



A Systematic Literature Review of Retrieval-Augmented Generation: Methods, Applications, and Future Research Directions

Muhammad Rizky Hajar¹, Ema Utami², Alva Hendi Muhammad³

¹Magister of Informatics Engineering, Universitas Amikom Yogyakarta, Indonesia

²Doctor of Informatics, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

³Department of Informatics, Postgraduate School, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

¹riskihajar@students.amikom.ac.id*, ²ema.u@amikom.ac.id, ³alva@amikom.ac.id

Abstract

Retrieval-Augmented Generation (RAG) represents a growing research direction in the advancement of large language models (LLMs) by incorporating external information sources into the response generation process. As LLM-based systems are increasingly deployed in information-sensitive domains such as healthcare, education, and law, the demand for responses that are not only fluent but also verifiable and context-aware has become more pronounced. This study conducts a systematic literature review (SLR) of 100 recent publications to examine methodological approaches, application domains, technical challenges, and research contributions related to RAG. The review draws on studies indexed in major academic databases, including IEEE, ACM, and Springer, and applies structured inclusion and exclusion criteria to ensure analytical rigor. The findings reveal a strong emphasis on architectural optimization, particularly in the interaction between retrieval and generation components, alongside widespread adoption in domain-specific contexts. Persistent challenges identified across the literature include limitations in retriever effectiveness, system integration complexity, and the absence of standardized evaluation benchmarks. Overall, this review provides a structured synthesis of current RAG research and highlights directions for future investigation and practical deployment.

Keywords: retrieval-augmented generation, systematic literature review, methods, applications, challenges, future research

Abstrak

Retrieval-Augmented Generation (RAG) menjadi salah satu arah penelitian yang berkembang dalam pemanfaatan model bahasa besar (LLM), dengan memadukan proses pengambilan informasi eksternal ke dalam mekanisme generasi respons. Seiring meningkatnya penggunaan LLM pada berbagai sektor yang memerlukan keandalan informasi, seperti kesehatan, pendidikan, dan hukum, kebutuhan akan sistem yang mampu menghasilkan respons yang akurat, dapat diverifikasi, dan sesuai konteks semakin meningkat. Penelitian ini menyajikan tinjauan literatur sistematis terhadap 100 publikasi terkini yang membahas pengembangan RAG dari sisi metode, domain penerapan, tantangan teknis, serta kontribusi penelitian. Proses kajian dilakukan dengan memanfaatkan basis data ilmiah utama, yaitu IEEE, ACM, dan Springer, melalui penerapan kriteria seleksi yang terstruktur. Hasil sintesis menunjukkan bahwa sebagian besar penelitian berfokus pada optimalisasi interaksi antara komponen retrieval dan generation, serta penerapan RAG pada domain spesifik. Tantangan yang konsisten ditemukan meliputi keterbatasan performa mekanisme retrieval, kompleksitas integrasi sistem, serta belum tersedianya kerangka evaluasi yang terstandarisasi. Kajian ini memberikan pemetaan terstruktur terhadap lanskap riset RAG serta arah pengembangan riset dan implementasi di masa mendatang..

Kata kunci: retrieval-augmented generation, tinjauan literatur sistematis, metode, aplikasi, tantangan, arah riset

1. Pendahuluan

Retrieval-Augmented Generation (RAG) merupakan salah satu pendekatan terkini dalam pengembangan Large Language Models (LLM) yang menggabungkan kemampuan pengambilan informasi (retrieval) dengan model generatif berbasis deep learning. Dengan memanfaatkan sumber eksternal yang relevan selama proses generasi teks, pendekatan ini mampu mengurangi fenomena hallucination dan meningkatkan akurasi serta factuality dari output model.

Metode RAG telah diimplementasikan dalam berbagai domain aplikasi, seperti penjawab pertanyaan berbasis dokumen, asisten medis, pencarian hukum, pendidikan, hingga sistem informasi publik. Popularitasnya terus meningkat seiring dengan kebutuhan akan sistem yang mampu memadukan generasi naratif dan penelusuran informasi yang akurat. Perkembangan penelitian terkait Retrieval-Augmented Generation menunjukkan kemajuan yang signifikan sejak diperkenalkannya pendekatan dasar integrasi retrieval dan generation pada studi awal [1], hingga berbagai aplikasi domain-spesifik



Lisensi

Lisensi Internasional Creative Commons Attribution-ShareAlike 4.0.

yang menunjukkan peningkatan performa substansial dalam akurasi dan relevansi output.

Meskipun penelitian RAG berkembang pesat dengan adopsi lintas berbagai domain, systematic literature reviews yang ada cenderung fokus pada domain spesifik seperti domain-specific knowledge enhancement [2], biomedicine [3], dan healthcare [4]. Belum terdapat tinjauan menyeluruh yang memetakan perkembangan metode, tren riset, tantangan, dan kontribusi dari pendekatan RAG secara cross-domain dan sistematis. Fragmentasi penelitian di berbagai venue publikasi dan kurangnya sintesis komprehensif menciptakan kesenjangan pemahaman tentang universal methodological patterns, best practices, dan arah penelitian masa depan dalam implementasi RAG.

Berbeda dengan existing systematic reviews yang menekankan pada aplikasi domain-spesifik, studi ini mengadopsi cross-domain analytical approach untuk mengidentifikasi pola metodologi universal dalam RAG implementation. Untuk menjawab kekosongan tersebut, penelitian ini melakukan Systematic Literature Review (SLR) berdasarkan pendekatan yang disarankan oleh Kitchenham [5], dengan tujuan memetakan pendekatan dan arsitektur RAG, mengidentifikasi domain aplikasi yang dominan, mengkaji tantangan teknis dan riset terbuka, serta merangkum kontribusi penelitian terhadap pengembangan sistem RAG di masa depan.

2. Metode Penelitian

Studi ini dilakukan menggunakan pendekatan Systematic Literature Review (SLR) sebagaimana dijelaskan oleh Kitchenham dan Charters. SLR dipilih karena kemampuannya dalam mengumpulkan, mengevaluasi, dan mensintesis temuan penelitian secara terstruktur dan dapat direplikasi. Pendekatan ini telah terbukti efektif dalam penelitian serupa, sebagaimana ditunjukkan dalam systematic reviews RAG untuk domain-specific knowledge enhancement [2] dan biomedicine [3].

Berbeda dengan existing reviews yang fokus pada domain spesifik, penelitian ini mengadopsi cross-domain methodology-focused approach dengan menggunakan database computer science dan engineering (IEEE, ACM, Springer) untuk mengcapture perspektif teknis yang lebih luas, dibandingkan dengan database medis (PubMed, Embase, PsycINFO) yang digunakan dalam domain-specific reviews.

Untuk menjaga fokus kajian, pertanyaan penelitian (Research Questions/RQ) dirumuskan seperti pada Tabel 1.

Kriteria inklusi dan eksklusi yang digunakan dalam proses seleksi studi dirangkum pada Tabel 2.

Dengan mengikuti prinsip transparansi dan replikasi dalam studi SLR, proses seleksi dilakukan secara sistematis melalui pencarian di tiga digital library utama:

IEEE Xplore, ACM Digital Library, dan SpringerLink. Kata kunci yang digunakan dalam proses pencarian adalah: ("Retrieval-Augmented Generation" OR "RAG") AND ("language model" OR "generation"). Studi ini diharapkan dapat memberikan kontribusi terhadap pemetaan penelitian RAG sekaligus menjadi acuan bagi pengembangan riset lanjutan di bidang ini.

Tabel 1. Tabel Pertanyaan Penelitian

ID	Pertanyaan Penelitian	Motivasi
RQ1	Bagaimana perkembangan metode dan pendekatan RAG dari waktu ke waktu, serta apa saja arsitektur utama yang digunakan? digunakan dalam pengembangan RAG?	Memetakan evolusi teknologi RAG dan mengidentifikasi teknik serta arsitektur model yang menjadi tren dalam penelitian
RQ2	Dalam domain apa saja pendekatan RAG diterapkan secara signifikan?	Mengetahui penerapan praktis RAG dan area yang paling terdampak oleh pendekatan ini
RQ3	Apa saja tantangan teknis dan riset yang sering diidentifikasi dalam studi RAG?	Memahami keterbatasan dan hambatan utama dalam implementasi RAG
RQ4	Apa kontribusi utama dari penelitian RAG yang telah dipublikasikan sejauh ini?	Merangkum kontribusi ilmiah untuk mendukung arah riset dan aplikasi RAG selanjutnya

Tabel 2. Tabel Kriteria Inklusi dan Eksklusi

Inklusi	Eksklusi
Studi yang membahas RAG atau integrasi retrieval dan generation	Studi yang hanya membahas retrieval atau generative model secara terpisah
Publikasi ilmiah di jurnal atau prosiding internasional (ACM, IEEE, Springer)	Artikel non-ilmiah, editorial, blog, atau white paper
Studi yang menyediakan penjelasan metode, aplikasi, tantangan, atau hasil evaluasi	Studi tanpa informasi teknis yang memadai (misal hanya ide konseptual tanpa implementasi)
Tahun Publikasi antar 2019-2024.	Duplikasi (hanya versi jurnal yang diambil jika tersedia)

3. Hasil dan Pembahasan

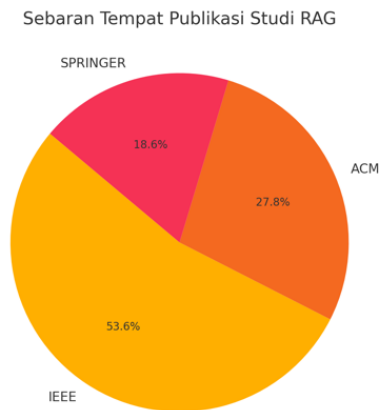
Bagian ini menyajikan hasil analisis sistematis terhadap 97 publikasi primer yang berhasil dikumpulkan berdasarkan kriteria inklusi. Tujuan utama dari bagian ini adalah untuk menjawab keempat pertanyaan penelitian (RQ) yang telah dirumuskan, yaitu mengenai metode yang digunakan, domain aplikasi yang ditargetkan, tantangan utama yang dihadapi, dan kontribusi ilmiah yang diberikan oleh masing-masing studi.

