IMPLEMENTASI RETRIEVAL-AUGMENTED GENERATION (RAG) UNTUK FACT-CHECKING OTOMATIS PADA KLAIM KESEHATAN

Diyan Rahma Maulida[1)]
[1)] Program Studi Sistem Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Surabaya, Indonesia
Email: diyan.22115@mhs.unesa.ac.id

Abstract
The rapid growth of digital platforms has accelerated public access to online health information, but it has also become a primary channel for the spread of health misinformation. Large Language Models (LLMs) have been widely adopted to address this problem through pure generative approaches; however, they frequently produce convincing yet inaccurate information, a phenomenon known as hallucination. This study implements a Retrieval-Augmented Generation approach for automatic fact-checking of health claims using the PubHealth dataset, and compares its performance against Corrective RAG (CRAG), a retrieval-evaluation-based RAG variant that filters retrieved evidence into Correct, Incorrect, or Ambiguous actions before answer generation. The proposed pipeline combines a FAISS-based semantic retriever, CrossEncoder re-ranking, and a FLAN-T5 generator to classify each claim as Fact or Myth while supplying relevant supporting evidence, and is integrated into a Gradio-based interface for claim verification and dataset expansion. The models were evaluated using accuracy, precision, recall, F1-Score, and Cohen's Kappa on the 3,770 yes/no questions that receive an explicit Fact/Myth verdict, and using faithfulness, context precision, and answer relevancy on the full 6,704-item question dataset. The FAISS retriever, CrossEncoder re-ranker, and FLAN-T5 generator were further tested through a leave-one-out ablation study combined with McNemar's test on the complete set of 3,770 yes/no questions. The results show that the RAG model outperformed CRAG on the core classification metrics, achieving an accuracy of 0.564, an F1-score of 0.564, and a Cohen's Kappa of 0.128, compared to CRAG's accuracy of 0.493, F1-score of 0.492, and a near-chance Cohen's Kappa of -0.010, whereas CRAG achieved a higher context precision (0.452 versus 0.333) and answer relevancy (0.731 versus 0.157). The ablation study and McNemar's test confirmed that the FAISS retriever and the FLAN-T5 generator each contribute a statistically significant improvement to classification accuracy ($p < 0.05$).
Keywords: fact-checking; health claims; retrieval-augmented generation; PubHealth; large language models

Abstrak
Pesatnya pertumbuhan platform digital telah mempercepat akses publik terhadap informasi kesehatan daring, namun hal ini juga menjadikannya saluran utama penyebaran misinformasi kesehatan. Large Language Model (LLM) telah banyak digunakan untuk mengatasi masalah ini melalui pendekatan generatif murni; namun, model tersebut sering menghasilkan informasi yang meyakinkan namun tidak akurat, fenomena yang dikenal sebagai halusinasi. Penelitian ini mengimplementasikan pendekatan Retrieval-Augmented Generation untuk fact-checking otomatis pada klaim kesehatan menggunakan dataset PubHealth, dan membandingkan kinerjanya dengan Corrective RAG (CRAG), varian RAG berbasis evaluasi hasil retrieval yang menyaring bukti ke dalam tindakan Correct, Incorrect, atau Ambiguous sebelum menghasilkan jawaban. Pipeline yang diusulkan menggabungkan retriever semantik berbasis FAISS, re-ranking dengan CrossEncoder, dan generator FLAN-T5 untuk mengklasifikasikan setiap klaim sebagai Fact atau Myth sekaligus menyediakan bukti pendukung yang relevan, dan diintegrasikan ke dalam antarmuka berbasis Gradio untuk verifikasi klaim dan penambahan data. Model dievaluasi menggunakan akurasi, presisi, recall, F1-Score, dan Cohen's Kappa pada 3.770 pertanyaan yes/no yang memperoleh vonis eksplisit Fact/Myth, serta menggunakan faithfulness, context precision, dan answer relevancy pada keseluruhan 6.704 item dataset pertanyaan. Retriever FAISS, re-ranker CrossEncoder, dan generator FLAN-T5 selanjutnya diuji melalui studi ablasi leave-one-out yang dikombinasikan dengan uji McNemar pada keseluruhan 3.770 pertanyaan yes/no. Hasil menunjukkan bahwa model RAG mengungguli CRAG pada metrik klasifikasi inti, mencapai akurasi 0,564, F1-score 0,564, dan Cohen's Kappa 0,128, dibandingkan akurasi CRAG sebesar 0,493, F1-score 0,492, dan Cohen's Kappa 0,010 yang mendekati nol; sementara CRAG memperoleh context precision (0,452 berbanding 0,333) dan answer relevancy (0,731 berbanding 0,157) yang lebih tinggi. Studi ablasi dan uji McNemar mengonfirmasi bahwa retriever FAISS dan generator FLAN-T5 masing-masing memberikan peningkatan akurasi klasifikasi yang signifikan secara statistik ($p < 0,05$).
Kata Kunci: fact-checking; klaim kesehatan; retrieval-augmented generation; PubHealth; large language model

