peringkat dokumen. Sistem dibangun menggunakan Google Gemini Embeddings, Supabase pgvector, orkestrasi n8n, model bahasa Cohere Aya Expanse 32B, serta Redis Cache untuk efisiensi waktu respons. Pengujian dilakukan terhadap 40 query menggunakan metrik Recall dan Mean Reciprocal Rank (MRR), serta pengujian waktu respons pada kondisi cache miss dan cache hit. Hasil pengujian menunjukkan bahwa RRF standar memperoleh Recall rata-rata 86,88% dan MRR 93,75%, sedangkan perbandingan skenario Semantic Retrieval + BM25 Reranking dengan skenario Hybrid Retrieval + RRF menunjukkan peningkatan Recall dari 66,67% menjadi 100,00% dan MRR dari 0,49 menjadi 0,92. Redis Cache mampu mempercepat waktu respons dari 18,326 detik pada kondisi cache miss menjadi 694 milidetik pada kondisi cache hit, atau sekitar 26,4 kali lebih cepat. Hasil ini membuktikan bahwa penerapan RRF pada hybrid retrieval mampu meningkatkan relevansi dan kualitas pemeringkatan dokumen, sementara Redis Cache secara signifikan meningkatkan efisiensi waktu respons chatbot layanan akademik.

Kata Kunci: Chatbot Akademik, Hybrid Retrieval, Reciprocal Rank Fusion, Retrieval-Augmented Generation, Redis Cache.

ABSTRACT

Academic information services in the Information Systems Study Program of Universitas Muhammadiyah Jember still largely rely on face-to-face services and manual communication. As a result, students often experience delays in obtaining information regarding the regulations governing the implementation of the Community Service Program (KKN), Internship Program, Final Project, and the applicable curriculum. This study aims to implement and evaluate an academic service chatbot based on Retrieval-Augmented Generation (RAG) that combines semantic retrieval and BM25 reranking using the reciprocal rank fusion (RRF) method. In its implementation, RRF was applied in its standard, unweighted form (RRF) with a constant of k = 60, giving semantic retrieval and BM25 an equal contribution when fusing document rankings. The system was built using Google Gemini Embeddings, Supabase pgvector, n8n orchestration, the Cohere Aya Expanse 32B language model, and Redis Cache for response-time efficiency. Testing was conducted on 40 queries using Recall and Mean Reciprocal Rank (MRR) metrics, along with response-time testing under cache-miss and cache-hit conditions. The results show that standard RRF achieved an average Recall of 86.88% and MRR of 93.75%, while comparison between the Semantic Retrieval + BM25 Reranking scenario and the Hybrid Retrieval + RRF scenario showed an increase in Recall from 66.67% to 100.00% and MRR from 0.49 to 0.92. Redis Cache accelerated response time from

18.326 seconds under cache-miss conditions to 694 milliseconds under cache-hit conditions, approximately 26.4 times faster. These results demonstrate that applying RRF in hybrid retrieval improves document relevance and ranking quality, while Redis Cache significantly improves the response-time efficiency of the academic service chatbot.
Keywords: *Academic Chatbot, Hybrid Retrieval, Reciprocal Rank Fusion, Retrieval-Augmented Generation, Redis Cache.*

PENDAHULUAN

Layanan informasi akademik merupakan salah satu aspek penting dalam mendukung kelancaran proses studi mahasiswa. Pada Program Studi Sistem Informasi Universitas Muhammadiyah Jember, sebagian besar layanan informasi terkait Kuliah Kerja Nyata (KKN), Kerja Praktik, Tugas Akhir, dan kurikulum masih dilakukan secara tatap muka atau melalui komunikasi manual dengan staf maupun dosen. Kondisi tersebut menyebabkan mahasiswa harus menunggu jam operasional layanan administrasi, sehingga proses pencarian informasi menjadi kurang efisien, khususnya untuk pertanyaan yang bersifat rutin dan berulang.