Untuk memperoleh gambaran menyeluruh, publikasi diklasifikasikan dan disintesis dalam bentuk tabel serta visualisasi yang mendukung. Analisis ini dimulai dengan menelaah sebaran publikasi berdasarkan tempat

publikasi ilmiah, dilanjutkan dengan klasifikasi metode dan pendekatan, domain aplikasi yang diadopsi, tantangan-tantangan teknis yang dihadapi, dan kontribusi masing-masing studi.

Distribusi sumber ini mencerminkan kepadatan dan kecenderungan kontribusi ilmiah dari masing-masing penerbit. IEEE mendominasi dengan 52 publikasi, diikuti oleh ACM dengan 27 dan Springer dengan 18. Beberapa studi yang diterbitkan oleh IEEE seperti [6], [7] dan [8] banyak mengeksplorasi pengembangan pipeline, sistem terbuka, serta struktur modular, terutama dalam konteks sistem tanya jawab dan rekomendasi. Sementara itu, publikasi dari ACM seperti [9], [10], dan [11] lebih menekankan pada integrasi retrieval dan generation dalam aplikasi praktis, serta membahas tantangan terkait performa retriever dan interpretabilitas. Springer cenderung menghadirkan pendekatan berbasis eksperimen dan aplikasi.

Gambar 1 menyajikan visualisasi sebaran publikasi primer berdasarkan tempat publikasi (IEEE, ACM, dan Springer).



Gambar 1. Sebaran Tempat Publikasi Berdasarkan Sumber

Gambar ini memperkuat fakta bahwa topik RAG mendapatkan perhatian yang cukup besar di forum-forum publikasi ilmiah ternama. Hal ini sejalan dengan pertumbuhan minat terhadap integrasi retrieval dan generative models dalam pengembangan Large Language Models (LLMs). Pendekatan RAG tidak hanya dibahas dalam konteks model generatif semata, tetapi juga dikaji dari aspek arsitektur sistem, efisiensi retrieval, hingga evaluasi factuality output model. Narasi berikut merinci metode yang digunakan, domain aplikasi, tantangan utama yang dihadapi, serta kontribusi signifikan dari studi-studi tersebut.

3.1. Metode dan Pendekatan yang Digunakan

Retrieval-Augmented Generation (RAG) menunjukkan variasi yang luas, mulai dari integrasi langsung antara retrieval dan generation, hingga pendekatan berbasis pipeline dan kerangka sistem terbuka. Dalam 97 paper primer yang dikaji, sebagian besar mengimplementasikan bentuk eksplisit dari RAG seperti

dalam studi [12], [13], dan [14], yang menggabungkan mekanisme pengambilan informasi eksternal secara real-time dengan kemampuan generatif untuk meningkatkan relevansi dan akurasi jawaban.

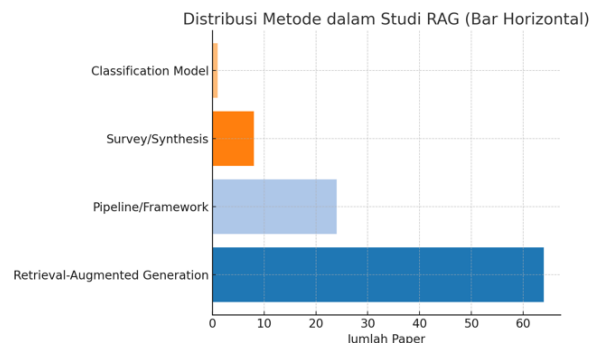
Beberapa studi lain, seperti [9], [15], dan [16], menggunakan integrasi retrieval-generation secara dua tahap tanpa menyebut RAG secara eksplisit, tetapi tetap memenuhi karakteristik arsitektur gabungan. Ada pula pendekatan yang dikembangkan dalam bentuk pipeline modular atau framework eksperimental sebagaimana ditemukan dalam [17], [18], dan [19], yang memfasilitasi eksperimen multi-komponen. Sebagian kecil studi memilih pendekatan klasifikasi terintegrasi seperti pada [20], serta kajian konseptual dan survei menyeluruh seperti [21], [22], dan [23].

Hasil pengelompokan metode secara keseluruhan disajikan pada Tabel 3.

Tabel 3. Tabel Klasifikasi Metode yang Digunakan

No	Metode	Jumlah Paper
1	Classification Model	1
2	Pipeline/Framework	24
3	Retrieval-Augmented Generation	64
4	Survey/Synthesis	8

Gambar 2 menunjukkan proporsi distribusi dari masing-masing metode tersebut dalam bentuk diagram pie. Seluruh paper telah berhasil diklasifikasikan tanpa menyisakan kategori 'Lainnya', hasil dari proses analisis manual yang mendalam terhadap 97 publikasi primer.



Gambar 2. Distribusi Metode dalam Studi RAG

Sebagian besar studi mengimplementasikan Retrieval-Augmented Generation (RAG) secara eksplisit (64 paper), yang menggabungkan komponen retrieval dan generation untuk mengurangi hallucination dan meningkatkan relevansi faktual. Pendekatan ini banyak diadopsi karena fleksibilitasnya dalam mengakses informasi eksternal selama proses generatif, sebagaimana dijelaskan pada studi [12], [13], dan [14].

Kategori Pipeline/Framework (24 paper) mencerminkan pendekatan berbasis sistem modular yang memungkinkan fleksibilitas eksperimen dalam arsitektur RAG, seperti terlihat pada [17], [18], dan [19]. Terdapat pula 8 paper yang termasuk ke dalam kategori Survey/Synthesis, yang memetakan pendekatan dan tren

penelitian RAG, sebagaimana ditunjukkan oleh [21], [22], dan [23]. Terakhir, hanya satu paper, yaitu [20], yang menggunakan pendekatan Classification Model sebagai komponen pendukung dalam pipeline retrieval-generation.

Pendekatan-pendekatan ini mencerminkan keberagaman strategi yang digunakan oleh para peneliti dalam mengembangkan dan mengevaluasi sistem berbasis RAG, baik dari sisi teknis, arsitektural, maupun konseptual.

3.2. Domain Aplikasi RAG

Domain aplikasi menjadi aspek penting dalam mengevaluasi sejauh mana pendekatan Retrieval-Augmented Generation (RAG) telah diterapkan dalam berbagai konteks dunia nyata. Berdasarkan hasil analisis terhadap 97 publikasi primer, ditemukan bahwa pendekatan RAG telah digunakan di berbagai sektor, mencerminkan fleksibilitas arsitektur ini dalam mendukung berbagai kebutuhan informasi dan interaksi.

Dua domain yang paling dominan adalah Sistem Rekomendasi dan Sistem Tanya Jawab, masing-masing dengan 28 publikasi. Ini menunjukkan bahwa pendekatan RAG sangat populer dalam membangun sistem yang memerlukan pencarian informasi akurat dan penyampaian jawaban yang kontekstual, seperti pada studi [24], [18], dan [16]. Dialog Interaktif merupakan domain terbesar berikutnya dengan 14 publikasi, menandakan penerapan yang luas dalam pengembangan chatbot dan sistem percakapan berbasis konteks, seperti yang dibahas dalam [25] dan [26].

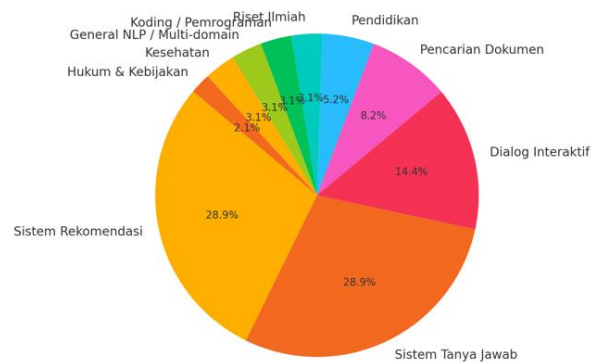
Selanjutnya, RAG juga diadopsi dalam Pencarian Dokumen (8 paper), Pendidikan (5 paper), serta beberapa domain spesifik lain seperti Kesehatan, Riset Ilmiah, Koding / Pemrograman, Hukum & Kebijakan, dan General NLP / Multi-domain — masing-masing dengan 2 hingga 3 publikasi. Hal ini menunjukkan bahwa RAG tidak hanya populer dalam bidang NLP umum, tetapi juga merambah ke sektor-sektor vertikal yang menuntut presisi informasi tinggi.

