Penerapan *Small Language Model* Berbasis *Retrieval-Augmented Generation* untuk Pemahaman Dokumen Teknis

I Gusti Ngurah Agung Putu Kresna Putra[1a)], Gede Angga Pradipta[2b)], Roy Rudolf Huizen [3c)]
[1)]Magister Sistem Informasi Institut Teknologi dan Bisnis STIKOM Bali Denpasar, Indonesia
e-mail: kresnaputra05@gmail.com[a], angga_pradipta@stikom-bali.ac.id[b], roy@stikom-bali.ac.id[c]

Abstrak

Dalam lingkungan organisasi, dokumen teknis internal seperti dokumentasi sistem, panduan operasional, dan spesifikasi perangkat lunak umumnya disajikan dalam bentuk PDF yang panjang dan kompleks. Kondisi ini menyulitkan pengguna dalam memperoleh informasi spesifik secara cepat dan akurat melalui pencarian manual. Sistem tanya jawab berbasis model bahasa pada dokumen teknis telah digunakan sebagai solusi, namun masih menghadapi permasalahan keterbatasan konteks dan kecenderungan menghasilkan jawaban yang tidak didukung oleh sumber informasi yang valid (hallucination). U ntuk mengatasi permasalahan tersebut, penelitian ini menerapkan pendekatan Retrieval-Augmented Generation (RAG) dengan memanfaatkan Small Language Model (SLM) sebagai mekanisme inferensi utama. Dokumen teknis diproses melalui tahapan ekstraksi teks, pemisahan berbasis struktur dokumen, dan chunking semantik untuk membentuk unit informasi yang koheren, yang selanjutnya direpresentasikan dalam bentuk embedding vektor dan disimpan dalam basis data vektor berbasis FAISS. Pada tahap inferensi, pertanyaan pengguna dipetakan ke dalam ruang vektor untuk memperoleh konteks yang paling relevan, yang kemudian digunakan oleh SLM dalam menghasilkan jawaban yang dibatasi secara ketat pada informasi yang tersedia dalam dokumen. Hasil eksperimen menunjukkan bahwa pendekatan SLM berbasis RAG mampu menghasilkan jawaban yang lebih relevan dan konsisten terhadap sumber dokumen serta secara efektif mengurangi kecenderungan hallucination dibandingkan dengan pendekatan generatif tanpa mekanisme retrieval.

Kata kunci: *Small Language Model, Retrieval-Augmented Generation, Question Answering, Dokumen Teknis, Sistem Informasi.*

1. Pendahuluan

Dokumen Kemajuan teknologi dalam bidang *Natural Language Processing* (NLP) telah membawa perkembangan signifikan terhadap kemampuan model bahasa besar (*Large Language Models* / LLM) seperti GPT, LLaMA, dan Gemini dalam memahami konteks serta menghasilkan teks alami yang mendekati kemampuan manusia. Namun demikian, penggunaan LLM masih menghadapi sejumlah kendala, terutama pada aspek biaya komputasi, privasi data, serta fenomena *hallucination* di mana model menghasilkan informasi yang tidak bersumber dari data aktual [1], [13].

Untuk mengatasi keterbatasan tersebut, pendekatan *Retrieval-Augmented Generation* (RAG) diperkenalkan sebagai solusi yang menggabungkan kemampuan generatif model bahasa dengan *retrieval system* berbasis konteks yang relevan [3], [5]. Pada mekanisme ini, sistem tidak hanya mengandalkan parameter model, tetapi juga melakukan pencarian semantik terhadap sumber informasi eksternal menggunakan *embedding* vektor dan *vector store* seperti FAISS. Hasil pencarian (*retrieved context*) kemudian disediakan sebagai masukan bagi model generatif untuk menghasilkan jawaban yang faktual dan berbasis sumber [4], [8]. Pendekatan ini terbukti meningkatkan akurasi dan transparansi sistem tanya jawab berbasis dokumen, sekaligus mengurangi ketergantungan terhadap model besar berbasis cloud [9], [16].

Seiring dengan berkembangnya penelitian di bidang ini, muncul tren baru yaitu penerapan Small Language Model (SLM) dalam sistem RAG. SLM menawarkan efisiensi komputasi dan privasi data yang lebih baik tanpa mengorbankan relevansi hasil [6], [7]. Studi oleh Shetty [5] dan Trangcasanchai [7] menunjukkan bahwa kombinasi SLM dengan RAG dapat menghasilkan performa mendekati LLM untuk kasus pemahaman dokumen teknis. Hasil serupa juga ditemukan oleh Muludi et al. [2] dan Alat & Hermawan [10], yang mengimplementasikan sistem tanya jawab berbasis FAISS dan model MiniLM pada dokumen akademik.