Perkembangan Large Language Model (LLM) mendorong pemanfaatan chatbot berbasis Retrieval-Augmented Generation (RAG) sebagai solusi penyediaan layanan informasi yang responsif dan tersedia sepanjang waktu. Pada pendekatan RAG, sistem terlebih dahulu melakukan pencarian dokumen relevan dari basis pengetahuan sebelum dokumen tersebut diuntukkan sebagai konteks bagi model bahasa dalam menghasilkan jawaban, sehingga respons yang diberikan tetap berpijak pada sumber informasi resmi dan tidak semata-mata bergantung pada pengetahuan internal model.

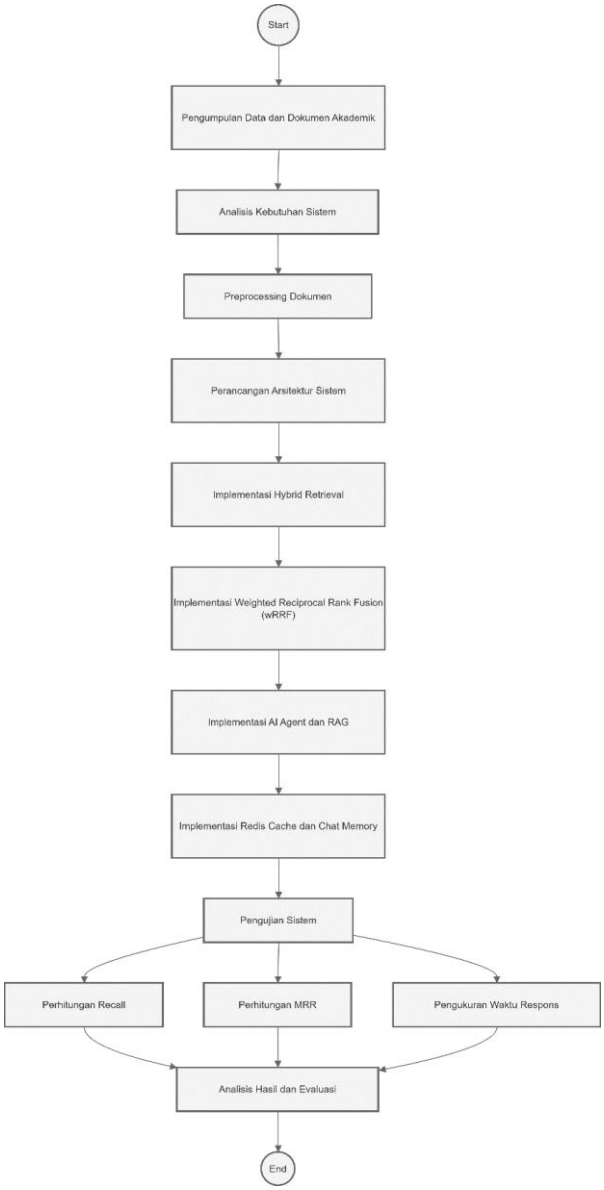
Kualitas jawaban chatbot berbasis RAG sangat ditentukan oleh kemampuan komponen retrieval dalam menemukan dokumen yang relevan. Semantic retrieval unggul dalam menangkap kesamaan makna antar-kalimat, namun kurang optimal ketika query memuat istilah teknis, nama dokumen, atau kode yang bersifat eksplisit. Sebaliknya, pendekatan lexical retrieval seperti BM25 unggul pada pencocokan kata kunci, tetapi kurang mampu menangkap variasi makna. Kombinasi kedua pendekatan melalui hybrid retrieval menjadi strategi yang umum diuntukkan untuk saling melengkapi kelemahan masing-masing metode, dengan reciprocal rank fusion (RRF) sebagai salah satu metode penggabungan peringkat yang populer karena kesederhanaannya.

RRF standar memberikan kontribusi yang sama besar kepada setiap retriever tanpa memerlukan proses tuning bobot secara manual, sehingga metode ini bersifat sederhana namun tetap mampu menggabungkan kelebihan semantic retrieval dan BM25 secara seimbang. Pada dokumen akademik yang banyak memuat istilah teknis, nomor dokumen, dan prosedur baku, karakteristik inilah yang mendasari penelitian ini untuk menerapkan reciprocal rank fusion (RRF), serta mengintegrasikan Redis Cache untuk meningkatkan efisiensi waktu respons chatbot layanan akademik Program Studi Sistem Informasi Universitas Muhammadiyah Jember.