Distribusi jumlah publikasi pada masing-masing domain aplikasi ditampilkan pada Tabel 4.

Tabel 4. Klasifikasi Domain Aplikasi Studi RAG

No	Domain Aplikasi	Jumlah Paper
1	Sistem Tanya Jawab	28
2	Sistem Rekomendasi	28
3	Dialog Interaktif	14
4	Hukum & Kebijakan	2
5	Pencarian Dokumen	8
6	Pendidikan	5
7	Kesehatan	3
8	Riset Ilmiah	3
9	Koding / Pemrograman	3
10	General NLP / Multi-domain	3

Sebaran distribusi domain aplikasi ditampilkan pada Gambar 3.



Gambar 3. Distribusi Metode dalam Studi RAG

Sebagian besar *paper* berada pada kategori yang konsisten dengan tren dominan seperti Sistem Tanya Jawab, Rekomendasi, dan Dialog Interaktif. Adapun studi-studi dalam bidang kesehatan (misalnya [27], [28], dan [29]) menitikberatkan pada kemampuan RAG dalam memberikan jawaban berbasis sumber literatur medis. Di sektor hukum dan kebijakan ([12], [30], [31]), RAG diterapkan untuk membantu penelusuran dokumen hukum atau menyusun argumen berbasis data peraturan. Domain pendidikan ([32], [33]) memanfaatkan RAG sebagai tutor virtual atau sebagai alat bantu pemahaman berbasis konteks materi ajar. Sementara itu, domain pencarian dokumen tetap menjadi salah satu penerapan paling dominan di luar aplikasi khusus, seperti ditunjukkan oleh [17], [24], dan [14]. Ini menunjukkan bahwa RAG tidak hanya diterapkan untuk keperluan sempit, tetapi sangat fleksibel dan terbuka untuk berbagai jenis implementasi.

3.3. Tantangan yang Dihadapi

Tantangan teknis dalam implementasi Retrieval-Augmented Generation (RAG) merupakan aspek penting yang banyak dibahas dalam publikasi primer. Studi-studi yang dikaji secara konsisten mengungkap sejumlah hambatan, baik dalam sisi integrasi komponen sistem maupun evaluasi performa akhir. Tantangan-tantangan ini tidak hanya bersifat teknis, tetapi juga menyentuh isu keberterimaan sistem, interpretabilitas hasil, serta efisiensi komputasi.

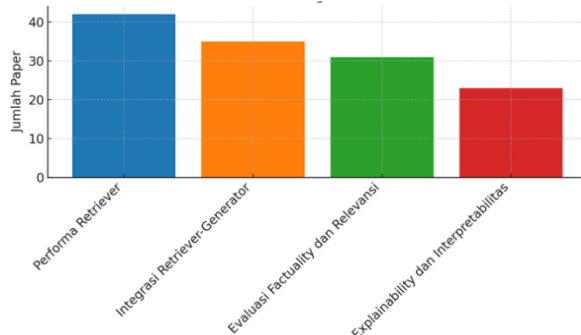
Secara umum, terdapat empat kelompok tantangan utama yang berhasil diidentifikasi dari 97 paper, yaitu: (1) performa retriever yang belum optimal, (2) integrasi retriever-generator yang kompleks dan belum distandarkan, (3) kesulitan dalam mengevaluasi factuality dan relevansi hasil, serta (4) keterbatasan pada aspek explainability dan transparansi sistem. Karena banyak paper membahas lebih dari satu tantangan, maka visualisasi yang ditampilkan tidak bersifat saling eksklusif.

Tabel 5 menyajikan ringkasan jumlah kemunculan masing-masing kategori tantangan berdasarkan hasil klasifikasi terhadap publikasi yang dikaji.

Tabel 5. Ringkasan Tantangan Studi RAG

No	Tantangan	Jumlah Paper
1	Performa Retriever	42
2	Integrasi Retriever-Generator	35
3	Evaluasi Factuality dan Relevansi	31
4	Explainability dan Interpretabilitas	23

Untuk memberikan gambaran visual terhadap penyebaran frekuensi masing-masing tantangan tersebut, Gambar 4 disajikan dalam bentuk histogram.



Gambar 4. Frekuensi Tantangan Studi RAG

Sebagian besar studi (42 dari 97) menyoroti bahwa efektivitas retrieval masih menjadi titik lemah utama dalam pipeline RAG. Ketergantungan pada metode retriever konvensional seperti BM25 dan DPR menyebabkan keterbatasan dalam fleksibilitas dan akurasi hasil. Tantangan integrasi retriever dan generator, sebagaimana dikemukakan dalam studi [16], [17], dan [32], juga menjadi perhatian, terutama dalam desain sistem end-to-end yang efisien.

Sementara itu, isu evaluasi factuality serta kesulitan dalam merancang metrik yang sesuai muncul pada 31 publikasi, termasuk studi [28], [34], dan [35]. Terakhir, 23 studi mengangkat pentingnya aspek explainability, dengan sorotan pada kebutuhan untuk menjelaskan asal-usul output serta meningkatkan kepercayaan pengguna terhadap hasil generatif. Hal ini menjadi krusial dalam aplikasi sensitif seperti hukum dan kesehatan.

3.4. Kontribusi Penelitian

Kontribusi penelitian dalam studi RAG secara umum dapat dikelompokkan ke dalam empat jenis utama. Pertama, kontribusi dalam bentuk pengembangan arsitektur dan teknik baru. Kategori ini mencakup studi yang mengusulkan arsitektur RAG baru, menggabungkan metode retrieval dengan pendekatan lain, atau melakukan optimalisasi pipeline retrieval-generation. Kontribusi ini ditemukan dalam 46 publikasi, seperti [16], [17], dan [36].

Kedua, kontribusi berupa penyediaan dataset dan benchmark evaluasi. Studi dalam kategori ini menyumbangkan dataset baru, merancang metrik evaluasi khusus, atau mengusulkan prosedur validasi yang relevan untuk mengukur performa sistem RAG secara objektif. Kontribusi ini muncul pada 28 paper, termasuk [8], [28], dan [37].

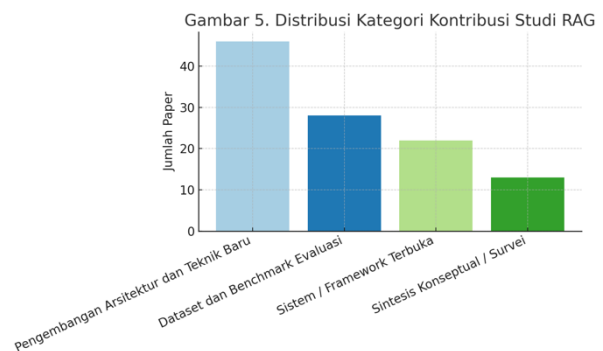
Ketiga, kontribusi dalam bentuk sistem atau framework terbuka. Paper dalam kategori ini menghadirkan implementasi sistem yang dapat diakses publik, seperti open-source tools dan arsitektur siap pakai, untuk memfasilitasi eksperimen lanjutan dan adopsi teknologi RAG. Contohnya ditemukan dalam 22 studi, seperti [11], [24], dan [35].

Terakhir, terdapat pula kontribusi dalam bentuk sintesis konseptual dan survei. Studi-studi dalam kategori ini berfokus pada pemetaan pendekatan, klasifikasi model, serta analisis tren riset dalam pengembangan RAG. Sebanyak 13 publikasi termasuk dalam kategori ini, seperti [21], [22], dan [23].

Tabel 6 merangkum jumlah publikasi berdasarkan masing-masing kategori kontribusi. Sementara itu, Gambar 5 menyajikan visualisasi distribusi kontribusi dalam bentuk grafik batang, untuk memberikan gambaran yang lebih jelas mengenai dominasi kontribusi teknis dalam studi-studi primer.

Tabel 6. Ringkasan Kontribusi dalam Studi RAG

No	Kontribusi	Jumlah Paper
1	Pengembangan Arsitektur dan Teknik Baru	46
2	Dataset dan Benchmark Evaluasi	28
3	Sistem / Framework Terbuka	22
4	Sintesis Konseptual / Survei	13



Gambar 5. Distribusi Kategori Kontribusi RAG

Kontribusi-kontribusi yang telah diidentifikasi dalam 97 publikasi diklasifikasikan ke dalam empat kategori utama, sebagaimana telah dijelaskan sebelumnya. Karena satu studi dapat memberikan lebih dari satu kontribusi, total jumlah kontribusi yang tercatat dalam klasifikasi ini melampaui jumlah publikasi.

Pendekatan ini penting untuk memberikan pemahaman yang menyeluruh mengenai peran setiap studi dalam

mendorong kemajuan penelitian RAG dari berbagai diidentifikasi, serta kontribusi utama dari masing-masing aspek—baik teknis, konseptual, maupun implementatif. masing studi.