Dalam konteks organisasi, pendekatan RAG dengan SLM menjadi solusi potensial bagi sistem knowledge management internal, terutama ketika data bersifat rahasia atau sumber daya komputasi terbatas [15], [19]. Implementasi RAG juga terbukti berhasil di berbagai domain lain, seperti sistem *Biomedical QA*

[18], *Enterprise Knowledge Management* [19], hingga *Automated Document Understanding* untuk laporan PDF [17]. Penelitian-penelitian tersebut memperkuat urgensi penerapan RAG dengan model kecil dalam lingkungan terkontrol, efisien, dan terverifikasi.

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan penerapan *Small Language Model* berbasis *Retrieval-Augmented Generation* (RAG) untuk pemahaman dokumen teknis internal. Studi kasus difokuskan pada *Sistem Manajemen Proyek Terpadu (SMPT)*, dengan tujuan untuk:

1. Mendesain pipeline RAG berbasis SLM yang memanfaatkan dokumen PDF teknis sebagai sumber utama.
2. Menerapkan *vector-based retrieval* menggunakan FAISS untuk pencarian semantik.
3. Mengevaluasi tingkat relevansi dan efisiensi sistem dalam menjawab pertanyaan pengguna berdasarkan konteks dokumen.

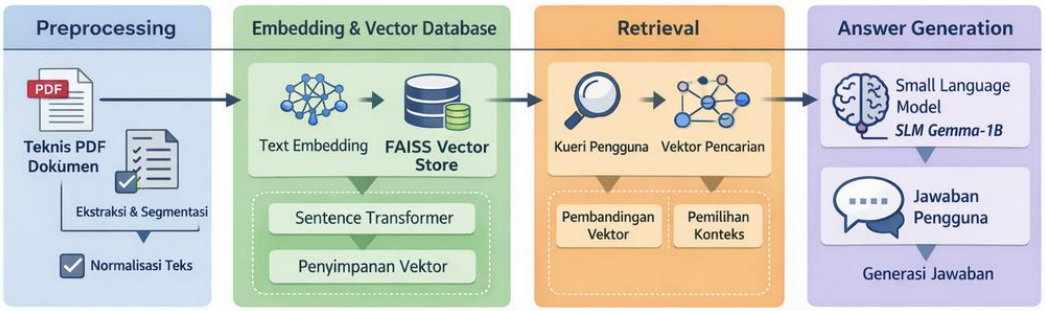
Pendekatan ini diharapkan dapat menjadi solusi efektif dan hemat sumber daya bagi pengembangan sistem tanya jawab internal berbasis dokumen teknis, sejalan dengan tren adopsi model bahasa kecil dan *retrieval system* efisien di berbagai sektor teknologi informasi [12], [14], [20].

2. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimental dengan tujuan mengevaluasi penerapan *Small Language Model* (SLM) *berbasis Retrieval-Augmented Generation* (RAG) pada sistem tanya jawab dokumen teknis. Metode penelitian difokuskan pada perancangan, implementasi, dan pengujian sistem end-to-end yang mampu membaca, memahami, dan menjawab pertanyaan berdasarkan isi dokumen teknis secara terkontrol dan kontekstual..

2.1 Desain Penelitian

Desain penelitian ini menggunakan pendekatan eksperimen rekayasa sistem (*engineering-based experimental approach*) untuk mengembangkan sistem *Small Language Model* (SLM) berbasis *Retrieval-Augmented Generation* (RAG) dalam konteks pemahaman dokumen teknis internal organisasi. Pendekatan ini dipilih karena memungkinkan perancangan sistem yang dapat diuji secara langsung terhadap efektivitas, efisiensi, dan relevansi hasilnya. Sistem yang dibangun terdiri dari empat komponen utama, yaitu tahap pra-pemrosesan dokumen, pembentukan representasi embedding, mekanisme *retrieval* berbasis kesamaan vektor, serta proses generasi jawaban menggunakan model bahasa kecil. Setiap komponen saling terhubung secara modular untuk membentuk satu pipeline terpadu dari input dokumen hingga keluaran jawaban pengguna. Secara konseptual, rancangan alur sistem digambarkan dalam, yang menunjukkan hubungan antar proses mulai dari ekstraksi teks hingga penyusunan jawaban berbasis konteks..