Berdasarkan uraian tersebut, penelitian ini dirumuskan ke dalam dua permasalahan utama, yaitu bagaimana menerapkan hybrid retrieval yang menggabungkan sparse dan dense retrieval menggunakan RRF pada chatbot layanan akademik, serta bagaimana mengukur kinerja chatbot tersebut berdasarkan metrik Recall, Mean Reciprocal Rank (MRR), dan waktu respons. Tujuan penelitian ini adalah menerapkan hybrid retrieval berbasis RRF pada chatbot layanan akademik dan mengevaluasi kinerjanya berdasarkan ketiga metrik tersebut, sehingga diperoleh gambaran objektif mengenai kontribusi RRF dan Redis Cache terhadap relevansi jawaban dan efisiensi sistem.

METODOLOGI

Penelitian ini dilaksanakan melalui beberapa tahapan, meliputi analisis kebutuhan sistem, pengumpulan data dan dataset, perancangan arsitektur hybrid retrieval, implementasi sistem, hingga pengujian dan evaluasi kinerja. Setiap tahapan dirancang secara berurutan agar hasil pada satu tahap dapat menjadi dasar bagi tahap berikutnya, sehingga sistem yang dibangun sesuai dengan kebutuhan layanan akademik dan dapat dievaluasi secara objektif.



HASIL DAN PEMBAHASAN

Implementasi Sistem

Sistem chatbot layanan akademik berhasil diimplementasikan dengan basis pengetahuan yang terdiri atas empat dokumen resmi Program Studi Sistem Informasi. Tahap preprocessing menghasilkan potongan dokumen (chunk) yang telah dilengkapi metadata;

1. fileName
2. fileType
3. processedAt
4. contentLength

5. documentHash
6. embedding vector

sehingga siap diuntukkan pada proses retrieval. Retrieval pipeline berhasil dijalankan secara end-to-end melalui orkestrasi n8n, mulai dari penerimaan query, pengecekan Redis Cache, semantic retrieval terhadap 50 dokumen kandidat, BM25 reranking, penggabungan RRF, hingga generasi jawaban oleh Cohere Aya Expanse 32B dan pencatatan metrik evaluasi.

Analisis Perhitungan reciprocal rank fusion (RRF)

Evaluasi terhadap 40 query menggunakan RRF (k = 60) menghasilkan Recall rata-rata sebesar 86,88% dan MRR rata-rata sebesar 93,75%. Karena setiap retriever diberi kontribusi yang setara, posisi akhir dokumen ditentukan murni oleh penjumlahan skor 1/(k + rank) dari semantic retrieval dan BM25 tanpa penyesuaian bobot tambahan. Untuk mengilustrasikan mekanisme penggabungan tersebut secara konkret, dipilih enam query yang paling representatif (powerful) dari keseluruhan 40 query pengujian, sebagaimana dipaparkan pada Tabel 1.

Keenam query tersebut dipilih karena secara konsisten menempatkan dokumen paling relevan pada peringkat atas, baik pada semantic retrieval maupun BM25, sehingga skor RRF yang dihasilkan mencerminkan secara jelas bagaimana kedua retriever saling melengkapi tanpa memerlukan penyesuaian bobot. Rincian peringkat dan skor RRF keenam query tersebut disajikan pada Tabel 1.1.

Tabel 1. Perhitungan Skor RRF pada Enam Query Terpilih

Query Terpilih	Peringkat Semantic	Peringkat Keyword (BM25)	Skor RRF
Siapa Dekan Fakultas Teknik dan Ketua Program Studi pada Lembar Pengesahan Panduan Tugas Akhir?	1	1	0,032787
Apa nama undang-undang nomor 14 tahun 2005 yang menjadi landasan hukum penyusunan panduan Tugas Akhir?	1	3	0,032266
Kemampuan apa yang diharapkan dari mahasiswa dalam memecahkan permasalahan nyata yang relevan dengan bidang sistem informasi?	5	2	0,031514
Apa tujuan utama Program Studi Sistem Informasi menyusun buku panduan TA yang memperhatikan standar nasional?	1	2	0,032522
Berapa lama mahasiswa biasanya melaksanakan proses pembelajaran Kerja Praktik (KP) di luar kampus?	1	1	0,032787
Manfaat apa yang bisa didapatkan oleh instansi eksternal dari pelaksanaan Kerja Praktik mahasiswa?	1	2	0,032522