Untuk memberikan gambaran yang lebih komprehensif Tabel 7 menyajikan ringkasan publikasi primer yang dan sebagai penutup bagian pembahasan, disusun diklasifikasikan berdasarkan sumber publikasi sebuah matriks yang merekap seluruh elemen penting (Springer, ACM, dan IEEE), serta dianalisis menurut dari 97 publikasi primer yang dikaji. Tabel ini tahun, metode, domain aplikasi, tantangan, dan mengintegrasikan berbagai informasi yang telah kontribusi utama untuk mengidentifikasi tren dan celah dianalisis sebelumnya, termasuk tahun publikasi, riset dalam pengembangan Retrieval-Augmented sumber publikasi, referensi, judul paper, metode yang Generation (RAG). digunakan, domain aplikasi, tantangan yang

Tabel 7. *Paper* yang telah di-review

Tahun	Paper	Publikasi	Metode	Domain	Tantangan	Kontribusi
2022	[12]	SPRINGER	Survey/Synthesis	Sistem Rekomendasi	Performa Retriever	Sintesis Konseptual / Survei
2023	[21]	SPRINGER	Survey/Synthesis	Sistem Tanya Jawab	Explainability dan Interpretabilitas	Sistem / Framework Terbuka, Sintesis Konseptual / Survei
2024	[22]	SPRINGER	Survey/Synthesis	General NLP / Multi-domain	Explainability dan Interpretabilitas	Sintesis Konseptual / Survei
2020	[17]	SPRINGER	Pipeline/Framework	Kesehatan	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru, Sistem / Framework Terbuka
2021	[38]	SPRINGER	Survey/Synthesis	Sistem Rekomendasi	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2022	[27]	SPRINGER	Pipeline/Framework	Sistem Rekomendasi	Performa Retriever	Dataset dan Benchmark Evaluasi
2023	[39]	SPRINGER	Survey/Synthesis	Sistem Rekomendasi	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2024	[40]	SPRINGER	Pipeline/Framework	Sistem Rekomendasi	Integrasi Retriever-Generator, Explainability dan Interpretabilitas	Dataset dan Benchmark Evaluasi
2020	[41]	SPRINGER	Pipeline/Framework	Sistem Rekomendasi	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2021	[15]	SPRINGER	Retrieval-Augmented Generation	Riset Ilmiah	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru
2022	[42]	SPRINGER	Survey/Synthesis	Sistem Tanya Jawab	Evaluasi Factuality dan Relevansi, Explainability dan Interpretabilitas	Dataset dan Benchmark Evaluasi
2023	[43]	SPRINGER	Pipeline/Framework	Sistem Rekomendasi	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2024	[30]	SPRINGER	Survey/Synthesis	Sistem Rekomendasi	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2020	[44]	SPRINGER	Pipeline/Framework	Sistem Tanya Jawab	Performa Retriever	Sistem / Framework Terbuka
2021	[36]	SPRINGER	Retrieval-Augmented Generation	Dialog Interaktif	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2022	[45]	SPRINGER	Pipeline/Framework	Sistem Rekomendasi	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2023	[46]	SPRINGER	Pipeline/Framework	Sistem Rekomendasi	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru, Sistem / Framework Terbuka
2024	[23]	SPRINGER	Retrieval-Augmented Generation	Riset Ilmiah	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi, Sintesis Konseptual / Survei
2020	[32]	ACM	Retrieval-Augmented Generation	Dialog Interaktif	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2021	[47]	ACM	Pipeline/Framework	Dialog Interaktif	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru
2022	[24]	ACM	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator	Sistem / Framework Terbuka

Tabel 7. (Lanjutan)

Tahun	Paper	Publikasi	Metode	Domain	Tantangan	Kontribusi
2023	[48]	ACM	Retrieval-Augmented Generation	Pendidikan	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2024	[9]	ACM	Retrieval-Augmented Generation	Riset Ilmiah	Explainability dan Interpretabilitas	Sistem / Framework Terbuka
2020	[49]	ACM	Retrieval-Augmented Generation	Koding / Pemrograman	Evaluasi Factuality dan Relevansi	Dataset dan Benchmark Evaluasi
2021	[50]	ACM	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru
2022	[51]	ACM	Retrieval-Augmented Generation	Pendidikan	Integrasi Retriever-Generator, Explainability dan Interpretabilitas	Dataset dan Benchmark Evaluasi
2023	[52]	ACM	Retrieval-Augmented Generation	Dialog Interaktif	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2024	[10]	ACM	Retrieval-Augmented Generation	Sistem Rekomendasi	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru
2020	[53]	ACM	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever	Dataset dan Benchmark Evaluasi
2021	[18]	ACM	Retrieval-Augmented Generation	Sistem Rekomendasi	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2022	[13]	ACM	Retrieval-Augmented Generation	Pencarian Dokumen	Evaluasi Factuality dan Relevansi	Sistem / Framework Terbuka
2023	[54]	ACM	Pipeline/Framework	Dialog Interaktif	Performa Retriever, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru
2024	[28]	ACM	Retrieval-Augmented Generation	Pencarian Dokumen	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2020	[33]	ACM	Retrieval-Augmented Generation	Dialog Interaktif	Performa Retriever, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru, Sintesis Konseptual / Survei
2021	[55]	ACM	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2022	[56]	ACM	Retrieval-Augmented Generation	Sistem Rekomendasi	Evaluasi Factuality dan Relevansi	Sistem / Framework Terbuka
2023	[25]	ACM	Retrieval-Augmented Generation	Dialog Interaktif	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru
2024	[57]	ACM	Survey/Synthesis	Dialog Interaktif	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2020	[58]	ACM	Retrieval-Augmented Generation	Pencarian Dokumen	Performa Retriever, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru
2021	[59]	ACM	Retrieval-Augmented Generation	Pencarian Dokumen	Integrasi Retriever-Generator, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru
2022	[60]	ACM	Retrieval-Augmented Generation	Dialog Interaktif	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi, Sintesis Konseptual / Survei
2023	[11]	ACM	Retrieval-Augmented Generation	Hukum & Kebijakan	Performa Retriever, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru, Sistem / Framework Terbuka
2024	[61]	ACM	Pipeline/Framework	Pencarian Dokumen	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2020	[62]	ACM	Retrieval-Augmented Generation	Dialog Interaktif	Performa Retriever, Explainability dan Interpretabilitas	Dataset dan Benchmark Evaluasi

Tabel 7. (Lanjutan)

Tahun	Paper	Publikasi	Metode	Domain	Tantangan	Kontribusi
2021	[31]	ACM	Retrieval-Augmented Generation	Sistem Rekomendasi	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2022	[6]	IEEE	Retrieval-Augmented Generation	Dialog Interaktif	Performa Retriever, Explainability dan Interpretabilitas	Dataset dan Benchmark Evaluasi
2023	[7]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru
2024	[63]	IEEE	Pipeline/Framework	Sistem Tanya Jawab	Performa Retriever	Sistem / Framework Terbuka
2020	[8]	IEEE	Pipeline/Framework	Sistem Rekomendasi	Integrasi Retriever-Generator, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru, Dataset dan Benchmark Evaluasi
2021	[64]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator	Sistem / Framework Terbuka
2022	[16]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru
2023	[65]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi, Sintesis Konseptual / Survei
2024	[66]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever, Evaluasi Factuality dan Relevansi	Sistem / Framework Terbuka
2020	[67]	IEEE	Pipeline/Framework	Pencarian Dokumen	Performa Retriever	Pengembangan Arsitektur dan Teknik Baru
2021	[68]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2022	[34]	IEEE	Retrieval-Augmented Generation	General NLP / Multi-domain	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru, Sintesis Konseptual / Survei
2023	[29]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2024	[69]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Performa Retriever, Evaluasi Factuality dan Relevansi	Dataset dan Benchmark Evaluasi
2020	[70]	IEEE	Pipeline/Framework	Dialog Interaktif	Performa Retriever, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru
2021	[71]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Integrasi Retriever-Generator	Sistem / Framework Terbuka
2022	[72]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2023	[73]	IEEE	Retrieval-Augmented Generation	Kesehatan	Integrasi Retriever-Generator, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru, Sintesis Konseptual / Survei
2024	[74]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever	Dataset dan Benchmark Evaluasi
2020	[75]	IEEE	Pipeline/Framework	Sistem Tanya Jawab	Integrasi Retriever-Generator	Sistem / Framework Terbuka
2021	[76]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2022	[77]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Performa Retriever, Explainability dan Interpretabilitas	Dataset dan Benchmark Evaluasi
2023	[78]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever	Sistem / Framework Terbuka

Tabel 7. (Lanjutan)