Gambar 1. Desain Metode Penelitian

2.2 Sumber Data Penelitian

Data utama dalam penelitian ini berasal dari dokumen internal berjudul **"Dokumentasi Teknis Sistem Manajemen Proyek Terpadu (SMPT)"**, yang disusun dalam format PDF dan terdiri dari delapan bab, yaitu pendahuluan, arsitektur sistem, peran pengguna, modul dan fungsionalitas, alur kerja proyek, aturan dan batasan sistem, penanganan *error*, serta keterbatasan sistem. Dokumen tersebut dijadikan sebagai korpus dasar untuk sistem RAG, di mana setiap bab diubah menjadi file teks Markdown agar mudah diproses secara semantik. Proses konversi dilakukan dengan menggunakan skrip *Python* pdf_to_md.py yang mengekstrak teks dari file PDF, kemudian memisahkannya berdasarkan struktur bab menggunakan pola ekspresi reguler. Hasil konversi disimpan dalam folder docs/ dengan delapan file Markdown, masing-masing mewakili satu bab dokumen. Struktur ini memudahkan proses embedding dan pencarian konteks sebagaimana juga diterapkan dalam penelitian serupa yang menggunakan korpus domain-spesifik untuk sistem RAG internal.

2.3 Pra-pemrosesan Dokumen

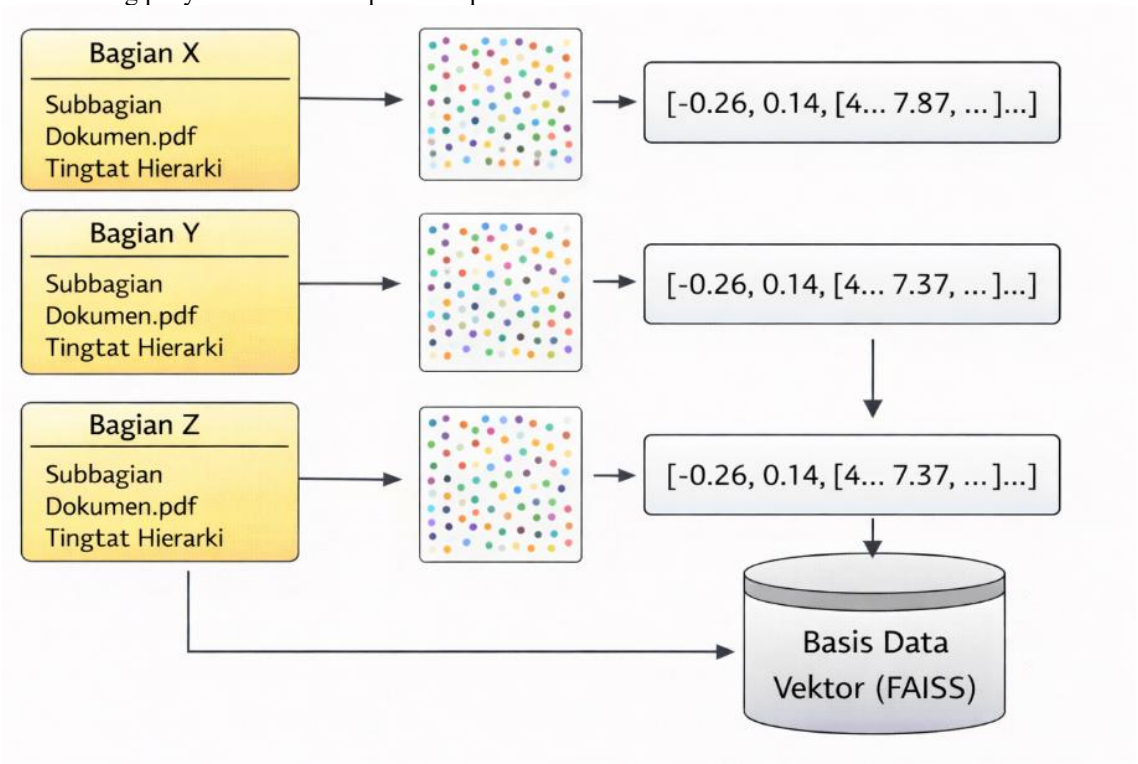
Untuk Tahap pra-pemrosesan bertujuan untuk mengubah dokumen mentah menjadi unit informasi yang terstruktur dan siap digunakan dalam proses retrieval. Tahapan pra-pemrosesan meliputi:

1. **Ekstraksi Teks**
   Dokumen PDF diekstraksi menjadi teks mentah dan dibersihkan dari karakter non-informatif, seperti pemisah halaman dan simbol kontrol.
2. **Pemisahan Berbasis Struktur Dokumen**
   Teks hasil ekstraksi dipisahkan berdasarkan struktur penomoran bagian dokumen, sehingga setiap bagian utama disimpan sebagai dokumen terpisah dalam format Markdown.
3. **Chunking Semantik**
   Setiap dokumen diproses menggunakan pendekatan chunking semantik dengan membagi teks berdasarkan judul bagian dan subbagian. Pendekatan ini bertujuan menjaga koherensi konteks dalam setiap unit informasi.

2.4 Representasi *Embedding* dan Penyimpanan Vektor

Setiap *chunk* hasil pra-pemrosesan direpresentasikan dalam bentuk ***embedding*** **vektor** menggunakan model sentence embedding. Representasi vektor ini memungkinkan sistem melakukan pencarian berbasis kemiripan semantik antara pertanyaan pengguna dan isi dokumen.

Embedding yang dihasilkan disimpan dalam basis data vektor menggunakan FAISS. Selain embedding, setiap chunk juga disertai metadata berupa judul bagian, sumber dokumen, dan tingkat hierarki untuk mendukung penyusunan konteks pada tahap inferensi.



Gambar 2 Representasi *Embedding* dan Penyimpanan Vetor

2.5 Mekanisme Retrieval

Pada tahap inferensi, pertanyaan pengguna diubah menjadi embedding vektor dan digunakan untuk melakukan pencarian terhadap basis data vektor. Sistem mengambil sejumlah *top-k* chunk dengan tingkat kemiripan tertinggi sebagai konteks relevan.

Mekanisme retrieval ini memastikan bahwa hanya informasi yang paling berkaitan dengan pertanyaan pengguna yang digunakan dalam proses generasi jawaban.

2.6 Proses Generasi Jawaban

Konteks hasil retrieval digabungkan dan disusun dalam bentuk prompt terstruktur yang kemudian diberikan kepada *Small Language Model*. Prompt dirancang dengan aturan ketat agar model hanya menghasilkan jawaban berdasarkan konteks yang disediakan. Apabila informasi yang ditanyakan tidak tersedia dalam konteks, sistem diarahkan untuk menyatakan bahwa informasi tersebut tidak ditemukan dalam dokumen..

2.7 Skenario Pengujian

Pengujian sistem dilakukan dengan memberikan serangkaian pertanyaan yang berkaitan langsung dengan isi dokumen teknis, mencakup aspek struktur sistem, peran pengguna, modul fungsionalitas, alur kerja, serta aturan sistem. Jawaban yang dihasilkan dianalisis secara kualitatif untuk menilai kesesuaian jawaban terhadap isi dokumen, konsistensi konteks, dan kejelasan informasi.

Hasil pengujian digunakan untuk mengevaluasi efektivitas penerapan *Small Language Model* berbasis *Retrieval-Augmented Generation* dalam mendukung pemahaman dokumen teknis secara interaktif.

3. Hasil dan Pembahasan

Bagian ini menyajikan hasil pengujian sistem tanya jawab dokumen teknis berbasis *Small Language Model* (SLM) dengan pendekatan *Retrieval-Augmented Generation* (RAG) serta pembahasan terhadap temuan yang diperoleh. Pengujian difokuskan pada kemampuan sistem dalam menghasilkan jawaban yang relevan, konsisten dengan dokumen sumber, dan terbebas dari informasi yang tidak didukung oleh konteks.

3.1 Hasil Penerapan Sistem *Retrieval-Augmented Generation*

Pada tahap *retrieval*, sistem melakukan pencarian konteks menggunakan pendekatan vector similarity berbasis embedding kalimat. Setiap paragraf dokumen direpresentasikan sebagai vektor embedding menggunakan model *SentenceTransformer* all-MiniLM-L6-v2. Pertanyaan pengguna juga diubah ke dalam ruang vektor yang sama, kemudian dihitung jaraknya menggunakan Euclidean distance.

Untuk meningkatkan relevansi hasil pencarian, sistem menerapkan mekanisme routing sederhana berbasis bab dokumen. Pertanyaan pengguna dianalisis menggunakan aturan berbasis kata kunci untuk membatasi ruang pencarian hanya pada bagian dokumen yang sesuai, sehingga proses retrieval tidak dilakukan pada seluruh korpus dokumen. Pendekatan ini terbukti mempersempit ruang pencarian dan meningkatkan kesesuaian konteks yang diambil.