Sebagai ilustrasi, pada query “Siapa Dekan Fakultas Teknik dan Ketua Program Studi pada Lembar Pengesahan Panduan Tugas Akhir?” dan query “Berapa lama mahasiswa biasanya melaksanakan proses pembelajaran Kerja Praktik (KP) di luar kampus?”, dokumen paling relevan berhasil menempati peringkat pertama baik pada semantic retrieval maupun BM25, sehingga skor RRF dihitung sebagai

$$\frac{1}{60+1}+\frac{1}{60+1}=\frac{1}{61}+\frac{1}{61}=0.032787$$

skor tertinggi di antara keenam query. Pada query "Kemampuan apa yang diharapkan dari mahasiswa dalam memecahkan permasalahan nyata yang relevan dengan bidang sistem informasi?", dokumen relevan berada pada peringkat kelima semantic retrieval namun peringkat kedua BM25, menghasilkan skor RRF sebesar

$$\frac{1}{60+5}+\frac{1}{60+2}=\frac{1}{65}+\frac{1}{62}\approx 0.031514$$

skor terendah di antara keenam query tersebut. Ketiga ini menunjukkan bahwa RRF tetap mampu mengangkat dokumen dengan kecocokan kuat pada salah satu metode retrieval meskipun tanpa pembobotan tambahan, karena kontribusi 1/(k+"rank") dari peringkat teratas sudah cukup signifikan terhadap skor akhir dan proporsional terhadap posisi peringkat pada setiap retriever.

Perbandingan Skenario Retrieval

Pengujian utama membandingkan Skenario 1 (Semantic Retrieval + BM25 Reranking) dan Skenario 2 (Hybrid Retrieval + RRF) terhadap enam pertanyaan yang identik. Ringkasan hasil pengujian disajikan pada Tabel 2.1.

Tabel 2. Ringkasan Perbandingan Kinerja Skenario 1 dan Skenario 2

Metrik	Skenario 1	Skenario 2	Selisih
Recall rata-rata	66,67%	100,00%	+33,33%
MRR rata-rata	0,49	0,92	+0,43
Query dengan Recall = 100%	2 dari 6 (33,33%)	6 dari 6 (100,00%)	+4 query
Query dengan MRR = 1,00	1 dari 6 (16,67%)	5 dari 6 (83,33%)	+4 query
Kategori Sangat Baik	1 dari 6 (16,67%)	6 dari 6 (100,00%)	+5 query

Hasil pengujian menunjukkan bahwa Skenario 2 secara konsisten memberikan kinerja yang lebih baik dibandingkan Skenario 1. Secara keseluruhan performa meningkat meskipun beberapa query mengalami penurunan MRR.

Pada Skenario 1, urutan akhir dokumen hanya ditentukan oleh BM25 reranking sehingga informasi semantic retrieval tidak dimanfaatkan pada penentuan peringkat akhir, kondisi ini terlihat dari nilai keyword recall dan keyword MRR yang identik dengan nilai Recall dan MRR utama pada Skenario 1.

Sebaliknya, Skenario 2 menggabungkan kedua peringkat melalui RRF sehingga dokumen dengan relevansi tinggi pada salah satu maupun kedua metode memiliki peluang lebih besar menempati peringkat atas. Rata-rata Recall Skenario 2 (100,00%) bahkan melampaui keyword recall maupun semantic recall yang masing-masing sebesar 75,00%, membuktikan bahwa penggabungan melalui RRF menghasilkan kinerja retrieval yang lebih baik dibandingkan penguntukan salah satu metode secara terpisah.