Tahun	Paper	Publikasi	Metode	Domain	Tantangan	Kontribusi
2024	[35]	IEEE	Retrieval-Augmented Generation	Koding / Pemrograman	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru, Sistem / Framework Terbuka
2020	[79]	IEEE	Pipeline/Framework	Sistem Rekomendasi	Performa Retriever, Evaluasi Factuality dan Relevansi	Dataset dan Benchmark Evaluasi
2021	[80]	IEEE	Pipeline/Framework	Dialog Interaktif	Performa Retriever, Explainability dan Interpretabilitas	Pengembangan Arsitektur dan Teknik Baru
2022	[81]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Performa Retriever, Explainability dan Interpretabilitas	Dataset dan Benchmark Evaluasi
2023	[82]	IEEE	Retrieval-Augmented Generation	General NLP / Multi-domain	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2024	[83]	IEEE	Retrieval-Augmented Generation	Pendidikan	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2020	[84]	IEEE	Pipeline/Framework	Sistem Rekomendasi	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2021	[85]	IEEE	Pipeline/Framework	Pendidikan	Integrasi Retriever-Generator	Dataset dan Benchmark Evaluasi
2022	[86]	IEEE	Pipeline/Framework	Koding / Pemrograman	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2023	[37]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever, Evaluasi Factuality dan Relevansi	Dataset dan Benchmark Evaluasi
2024	[87]	IEEE	Pipeline/Framework	Sistem Tanya Jawab	Integrasi Retriever-Generator	Sistem / Framework Terbuka
2020	[88]	IEEE	Pipeline/Framework	Pencarian Dokumen	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2021	[89]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2022	[90]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Integrasi Retriever-Generator	Sistem / Framework Terbuka
2023	[19]	IEEE	Pipeline/Framework	Sistem Tanya Jawab	Performa Retriever, Evaluasi Factuality dan Relevansi	Dataset dan Benchmark Evaluasi
2024	[91]	IEEE	Retrieval-Augmented Generation	Hukum & Kebijakan	Integrasi Retriever-Generator	Sistem / Framework Terbuka
2020	[92]	IEEE	Retrieval-Augmented Generation	Dialog Interaktif	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2021	[14]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Evaluasi Factuality dan Relevansi	Dataset dan Benchmark Evaluasi
2022	[93]	IEEE	Retrieval-Augmented Generation	Pendidikan	Explainability dan Interpretabilitas	Sintesis Konseptual / Survei
2023	[94]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2024	[95]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2020	[96]	IEEE	Retrieval-Augmented Generation	Kesehatan	Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2021	[26]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Explainability dan Interpretabilitas	Sintesis Konseptual / Survei
2022	[97]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Evaluasi Factuality dan Relevansi	Sistem / Framework Terbuka

Tabel 7. (Lanjutan)

Tahun	Paper	Publikasi	Metode	Domain	Tantangan	Kontribusi
2023	[98]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Performa Retriever, Evaluasi Factuality dan Relevansi	Pengembangan Arsitektur dan Teknik Baru
2024	[99]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Explainability dan Interpretabilitas	Sistem / Framework Terbuka
2020	[100]	IEEE	Retrieval-Augmented Generation	Sistem Tanya Jawab	Integrasi Retriever-Generator	Pengembangan Arsitektur dan Teknik Baru
2021	[20]	IEEE	Classification Model	Pencarian Dokumen	Integrasi Retriever-Generator, Performa Retriever	Sintesis Konseptual / Survei
2022	[101]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Explainability dan Interpretabilitas	Sintesis Konseptual / Survei
2023	[102]	IEEE	Retrieval-Augmented Generation	Sistem Rekomendasi	Performa Retriever, Evaluasi Factuality dan Relevansi	Sistem / Framework Terbuka

4. Kesimpulan

Penelitian ini menyajikan tinjauan sistematis terhadap 97 publikasi primer yang membahas pendekatan Retrieval-Augmented Generation (RAG), dengan tujuan untuk memetakan metode yang digunakan, domain aplikasi, tantangan utama, serta kontribusi ilmiah dari masing-masing studi. Berdasarkan hasil kajian terhadap lima research questions yang dirumuskan, sejumlah temuan utama dapat disimpulkan sebagai berikut:

Sebagian besar studi (66% atau 64 dari 97 paper) mengimplementasikan RAG secara eksplisit, menunjukkan dominasi pendekatan yang mengintegrasikan retrieval dan generation secara langsung. Pipeline/Framework menjadi pendekatan kedua terbanyak (25%), mencerminkan kebutuhan akan sistem modular dan fleksibel. Survey/Synthesis (8%) berperan penting dalam membangun fondasi teoritis, sementara Classification Model (1%) menunjukkan eksplorasi pendekatan alternatif.

RAG menunjukkan adopsi yang merata di dua domain utama: Sistem Tanya Jawab dan Sistem Rekomendasi (masing-masing 28 paper atau 29%). Dialog Interaktif menempati posisi ketiga (14%), diikuti oleh Pencarian Dokumen (8%) dan Pendidikan (5%). Domain spesifik seperti Kesehatan, Riset Ilmiah, Koding/Pemrograman, dan Hukum & Kebijakan menunjukkan adopsi yang konsisten namun terbatas (2-3 paper per domain), mengindikasikan potensi eksplorasi lebih lanjut.

Analisis mengidentifikasi empat kategori tantangan utama: (1) Performa Retriever yang belum optimal (43% paper), (2) Integrasi Retriever-Generator yang kompleks (36%), (3) Evaluasi Factuality dan Relevansi yang sulit (32%), dan (4) Explainability dan Interpretabilitas yang terbatas (24%). Dominasi masalah performa retriever menunjukkan bahwa komponen retrieval masih menjadi bottleneck utama dalam pipeline RAG.

Kontribusi penelitian didominasi oleh Pengembangan Arsitektur dan Teknik Baru (47% paper), mencerminkan

fase eksploratif dalam penelitian RAG. Dataset dan Benchmark Evaluasi (29%) menunjukkan upaya komunitas untuk standarisasi evaluasi. Sistem/Framework Terbuka (23%) mengindikasikan komitmen terhadap reproducibility dan open science. Sintesis Konseptual/Survei (13%) berperan dalam konsolidasi pengetahuan.

Distribusi temporal publikasi menunjukkan pertumbuhan konsisten dari 2020-2024, dengan puncak aktivitas penelitian pada 2023-2024. IEEE mendominasi publikasi (54%), diikuti ACM (28%) dan Springer (18%), mencerminkan fokus teknis dan aplikatif penelitian RAG. Evolusi menunjukkan transisi dari proof-of-concept awal menuju implementasi praktis dan evaluasi sistematis.

Berbeda dengan systematic reviews domain-spesifik seperti domain-specific enhancement dan biomedicine, studi ini mengungkap universal methodological patterns yang berlaku lintas domain. Temuan cross-domain ini melengkapi dan memperluas understanding dari existing domain-specific reviews dengan memberikan perspektif metodologi yang lebih luas.

Temuan ini menggarisbawahi bahwa RAG telah matang dari pendekatan eksperimental menjadi metodologi yang dapat diandalkan. Namun, tantangan fundamental dalam performa retrieval, integrasi komponen, dan evaluasi factuality masih memerlukan perhatian prioritas. Penelitian masa depan perlu fokus pada: (1) pengembangan retriever yang lebih akurat dan efisien, (2) standarisasi metrik evaluasi factuality, (3) peningkatan explainability sistem RAG, dan (4) eksplorasi domain aplikasi yang belum optimal seperti kesehatan dan hukum.

Secara keseluruhan, studi ini memberikan pemetaan komprehensif landscape penelitian RAG dan menjadi referensi penting bagi pengembangan riset lanjutan dalam bidang ini.