3.2 Kualitas Jawaban yang Dihasilkan oleh *Small Language Model*

Konteks hasil retrieval kemudian digabungkan dan diberikan kepada *Small Language Model* melalui prompt terstruktur. *Prompt* dirancang dengan aturan eksplisit yang membatasi model untuk hanya menjelaskan dan merangkum informasi yang tersedia dalam konteks, serta melarang penambahan pengetahuan di luar dokumen.

Pendekatan *prompt* engineering ini berfungsi sebagai mekanisme kontrol generatif yang mencegah model menghasilkan jawaban spekulatif. Ketika konteks tersedia, model menghasilkan jawaban yang konsisten dengan isi dokumen, sedangkan ketika konteks tidak tersedia, sistem tidak memaksakan generasi jawaban.

3.3 Analisis Pengaruh Retrieval terhadap Hallucination

Berkurangnya kecenderungan hallucination pada sistem disebabkan oleh dua faktor teknis utama. Pertama, mekanisme retrieval memastikan bahwa konteks yang diberikan kepada model berasal langsung dari dokumen sumber, bukan dari memori parametrik model. Kedua, pembatasan prompt secara eksplisit melarang model untuk menambahkan informasi di luar konteks yang disediakan.

Dengan kombinasi retrieval berbasis embedding dan prompt terkontrol, *Small Language Model* yang memiliki kapasitas parameter terbatas tetap mampu menghasilkan jawaban yang akurat dan konsisten dengan sumber dokumen.

3.4 Efektivitas Representasi Embedding dan Basis Data Vektor

Penggunaan *embedding* vektor sebagai representasi semantik dokumen memungkinkan sistem melakukan pencarian informasi secara lebih fleksibel dibandingkan pencarian berbasis kata kunci. Sistem mampu menemukan konteks yang relevan meskipun terdapat perbedaan redaksi antara pertanyaan pengguna dan teks dokumen.

Basis data vektor berbasis FAISS juga menunjukkan performa yang efisien dalam proses pencarian *nearest neighbor*, sehingga waktu respons sistem tetap relatif cepat meskipun jumlah chunk dokumen meningkat. Selain itu, penyertaan metadata pada setiap embedding membantu sistem dalam menyusun konteks yang lebih terstruktur dan mudah dipahami oleh model bahasa.

3.5 Pembahasan Temuan Penelitian

Hasil penelitian ini menunjukkan bahwa pendekatan *Small Language Model* berbasis *Retrieval-Augmented Generation* efektif diterapkan pada sistem tanya jawab dokumen teknis. Integrasi retrieval tidak hanya meningkatkan relevansi jawaban, tetapi juga berfungsi sebagai mekanisme pengendalian generasi bahasa yang penting untuk menjaga keakuratan informasi.

Dibandingkan dengan pendekatan yang sepenuhnya mengandalkan kemampuan generatif model bahasa, sistem yang dikembangkan dalam penelitian ini lebih sesuai untuk lingkungan yang menuntut keandalan informasi, seperti dokumentasi teknis internal. Selain itu, penggunaan *Small Language Model* memberikan keuntungan dari sisi efisiensi komputasi, sehingga sistem dapat dijalankan pada lingkungan dengan sumber daya terbatas tanpa mengorbankan kualitas jawaban secara signifikan.

Hasil penelitian ini menunjukkan bahwa pendekatan teknis berbasis routing dokumen, vector similarity, dan prompt terkontrol berperan penting dalam menjaga kualitas jawaban pada sistem RAG berbasis Small Language Model.

4. Kesimpulan

Berdasarkan Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan *Small Language Model* (SLM) berbasis *Retrieval-Augmented Generation* (RAG**)** efektif digunakan untuk mendukung sistem tanya jawab pada dokumen teknis. Integrasi mekanisme retrieval berbasis embedding vektor memungkinkan sistem memperoleh konteks yang relevan dari dokumen sumber, sehingga jawaban yang dihasilkan oleh model bahasa lebih akurat dan konsisten terhadap informasi yang tersedia.

Hasil pengujian menunjukkan bahwa pendekatan RAG mampu meningkatkan relevansi jawaban serta secara signifikan mengurangi kecenderungan hallucination yang umum terjadi pada sistem berbasis model bahasa generatif tanpa retrieval. Dengan adanya pembatasan konteks yang eksplisit, model bahasa hanya menghasilkan jawaban berdasarkan informasi yang ditemukan dalam dokumen, dan mampu menyatakan ketidaksediaan informasi ketika jawaban tidak didukung oleh sumber data.