Pengujian Waktu Respons dan Efisiensi Redis Cache

Pengujian waktu respons dilakukan pada dua kondisi, yaitu cache miss dan cache hit. Pada kondisi cache miss, sistem menjalankan seluruh pipeline mulai dari pembentukan embedding query, similarity search terhadap 50 dokumen kandidat, perhitungan BM25, penggabungan RRF, hingga generasi jawaban oleh Cohere Aya Expanse 32B, dengan total waktu respons 18,326 detik dan konsumsi 2.767 token. Pada kondisi cache hit, sistem hanya menjalankan enam node awal (penerimaan pesan, pembentukan cache key, pengecekan Redis, pemrosesan hasil cache, cache router, dan pengembalian jawaban) tanpa menjalankan ulang retrieval maupun generasi jawaban, dengan waktu respons 694 milidetik. Ringkasan perbandingan disajikan pada Tabel 3.1.

Tabel 3. Perbandingan Waktu Respons pada Kondisi Cache Miss dan Cache Hit

Kondisi	Waktu Respons	Node yang Dieksekusi	Keterangan
Cache Miss	18,326 detik	Seluruh pipeline (retrieval + LLM)	Query baru, tidak ditemukan di cache
Cache Hit	694 milidetik	6 node awal saja	Jawaban langsung diambil dari Redis
Peningkatan Efisiensi	≈ 26,4x lebih cepat	-	-

Hasil pengujian menunjukkan bahwa Redis Cache mampu mempercepat waktu respons sekitar 26,4 kali lipat pada kondisi cache hit dibandingkan cache miss, karena sistem tidak perlu mengulang proses pembentukan embedding, pencarian vektor, reranking, maupun pemanggilan model bahasa. Mekanisme ini sangat relevan bagi chatbot akademik yang sering menerima pertanyaan berulang, seperti persyaratan skripsi, prosedur KKN, atau kerja praktik. Setiap data cache disimpan dengan TTL selama 3.600 detik sehingga akan terhapus otomatis setelah masa berlakunya berakhir, menjaga keseimbangan antara efisiensi waktu respons dan kemutakhiran informasi yang diberikan kepada penguntuk.

Pembahasan Umum

Secara keseluruhan, hasil pengujian membuktikan bahwa kombinasi semantic retrieval dan BM25 reranking yang digabungkan melalui RRF memberikan hasil retrieval yang lebih relevan dibandingkan penguntukan BM25 reranking maupun semantic retrieval secara terpisah tanpa fusion. Kontribusi yang setara antara semantic retrieval dan BM25 pada RRF terbukti cukup untuk mengangkat dokumen yang unggul pada salah satu metode retrieval, tanpa memerlukan proses tuning bobot tambahan yang berpotensi menambah kompleksitas maupun risiko overfitting terhadap karakteristik dataset tertentu. Di sisi lain, integrasi Redis Cache terbukti menjadi komponen penting dalam menjaga efisiensi waktu respons chatbot tanpa mengorbankan konsistensi jawaban yang dihasilkan, sehingga arsitektur hybrid retrieval berbasis RRF dengan Redis Cache ini layak dijadikan acuan pengembangan chatbot layanan akademik pada program studi atau fakultas lain.[1]

KESIMPULAN

Berdasarkan hasil implementasi dan pengujian, dapat disimpulkan bahwa hybrid retrieval yang menggabungkan semantic retrieval dan BM25 reranking menggunakan reciprocal rank fusion (RRF) dengan RRF (k = 60) berhasil diterapkan pada chatbot layanan akademik Program Studi Sistem Informasi Universitas Muhammadiyah Jember. Penerapan RRF, yang memberikan kontribusi setara pada setiap retriever tanpa skema pembobotan tambahan, terbukti tetap mampu memprioritaskan dokumen dengan kecocokan kata kunci kuat maupun kesamaan makna, sebagaimana diilustrasikan pada enam query terpilih dari 40 query pengujian.

Hasil evaluasi kinerja menunjukkan bahwa RRF standar mencapai Recall rata-rata 86,88% dan MRR 93,75% pada 40 query pengujian, sedangkan perbandingan skenario menunjukkan bahwa Hybrid Retrieval dengan RRF (Skenario 2) meningkatkan Recall dari 66,67% menjadi 100,00% dan MRR dari 0,49 menjadi 0,92 dibandingkan skenario Semantic Retrieval + BM25 Reranking tanpa fusion (Skenario 1), meskipun pada beberapa query masih ditemukan penurunan nilai MRR. Dari sisi efisiensi, integrasi Redis Cache berhasil mempercepat waktu respons sistem dari 18,326 detik pada kondisi cache miss

[1] Zainal Abidin Bagir, *Sains, Teknologi, dan Agama dalam Perspektif Islam Kontemporer* (Yogyakarta: CRCS UGM, 2023), hlm. 98-102.

menjadi 694 milidetik pada kondisi cache hit, atau sekitar 26,4 kali lebih cepat, sehingga chatbot mampu memberikan layanan informasi akademik yang lebih relevan, akurat, dan responsif bagi mahasiswa.