Daftar Rujukan

- [1] P. S. H. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv Comput. Lang.*, 2020, [Online]. Available: <https://arxiv.org/pdf/2005.11401>
- [2] M. Murtiyoso, I. Tahyudin, and B. Berilana, "A Systematic Review of Retrieval-Augmented Generation for Enhancing Domain-Specific Knowledge in Large Language Models," *Sinkron*, vol. 9, no. 2, pp. 969–977, 2025, doi: 10.33395/sinkron.v9i2.14824.
- [3] S. Liu, A. B. McCoy, and A. Wright, "Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines," *J. Am. Med. Informatics Assoc.*, vol. 32, no. 4, pp. 605–615, Apr. 2025, doi: 10.1093/jamia/ocaf008.
- [4] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digit. Heal.*, vol. 4, no. 6, p. e0000877, Jun. 2025, doi: 10.1371/journal.pdig.0000877.
- [5] B. Kitchenham, "Procedures for Performing Systematic Reviews," Keele University, 2004.
- [6] R. Pal, S. Kar, and D. K. Prasad, "A Self-Supervised Generative Fine-Tune of CLIP for VQA for Visually Impaired," in *2024 IEEE 21st India Council International Conference (INDICON)*, IEEE, 2024, pp. 1–6, doi: 10.1109/INDICON63790.2024.10958384.
- [7] C.-H. Chou and C.-H. Wu, "A Translative Research Assistant on Retrieval-Augmented Generation Enhanced by Keyword Extraction in the Case of Asian Studies," in *2024 Artificial Intelligence x Humanities, Education, and Art (AIxHEART)*, IEEE, 2024, pp. 76–83, doi: 10.1109/AIxHeart62327.2024.00025.
- [8] M. Zhang, Y. Shen, J. Yin, S. Lu, and X. Wang, "ADAGENT: Anomaly Detection Agent With Multimodal Large Models in Adverse Environments," *IEEE Access*, vol. 12, pp. 172061–172074, 2024, doi: 10.1109/ACCESS.2024.3480250.
- [9] J. C. Dos Santos Junior, R. Hu, R. Song, and Y. Bai, "Domain-Driven LLM Development: Insights into RAG and Fine-Tuning Practices," in *KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, 2024, pp. 6416–6417, doi: 10.1145/3637528.3671445.
- [10] S. Wang, E. Khramtsova, S. Zhuang, and G. Zuccon, "FeB4RAG: Evaluating Federated Search in the Context of Retrieval Augmented Generation," in *SIGIR 2024: The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2024, pp. 763–773, doi: 10.1145/3626772.3657853.
- [11] S. Yang *et al.*, "Retrieval-Augmented Generation with Quantized Large Language Models: A Comparative Analysis," *ACM Int. Conf. Proceeding Ser.*, pp. 120–124, Nov. 2023, doi: 10.1145/3653081.3653102;WGROU:STRING:ACM.
- [12] G. Hill, M. Waddington, and L. Qiu, "From pen to algorithm: optimizing legislation for the future with artificial intelligence," *AI Soc.*, 2024, doi: 10.1007/s00146-024-02062-3.
- [13] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, "HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction," in *ICAIF '24: 5th ACM International Conference on AI in Finance*, ACM, 2024, pp. 608–616, doi: 10.1145/3677052.3698671.
- [14] B. Zhan, A. Li, X. Yang, D. He, Y. Duan, and S. Yan, "RARoK: Retrieval-Augmented Reasoning on Knowledge for Medical Question Answering," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2024, pp. 2837–2843, doi: 10.1109/BIBM62325.2024.10822341.
- [15] S. Toro *et al.*, "Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI)," *J. Biomed. Semantics*, vol. 15, no. 1, p. 19, 2024, doi: 10.1186/s13326-024-00320-3.
- [16] M. Mudassar Yamin, E. Hashmi, M. Ullah, and B. Katt, "Applications of LLMs for Generating Cyber Security Exercise Scenarios," *IEEE Access*, vol. 12, pp. 143806–143822, 2024, doi: 10.1109/ACCESS.2024.3468914.
- [17] S. Sen, V. Gonzalez, E. J. Husom, S. Tverdal, S. Tokas, and S. O. Tjøsøvoll, "ERG-AI: enhancing occupational ergonomics with uncertainty-aware ML and LLM feedback," *Appl. Intell.*, vol. 54, no. 23, pp. 12128–12155, 2024, doi: 10.1007/s10489-024-05796-1.
- [18] U. Maes, L. Michiels, and A. Smets, "GenUI(ne) CRS: UI Elements and Retrieval-Augmented Generation in Conversational Recommender Systems with LLMs," in *RecSys '24: 18th ACM Conference on Recommender Systems*, ACM, 2024, pp. 1177–1179, doi: 10.1145/3640457.3691697.
- [19] Q. Wang, L.-I. Wu, and G. Li, "PIRTRE-C: A Two-Stage Retrieval and Reranking Enhanced Framework for Improving Chinese Psychological Counseling," in *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, IEEE, 2024, pp. 337–347, doi: 10.1109/MedAI62885.2024.00051.
- [20] H. Xiong *et al.*, "When Search Engine Services Meet Large Language Models: Visions and Challenges," *IEEE Trans. Serv. Comput.*, vol. 17, no. 6, pp. 4558–4577, Nov. 2024, doi: 10.1109/TSC.2024.3451185.
- [21] E. J. Robinson *et al.*, "Physician vs. AI-generated messages in urology: evaluation of accuracy, completeness, and preference by patients and physicians," *World J. Urol.*, vol. 43, no. 1, p. 48, 2024, doi: 10.1007/s00345-024-05399-y.
- [22] J. Schneider, "Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda," *Artif. Intell. Rev.*, vol. 57, no. 11, p. 289, 2024, doi: 10.1007/s10462-024-10916-x.
- [23] H. Xu, F. Omiaomou, S. Sabri, S. Zlatanova, X. Li, and Y. Song, "Leveraging generative AI for urban digital twins: a scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement," *Urban Informatics*, vol. 3, no. 1, p. 29, 2024, doi: 10.1007/s44212-024-00060-w.
- [24] J. Jang and W.-S. Li, "AU-RAG: Agent-based Universal Retrieval Augmented Generation," in *SIGIR-AP 2024: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, ACM, 2024, pp. 2–11, doi: 10.1145/3673791.3698416.
- [25] S. Amri, S. Bani, and R. Bani, "Moroccan legal assistant Enhanced by Retrieval-Augmented Generation Technology," in *NISS 2024: The 7th International Conference On Networking, Intelligents Systems and Security*, ACM, 2024, pp. 1–5, doi: 10.1145/3659677.3659737.
- [26] A. T. Strand, S. Gautam, C. Midoglu, and P. Halvorsen, "SoccerRAG: Multimodal Soccer Information Retrieval via Natural Queries," in *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, 2024, pp. 1–7, doi: 10.1109/CBMI62980.2024.10859209.
- [27] D. Malerba and V. Pasquadibisceglie, "Data-Centric AI," *J. Intell. Inf. Syst.*, vol. 62, no. 6, pp. 1493–1502, 2024, doi: 10.1007/s10844-024-00901-9.
- [28] M. A. Arefeen, B. Debnath, M. Y. S. Uddin, and S. Chakradhar, "iRAG: Advancing RAG for Videos with an Incremental Approach," in *CIKM '24: The 33rd ACM International Conference on Information and Knowledge Management*, ACM, 2024, pp. 4341–4348, doi: 10.1145/3627673.3680088.
- [29] J. I. Ahad *et al.*, "Empowering Meta-Analysis: Leveraging Large Language Models for Scientific Synthesis," in *2024 IEEE International Conference on Big Data (BigData)*, IEEE, 2024, pp. 1615–1624, doi: 10.1109/BigData62323.2024.10825310.
- [30] M. Ding, S. Dong, and R. Grewal, "Generative AI and Usage in Marketing Classroom," *Cust. Needs Solut.*, vol. 11, no. 1, p. 5, 2024, doi: 10.1007/s40547-024-00145-2.
- [31] S. Han, H. Kang, B. Jin, X.-Y. Liu, and S. Y. Yang, "XBRL Agent: Leveraging Large Language Models for Financial Report Analysis," in *ICAIF '24: 5th ACM International Conference on AI in Finance*, ACM, 2024, pp. 856–864, doi: 10.1145/3677052.3698614.
- [32] B. Peng, X. Wang, and L. Xu, "'TILSE' Framework for RAG-Based AIGC Feedback Prompts: A Modular and Personalized Intelligent Feedback Generation Method," *ACM Int. Conf.*