Selain itu, penggunaan *Small Language Model* memberikan keuntungan dari sisi efisiensi komputasi, sehingga sistem dapat diimplementasikan pada lingkungan dengan keterbatasan sumber daya tanpa mengorbankan keandalan informasi. Pemanfaatan basis data vektor berbasis FAISS juga terbukti mendukung proses pencarian semantik yang cepat dan efektif terhadap kumpulan dokumen teknis.

Secara keseluruhan, penelitian ini membuktikan bahwa pendekatan SLM berbasis RAG merupakan solusi yang layak dan praktis untuk pemahaman dokumen teknis secara interaktif. Penelitian selanjutnya dapat diarahkan pada evaluasi kuantitatif yang lebih mendalam, pengujian pada berbagai jenis dokumen teknis, serta pengembangan mekanisme interaksi yang lebih adaptif sesuai kebutuhan pengguna.

Daftar Pustaka

[1] H. N. Patel, A. Surti, P. Goel, & B. Patel, “A Comparative Analysis of Large Language Models with Retrieval-Augmented Generation Based Question Answering System,” *IEEE I-SMAC*, 2024.

[2] K. Muludi, K. M. Fitria, & J. Triloka, “Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model,” *Int. J. ICT Research*, 2024.

[3] O. Ali, “Retrieval-Augmented Generation for Intelligent Querying of Databases and Documents,” *Metropolia Univ. of Applied Sciences*, 2025.

[4] V. Reddy & N. Veeranjaneyulu, “An Optimized Content Retriever from Web Articles using LLMs and FAISS Indexing,” *ResearchSquare*, 2025.

[5] R. Shetty, “Enhancing Context-Aware Search with Retrieval-Augmented Generation,” *TechRxiv*, 2025.

[6] M. Mickel, “Development and Optimization of a Retrieval Augmented Generation System for Enhanced Conversational AI Assistance,” *Univ. of Padua*, 2024.

[7] S. Trangcasanchai, “Improving Question Answering Systems with Retrieval Augmented Generation,” *Univ. of Helsinki*, 2024.

[8] R. Rolle, “Retrieval-Augmented Generation Optimizations,” *Theseus Repository*, 2024.

[9] R. Korkee, “Development and Evaluation of Retrieval-Augmented Generation Methods for Document Search and QA,” *Tampere Univ.*, 2025.

[10] L. Alat & H. Hermawan, "Development of Academic QA System Based on RAG Using Local LLM and FAISS Index," *JANAPATI Journal*, 2025.

[11] I. Siragusa, "Towards the Usage of Large Language Models in Information Retrieval and Question Answering," *Univ. of Bologna*, 2025.

[12] S. Hasan & A. Rezai, "LLM Retrieval vs Parametric Memory: A Comparison Using RAGAS Answer Evaluation," *DiVA Portal*, 2025.

[13] F. Bianchini, "Retrieval-Augmented Generation," in *Advances in Information Systems with LLMs*, Springer, 2025.

[14] M. Zeller, "Disaggregated Heterogeneous System for Retrieval-Augmented Language Models," *ETH Zurich*, 2023.

[15] M. A. Kartiyanta & E. Ancilla, "Performance Evaluation for Cost-Effective Retrieval Process in Multi-Document RAG," *IEEE AIAICT*, 2025.

[16] L. Xu, L. Lu, M. Liu, C. Song, & L. Wu, "Nanjing Yunjin Intelligent QA System Based on Knowledge Graphs and RAG," *Heritage Science*, Nature, 2024.

[17] B. Chandra, P. Preethika, S. Challagundla, & Y. Gogireddy, "End-to-End Neural Embedding Pipeline for Large-Scale PDF Retrieval Using Distributed FAISS," *ResearchGate*, 2024.

[18] L. Y. Panchumarthi, S. P. Gudari, & A. Negi, "RAG-BioQA: Retrieval-Augmented Generation for Long-Form Biomedical QA," *arXiv:2510.01612*, 2025.

[19] E. Karakurt & A. Akbulut, "Retrieval-Augmented Generation and LLMs for Enterprise Knowledge Management," *Applied Sciences*, vol. 16, no. 1, p. 368, 2025.

[20] M. Bucur, "Exploring Large Language Models and RAG for Automated Form Filling," *Univ. of Twente*, 2023.