Untuk penelitian selanjutnya, disarankan dilakukan eksperimen terhadap variasi nilai konstanta k pada RRF secara sistematis untuk mengetahui pengaruhnya terhadap kualitas peringkat dokumen, perluasan dataset dan jumlah query pengujian agar hasil evaluasi lebih representatif, integrasi langsung dengan Sistem Informasi Akademik (SIA) universitas, serta penambahan evaluasi kualitatif terhadap kepuasan dan persepsi penguntuk terhadap chatbot layanan akademik.

DAFTAR PUSTAKA

Abdurrazzaq, M. A., Tjiong, E. L., & Fasya, A. (2025). An Indonesian Chatbot for Disease Diagnosis Using Retrieval-Augmented Generation. Jurnal Ilmiah, 10(3), 1877–1887.

Aljohani, B., & Alsanoosy, T. (2026). Enhancing Medical Question Answering with LLMs via a Hybrid Retrieval-Augmented Generation Framework. Information (Switzerland), 17(2). https://doi.org/10.3390/info17020133

Bruch, S., Gai, S., & Ingber, A. (2023). An Analysis of Fusion Functions for Hybrid Retrieval. ACM Transactions on Information Systems, 42(1). https://doi.org/10.1145/3596512

Davar, N. F., Dewan, M. A. A., & Zhang, X. (2025). AI Chatbots in Education: Challenges and Opportunities. Information (Switzerland), 16(3). https://doi.org/10.3390/info16030235

Efendi, G., Mutawalli, L., Taufan, M., & Zaen, A. (2025). Layanan Informasi Akademik Stmik Lombok. ZONAsi: Jurnal Sistem Informasi, 7(3), 956–969.

Ennion, M., & McLellan, R. (2025). Large Language Model Chatbots in Education: Exploring Literature Insights on Their Impact and Influence on Learning Behaviours. Studies in Technology Enhanced Learning, 4(1). https://doi.org/10.21428/8c225f6e.540b41b5

Hambarde, K. A., & Proença, H. (2023). Information Retrieval: Recent Advances and Beyond. IEEE Access, 11, 76581–76604. https://doi.org/10.1109/ACCESS.2023.3295776

Irany, F. A., & Akwafuo, S. (2026). A Hybrid Retrieval and Reranking Framework for Evidence-Grounded Retrieval-Augmented Generation. arXiv. http://arxiv.org/abs/2605.01664

Jani, S., Heidari, A., Anvari, A., & Rahimi, Z. (2025). Intelligent Scientific Literature Explorer using Machine Learning (ISLE). arXiv. http://arxiv.org/abs/2512.12760

Kaptosv, L. (2025). Using Redis for Caching Optimization in High-Traffic Web Applications. International Journal of Advanced Multidisciplinary Research and Studies, 5, 1714–1722. https://doi.org/10.62225/2583049X.2025.5.4.4839

Mala, C. S., Gezici, G., & Giannotti, F. (2025). Hybrid Retrieval for Hallucination Mitigation in Large Language Models: A Comparative Analysis. arXiv. http://arxiv.org/abs/2504.05324

Melati, L. A., Swanjaya, D., & Pamungkas, D. P. (2025). Pencarian dan Evaluasi Relevansi Dongeng Bahasa Indonesia Berbasis Semantik Menguntukkan Latent Semantic Indexing (LSI). Rabit: Jurnal Teknologi dan Sistem Informasi Univrab, 10(2), 361–372. https://doi.org/10.36341/rabit.v10i2.6093

Minukhin, S., & Bukhalo, V. (2026). A Study of the Impact of the Weighted RRF Method on the Quality of Recommender Systems. Nauka i Tekhnika S′ohodni, 3(57). https://doi.org/10.52058/2786-6025-2026-3(57)-1865-1879

Muhammad Dzaki Salman, Rahmaddeni, Torkis Nasution, & Susanti. (2026). Implementation of Retrieval-Augmented Generation Method on Large Language Model for Development of Campus Service and Information Chatbot. INOVTEK Polbeng - Seri Informatika, 11(1), 298–309. https://doi.org/10.35314/3y9hy151

Ningsih. (2025). Evaluasi Tingkat Kepuasan Mahasiswa Terhadap Layanan Sistem Informasi Akademik atau Siakad di Universitas Tulungagung. Jurnal Ilmiah, 4(1), 23–34.