- Proceeding Ser.*, pp. 398–403, Nov. 2024, doi: 10.1145/3700297.3700365;CSUBTYPE:STRING:CONFEREN CE.
- [33] H. Chandrasekar, S. Gupta, C.-T. Liu, and C.-H. Tsai, “Leveraging Large Language Models for Effective Organizational Navigation,” in *dg.o 2024: 25th Annual International Conference on Digital Government Research*, ACM, 2024, pp. 1020–1022. doi: 10.1145/3657054.3657272.
- [34] R. C. Barron *et al.*, “Domain-Specific Retrieval-Augmented Generation Using Vector Stores, Knowledge Graphs, and Tensor Factorization,” in *2024 International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2024, pp. 1669–1676. doi: 10.1109/ICMLA61862.2024.00258.
- [35] C. Arora, T. Herda, and V. Himm, “Generating Test Scenarios from NL Requirements Using Retrieval-Augmented LLMs: An Industrial Study,” *2024 IEEE 32nd Int. Requir. Eng. Conf.*, pp. 240–251, Jun. 2024, doi: 10.1109/RE59067.2024.00031.
- [36] K. Rangan and Y. Yin, “A fine-tuning enhanced RAG system with quantized influence measure as AI judge,” *Sci. Rep.*, vol. 14, no. 1, p. 27446, 2024, doi: 10.1038/s41598-024-79110-x.
- [37] X. Liu, Y. Zhu, T. Pang, K. Xue, X. Zhang, and C. Fan, “Medical Document Embedding Enhancement with Heterogeneous Mixture-of-Experts,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2024, pp. 2238–2243. doi: 10.1109/BIBM62325.2024.10822374.
- [38] D. Leuthe, T. Meyer-Hollatz, T. Plank, and A. Senkmüller, “Towards Sustainability of AI – Identifying Design Patterns for Sustainable Machine Learning Development,” *Inf. Syst. Front.*, vol. 26, no. 6, pp. 2103–2145, 2024, doi: 10.1007/s10796-024-10526-6.
- [39] C. Novelli, G. Formisano, P. Juneja, G. Sandri, and L. Floridi, “Artificial Intelligence for the Internal Democracy of Political Parties,” *Minds Mach.*, vol. 34, no. 4, p. 36, 2024, doi: 10.1007/s11023-024-09693-x.
- [40] D. Grewal, C. B. Saturnino, T. Davenport, and A. Guha, “How generative AI is shaping the future of marketing,” *J. Acad. Mark. Sci.*, 2024, doi: 10.1007/s11747-024-01064-3.
- [41] S. Beck, M. Kuhner, M. Haar, A. Daubmann, M. Semmann, and S. Kluge, “Evaluating the accuracy and reliability of AI chatbots in disseminating the content of current resuscitation guidelines: a comparative analysis between the ERC 2021 guidelines and both ChatGPTs 3.5 and 4,” *Scand. J. Trauma. Resusc. Emerg. Med.*, vol. 32, no. 1, p. 95, Sep. 2024, doi: 10.1186/s13049-024-01266-2.
- [42] C. C. Y. Gue *et al.*, “Evaluating the OpenAI’s GPT-3.5 Turbo’s performance in extracting information from scientific articles on diabetic retinopathy,” *Syst. Rev.*, vol. 13, no. 1, p. 135, 2024, doi: 10.1186/s13643-024-02523-2.
- [43] J. Zhang, R. Xiang, Z. Kuang, B. Wang, and Y. Li, “ArchGPT: harnessing large language models for supporting renovation and conservation of traditional architectural heritage,” *Herit. Sci.*, vol. 12, no. 1, p. 220, 2024, doi: 10.1186/s40494-024-01334-x.
- [44] X. Luo *et al.*, “Large language models surpass human experts in predicting neuroscience results,” *Nat. Hum. Behav.*, vol. 9, no. 2, pp. 305–315, 2024, doi: 10.1038/s41562-024-02046-9.
- [45] A. Pahud De Mortanges *et al.*, “Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging,” *npj Digit. Med.*, vol. 7, no. 1, p. 195, 2024, doi: 10.1038/s41746-024-01190-w.
- [46] C. Silvestri *et al.*, “Evaluation of a novel large language model (LLM)-powered chatbot for oral boards scenarios,” *Glob. Surg. Educ. - J. Assoc. Surg. Educ.*, vol. 3, no. 1, p. 112, 2024, doi: 10.1007/s44186-024-00303-z.
- [47] Y. Hu *et al.*, “APTESS: Incorporating Appraisal Theory and Emotion Support Strategies for Empathetic Response Generation,” in *CIKM ’24: The 33rd ACM International Conference on Information and Knowledge Management*, ACM, 2024, pp. 900–909. doi: 10.1145/3627673.3679687.
- [48] J. Dou and X. Zhao, “Design and Application of Online Teaching Resource Platform for College English Based on Retrieval-Augmented Generation,” in *EKI 2024: The 2nd International Conference on Educational Knowledge and Informatization*, ACM, 2024, pp. 111–115. doi: 10.1145/3691720.3691739.
- [49] X. Yu, C. Li, M. Pan, and X. Li, “DroidCoder: Enhanced Android Code Completion with Context-Enriched Retrieval-Augmented Generation,” in *ASE ’24: 39th IEEE/ACM International Conference on Automated Software Engineering*, ACM, 2024, pp. 681–693. doi: 10.1145/3691620.3695063.
- [50] J. Guo, J. Huang, Z. Zhao, J. Chen, Y. Weng, and G. Lin, “EFRAG: Embedding-FineTuning Retrieval Augmented Generation for QA in power projects,” in *PRIS 2024: 2024 6th International Conference on Pattern Recognition and Intelligent Systems*, ACM, 2024, pp. 21–27. doi: 10.1145/3689218.3689225.
- [51] Y. Zhu *et al.*, “EMERGE: Enhancing Multimodal Electronic Health Records Predictive Modeling with Retrieval-Augmented Generation,” in *CIKM ’24: The 33rd ACM International Conference on Information and Knowledge Management*, ACM, 2024, pp. 3549–3559. doi: 10.1145/3627673.3679582.
- [52] Y. Xu, T. Cai, J. Jiang, and X. Song, “Face4Rag: Factual Consistency Evaluation for Retrieval Augmented Generation in Chinese,” in *KDD ’24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, 2024, pp. 6083–6094. doi: 10.1145/3637528.3671656.
- [53] T. Cai *et al.*, “FoRAG: Factuality-optimized Retrieval Augmented Generation for Web-enhanced Long-form Question Answering,” in *KDD ’24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, 2024, pp. 199–210. doi: 10.1145/3637528.3672065.
- [54] Z. Zhou *et al.*, “Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation,” in *SIGIR 2024: The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2024, pp. 2749–2754. doi: 10.1145/3626772.3657673.
- [55] X. Bai, X. Wu, I. Stojkovic, and K. Tsioutsoulis, “Leveraging Large Language Models for Improving Keyphrase Generation for Contextual Targeting,” in *CIKM ’24: The 33rd ACM International Conference on Information and Knowledge Management*, ACM, 2024, pp. 4349–4357. doi: 10.1145/3627673.3680093.
- [56] S. Xu, L. Pang, J. Xu, H. Shen, and X. Cheng, “List-aware Reranking-Truncation Joint Model for Search and Retrieval-augmented Generation,” in *WWW ’24: The ACM Web Conference 2024*, ACM, 2024, pp. 1330–1340. doi: 10.1145/3589334.3645336.
- [57] J. Fang, N. Arechiga, K. Namikoshi, N. Bravo, C. Hogan, and D. A. Shamma, “On LLM Wizards: Identifying Large Language Models’ Behaviors for Wizard of Oz Experiments,” in *IVA ’24: ACM International Conference on Intelligent Virtual Agents*, ACM, 2024, pp. 1–11. doi: 10.1145/3652988.3673967.
- [58] W. Xu, M. Wang, W. Zhou, and H. Li, “P-RAG: Progressive Retrieval Augmented Generation For Planning on Embodied Everyday Task,” in *MM ’24: The 32nd ACM International Conference on Multimedia*, ACM, 2024, pp. 6969–6978. doi: 10.1145/3664647.3680661.
- [59] Z. Hu, C. Wang, Y. Shu, H.-Y. Paik, and L. Zhu, “Prompt Perturbation in Retrieval-Augmented Generation based Large Language Models,” in *KDD ’24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, 2024, pp. 1119–1130. doi: 10.1145/3637528.3671932.
- [60] J. Yu *et al.*, “RAG-Guided Large Language Models for Visual Spatial Description with Adaptive Hallucination Corrector,” in *MM ’24: The 32nd ACM International Conference on Multimedia*, ACM, 2024, pp. 11407–11413. doi: 10.1145/3664647.3688990.
- [61] Y. Shi, Q. Tan, X. Wu, S. Zhong, K. Zhou, and N. Liu, “Retrieval-enhanced Knowledge Editing in Language Models for Multi-Hop Question Answering,” in *CIKM ’24: The 33rd ACM International Conference on Information and Knowledge Management*, ACM, 2024, pp. 2056–2066. doi: 10.1145/3627673.3679722.
- [62] F. Wu *et al.*, “Time-Sensitive Retrieval-Augmented Generation for Question Answering,” in *CIKM ’24: The 33rd ACM International Conference on Information and Knowledge*