Pramudia, F. A., Zulfa, M. I., & Aliim, M. S. (2025). Effect In-Memory Based Cache System on Web Applications in Improving Data Access Performance. Dinamika Rekayasa, 21(2), 166–174. https://doi.org/10.20884/1.jidr.2025.21.2.12

Pratama, I. I. R., & Sisephaputra, B. (2024). Pengembangan Sistem Helpdesk Menguntukkan Chatbot dengan Metode Retrieval-Augmented Generation (RAG). Journal of Informatics and Computer Science (JINACS), 6(03), 696–710. https://doi.org/10.26740/jinacs.v6n03.p696-710

Procko, T. T., & Ochoa, O. (2024). Graph Retrieval-Augmented Generation for Large Language Models: A Survey. Proceedings - 2024 Conference on AI, Science, Engineering, and Technology (AIxSET), 166–169. https://doi.org/10.1109/AIxSET62544.2024.00030

Rackauckas, Z. (2024). RAG-Fusion: A New Take on Retrieval Augmented Generation. International Journal on Natural Language Computing, 13(1), 37–47. https://doi.org/10.5121/ijnlc.2024.13103

Ramadhan, T. I., Supriatman, A., & Kurniawan, T. R. (2024). Evaluasi dan Implementasi IndoBERT Question Answering (QA) pada Domain Spesifik Menguntukkan Mean Reciprocal Rank. 180–188. https://doi.org/10.33364/algoritma/v.21-1.1542

Ramadhani, T. Q., Nada, N. Q., & S, N. D. (2025). Penerapan Metode Retrieval-Augmented Generation (RAG) pada Chatbot E-Commerce Berbasis Gemini AI. Jurnal Ilmiah ILKOMINFO, 8(2), 301–313. https://doi.org/10.47324/ilkominfo.v8i2.384

Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. African Journal of Biomedical Research, 27(4), 4023–4031. https://doi.org/10.53555/AJBR.v27i4S.4345

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. Proceedings of the International Conference on Multimedia Information Processing and Retrieval (MIPR), 155–161. https://doi.org/10.1109/MIPR62202.2024.00031

Sivanandan, M. (2025). Balancing the Scales: Making reciprocal rank fusion (RRF) Smarter with Weights. Elasticsearch. https://www.elastic.co/search-labs/blog/weighted-reciprocal-rank-fusion-rrf

Suasnawa, I., Wiratama, I., Sudiartha, I., Caturbawa, I., Sapteka, A., & Indrayana, I. (2023). Chatbot-Based Student Information Service in Indonesian Language. 223–227. https://doi.org/10.5220/0011753800003575

Suharti, & Sholihah, H. A. (2025). Perbandingan Recall and Precision pada Sistem Temu Kembali Informasi Lama (Simpus) dan Baru (Simpustaka) di Perpustakaan Universitas Islam Indonesia. Jurnal Pustaka Ilmiah, 11(2), 187–199. https://doi.org/10.20961/jpi.v11i2.109403

Suharyadi, & Saputra, I. (2025). Hybrid Ensemble Retrieval-Augmented Generation for Indonesian Legal Consultation with Keyword Boosting. Journal of Novel Engineering Science and Technology, 4(02), 71–85. https://doi.org/10.56741/jnest.v4i02.1042

Youseff, A. S., et al. (2026). Implementation of Redis-Based Caching to Reduce Latency in

Web Applications.
Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., & Wen, J.-R. (2026). Large Language Models for Information Retrieval: A Survey. ACM Transactions on Information Systems, 44(1), 1–54. https://doi.org/10.1145/3748304