- Management*, ACM, 2024, pp. 2544–2553. doi: 10.1145/3627673.3679800.
- [63] A. Barghi, “Accelerating Graph Query Languages for Machine Learning and Retrieval Augmented Generation,” in *2024 IEEE International Conference on Big Data (BigData)*, IEEE, 2024, pp. 3367–3374. doi: 10.1109/BigData62323.2024.10825641.
- [64] R. N. Barros, K. K. Arguello, and J. Wehrmann, “Anchor Your Embeddings Through the Storm: Mitigating Instance-to-Document Semantic Gap,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10650518.
- [65] M. Fateen, B. Wang, and T. Mine, “Beyond Scores: A Modular RAG-Based System for Automatic Short Answer Scoring with Feedback,” *IEEE Access*, p. 1, 2024, doi: 10.1109/ACCESS.2024.3508747.
- [66] C. Samarajeewa, D. De Silva, E. Osipov, D. Alahakoon, and M. Manic, “Causal Reasoning in Large Language Models using Causal Graph Retrieval Augmented Generation,” in *2024 16th International Conference on Human System Interaction (HSI)*, IEEE, 2024, pp. 1–6. doi: 10.1109/HSI61632.2024.10613566.
- [67] Z. Zhu *et al.*, “Conversational Crowdsensing in the Age of Industry 5.0: A Parallel Intelligence and Large Models Powered Novel Sensing Approach,” *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 6, pp. 8046–8063, 2024, doi: 10.1109/TCSS.2024.3451649.
- [68] A. Golatkar, A. Achille, L. Zancato, Y.-X. Wang, A. Swaminathan, and S. Soatto, “CPR: Retrieval Augmented Generation for Copyright Protection,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2024, pp. 12374–12384. doi: 10.1109/CVPR52733.2024.01176.
- [69] A. Darji, F. Kheni, D. Chodvadia, P. Goel, D. Garg, and B. Patel, “Enhancing Financial Risk Analysis using RAG-based Large Language Models,” in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, 2024, pp. 754–760. doi: 10.1109/ICACRS62842.2024.10841711.
- [70] J. Martínez, B. Llinas, J. G. Botello, J. J. Padilla, and E. Frydenlund, “Enhancing GPT-3.5’s Proficiency in Netlogo Through Few-Shot Prompting and Retrieval-Augmented Generation,” in *2024 Winter Simulation Conference (WSC)*, IEEE, 2024, pp. 666–677. doi: 10.1109/WSC63780.2024.10838967.
- [71] Z. Zhu *et al.*, “Enhancing Large Language Models with Knowledge Graphs for Robust Question Answering,” in *2024 IEEE 30th International Conference on Parallel and Distributed Systems (ICPADS)*, IEEE, 2024, pp. 262–269. doi: 10.1109/ICPADS63350.2024.00042.
- [72] Z. Jiang, Z. Shi, X. Gao, and Y. Xiong, “Enhancing Code Generation Through Retrieval of Cross-Lingual Semantic Graphs,” in *2024 31st Asia-Pacific Software Engineering Conference (APSEC)*, IEEE, 2024, pp. 151–160. doi: 10.1109/APSEC65559.2024.00026.
- [73] M. Shayaninasab, M. Zahoor, and Ö. N. Yalçın, “Enhancing Patient Intake Process in Mental Health Consultations Using RAG-Driven Chatbot,” in *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2024, pp. 256–264. doi: 10.1109/ACIIW63320.2024.00053.
- [74] D. Hamzic, M. Wurzenberger, F. Skopik, M. Landauer, and A. Rauber, “Evaluation and Comparison of Open-Source LLMs Using Natural Language Generation Quality Metrics,” in *2024 IEEE International Conference on Big Data (BigData)*, IEEE, 2024, pp. 5342–5351. doi: 10.1109/BigData62323.2024.10825576.
- [75] M. V. Nezafat and S. Samet, “Fake News Detection with Retrieval Augmented Generative Artificial Intelligence,” in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, IEEE, 2024, pp. 160–167. doi: 10.1109/FLLM63129.2024.10852474.
- [76] H. Zeng, Z. Yue, Q. Jiang, and D. Wang, “Federated Recommendation via Hybrid Retrieval Augmented Generation,” in *2024 IEEE International Conference on Big Data (BigData)*, IEEE, 2024, pp. 8078–8087. doi: 10.1109/BigData62323.2024.10825302.
- [77] S. Neupane *et al.*, “From Questions to Insightful Answers: Building an Informed Chatbot for University Resources,” in *2024 IEEE Frontiers in Education Conference (FIE)*, IEEE, 2024, pp. 1–9. doi: 10.1109/FIE61694.2024.10892994.
- [78] B. H. Rappazzo, Y. Wang, A. Ferber, and C. Gomes, “GEM-RAG: Graphical Eigen Memories for Retrieval Augmented Generation,” in *2024 International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2024, pp. 1259–1264. doi: 10.1109/ICMLA61862.2024.00196.
- [79] Y. Xia, Z. Xiao, N. Jazdi, and M. Weyrich, “Generation of Asset Administration Shell With Large Language Model Agents: Toward Semantic Interoperability in Digital Twins in the Context of Industry 4.0,” *IEEE Access*, vol. 12, pp. 84863–84877, 2024, doi: 10.1109/ACCESS.2024.3415470.
- [80] R. Zhang *et al.*, “Generative AI Agents With Large Language Model for Satellite Networks via a Mixture of Experts Transmission,” *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3581–3596, 2024, doi: 10.1109/JSAC.2024.3459037.
- [81] L.-B. Hernandez-Salinas *et al.*, “IDAS: Intelligent Driving Assistance System Using RAG,” *IEEE Open J. Veh. Technol.*, vol. 5, pp. 1139–1165, 2024, doi: 10.1109/OJVT.2024.3447449.
- [82] A. Abdulnazar, R. Roller, S. Schulz, and M. Kreuzthaler, “Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization,” *IEEE Access*, vol. 12, pp. 147981–147990, 2024, doi: 10.1109/ACCESS.2024.3472500.
- [83] W. Zhang *et al.*, “Leveraging RAG-Enhanced Large Language Model for Semi-Supervised Log Anomaly Detection,” in *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*, IEEE, 2024, pp. 168–179. doi: 10.1109/ISSRE62328.2024.00026.
- [84] A. Rachha and M. Seyam, “LLM-Enhanced Learning Environments for CS: Exploring Data Structures and Algorithms with Gurukul,” in *2024 IEEE Frontiers in Education Conference (FIE)*, IEEE, 2024, pp. 1–9. doi: 10.1109/FIE61694.2024.10893211.
- [85] S. Irtiza, K. A. Akbar, A. Yasmeen, L. Khan, O. Daescu, and B. Thuraishingham, “LLM-Sentry: A Model-Agnostic Human-in-the-Loop Framework for Securing Large Language Models,” in *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, IEEE, 2024, pp. 245–254. doi: 10.1109/TPS-ISA62245.2024.00036.
- [86] D. Liao *et al.*, “A3-CodGen: A Repository-Level Code Generation Framework for Code Reuse With Local-Aware, Global-Aware, and Third-Party-Library-Aware,” *IEEE Trans. Softw. Eng.*, vol. 50, no. 12, pp. 3369–3384, 2024, doi: 10.1109/TSE.2024.3486195.
- [87] G. Wang *et al.*, “MeTMaP: Metamorphic Testing for Detecting False Vector Matching Problems in LLM Augmented Generation,” in *FORGE ’24: 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*, ACM, 2024, pp. 12–23. doi: 10.1145/3650105.3652297.
- [88] C. Zhang *et al.*, “MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2024, pp. 13647–13657. doi: 10.1109/CVPR52733.2024.01295.
- [89] T. Sun, A. Somalwar, and H. Chan, “Multimodal Retrieval Augmented Generation Evaluation Benchmark,” in *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, IEEE, 2024, pp. 1–5. doi: 10.1109/VTC2024-Spring62846.2024.10683437.
- [90] S. Kukreja, T. Kumar, V. Bharate, A. Purohit, A. Dasgupta, and D. Guha, “Performance Evaluation of Vector Embeddings with Retrieval-Augmented Generation,” in *2024 9th International Conference on Computer and Communication Systems (ICCCS)*, IEEE, 2024, pp. 333–340. doi: 10.1109/ICCCS61882.2024.10603291.
- [91] L. Garza, L. Elluri, A. Piplai, A. Kotal, D. Gupta, and A. Joshi, “PrivComp-KG: Leveraging KG and LLM for Compliance

- Verification,” in *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, IEEE, 2024, pp. 97–106. doi: 10.1109/TPS-ISA62245.2024.00021.
- [92] B. Saha, U. Saha, and M. Z. Malik, “Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance,” *IEEE Access*, p. 1, 2024, doi: 10.1109/ACCESS.2024.3513155.
- [93] B. Başaran, “Rethinking AI in Education: A Comparative Study of Custom and External Systems,” in *2024 6th International Workshop on Artificial Intelligence and Education (WAIE)*, IEEE, 2024, pp. 16–28. doi: 10.1109/WAIE63876.2024.00011.
- [94] D. Horita, N. Inoue, K. Kikuchi, K. Yamaguchi, and K. Aizawa, “Retrieval-Augmented Layout Transformer for Content-Aware Layout Generation,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2024, pp. 67–76. doi: 10.1109/CVPR52733.2024.00015.
- [95] W. Sun, H. Li, M. Yan, Y. Lei, and H. Zhang, “Revisiting and Improving Retrieval-Augmented Deep Assertion Generation,” in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, IEEE, 2023, pp. 1123–1135. doi: 10.1109/ASE56229.2023.00090.
- [96] A. Tommasel and I. Assent, “Semantic grounding of LLMs using knowledge graphs for query reformulation in medical information retrieval,” in *2024 IEEE International Conference on Big Data (BigData)*, IEEE, 2024, pp. 4048–4057. doi: 10.1109/BigData62323.2024.10826117.
- [97] H. Qiu *et al.*, “SteLLA: A Structured Grading System Using LLMs with RAG,” in *2024 IEEE International Conference on Big Data (BigData)*, IEEE, 2024, pp. 8154–8163. doi: 10.1109/BigData62323.2024.10825385.
- [98] J. Pasaribu, N. Yudistira, and W. F. Mahmudy, “Tabular Data Classification and Regression : XGBoost or Deep Learning with Retrieval-Augmented Generation,” *IEEE Access*, p. 1, 2024, doi: 10.1109/ACCESS.2024.3518205.
- [99] J. Yang, X. Wang, Q. Ni, Y. Liu, J. Wang, and F.-Y. Wang, “Towards Intelligent and Trustworthy Crowdsourcing Systems with Cyber-physical Social Intelligence via Retrieval-augmented Generation,” in *2024 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, IEEE, 2024, pp. 1–6. doi: 10.1109/ICCSI62669.2024.10799393.
- [100] Y. Tevissen, K. Guetari, and F. Petitpont, “Towards Retrieval Augmented Generation over Large Video Libraries,” in *2024 16th International Conference on Human System Interaction (HSI)*, IEEE, 2024, pp. 1–4. doi: 10.1109/HSI61632.2024.10613524.
- [101] Tamanna, S. Rana, and P. Choudhary, “Navigating the Roadblocks: Lessons Learned from a Gardening Society Conversational RAG Bot,” in *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, 2024, pp. 618–626. doi: 10.1109/AIC61668.2024.10731053.
- [102] V. Patel, P. Tejani, J. Parekh, K. Huang, and X. Tan, “Developing A Chatbot: A Hybrid Approach Using Deep Learning and RAG,” *Proc. - 2024 IEEE/WIC Int. Conf. Web Intell. Intell. Agent Technol. WI-IAT 2024*, pp. 273–280, 2024, doi: 10.1109/WI-IAT62293.2024.00043.