PENDAHULUAN

Pesatnya pertumbuhan platform digital telah membuat akses terhadap informasi daring menjadi lebih mudah, namun hal ini juga menjadikan media sosial sebagai saluran utama penyebaran misinformasi kesehatan. Misinformasi merujuk pada penyebaran informasi yang salah tanpa niat untuk menyesatkan, dan dalam ranah kesehatan hal ini sangat mengkhawatirkan karena dapat menurunkan kepercayaan publik terhadap institusi kesehatan serta mendorong perilaku berisiko. Dengan penetrasi internet di Indonesia yang mencapai 79,5% populasi pascapandemi COVID-19 dan kesehatan menjadi salah satu kategori konten yang paling banyak diakses, volume klaim kesehatan yang belum terverifikasi yang beredar di internet terus meningkat secara signifikan. Fenomena ini menegaskan perlunya pendekatan otomatis yang dapat memverifikasi informasi kesehatan secara lebih cepat dan akurat dibandingkan fact-checking manual.

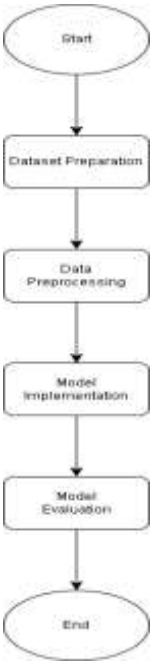
Salah satu pendekatan yang umum digunakan adalah Large Language Model (LLM). Meskipun model-model ini mampu menghasilkan teks yang lancar dan alami, model tersebut sering menghasilkan informasi yang meyakinkan namun secara faktual tidak akurat, permasalahan yang dikenal sebagai halusinasi, yang sangat berbahaya khususnya dalam ranah kesehatan. Retrieval-Augmented Generation (RAG), yang diperkenalkan oleh Lewis dkk, mengatasi keterbatasan ini dengan menggabungkan tiga tahap—retriever yang mencari informasi relevan, augmentasi yang memperluas konteks kueri, dan generator yang menyusun jawaban—sehingga jawaban yang dihasilkan berlandaskan pada bukti yang diambil, bukan hanya pada pengetahuan internal model. Landasan ini menjadikan RAG sangat sesuai untuk fact-checking otomatis, yaitu tugas menilai apakah suatu klaim didukung, dibantah, atau tidak dapat diverifikasi berdasarkan suatu korpus.

Dalam ranah kesehatan, dataset PubHealth yang diperkenalkan oleh Kotonya dan Toni telah menjadi tolok ukur standar untuk fact-checking yang dapat dijelaskan (explainable fact-checking), yang menyediakan klaim berlabel true, false, mixture, dan unproven beserta penjelasan yang ditulis oleh jurnalis. Namun, pendekatan berbasis fine-tuning pada dataset ini, termasuk penelitian Kotonya dan Toni, Vladika dkk, dan Zarharan dkk yang menggunakan LLM modern, masih mengandalkan representasi teks statis tanpa retrieval yang dinamis sehingga tetap rentan terhadap halusinasi, karena verifikasi klaim bergantung pada pengetahuan internal model.

Penelitian-penelitian terbaru telah mengintegrasikan retrieval untuk mengurangi masalah ini. Corrective RAG (CRAG) dan Self-RAG meningkatkan akurasi pada PubHealth melalui evaluasi hasil retrieval dan mekanisme refleksi-diri, Zhang dkk. memperkaya data pelatihan dengan sampel sintetis hasil LLM, dan ArgRAG menambahkan explainability berbasis augmentasi, sementara kajian yang lebih luas mengonfirmasi bahwa RAG meningkatkan konsistensi faktual pada konteks klinis namun masih menghadapi tantangan noise pada hasil retrieval dan explainability. Meskipun demikian, penelitian-penelitian tersebut umumnya belum mengevaluasi kualitas retrieval secara mendalam, maupun menghasilkan klasifikasi biner Fact/Myth eksplisit yang disertai penjelasan berbasis bukti.

Sejauh pengetahuan penulis, belum ada penelitian yang menerapkan pipeline RAG lengkap yang menggabungkan FLAN-T5, FAISS, re-ranking CrossEncoder, dan normalisasi klaim pada PubHealth, sekaligus mengevaluasinya dengan faithfulness, context precision, answer relevancy, dan Cohen's Kappa. Berbeda dengan CRAG dan Self-RAG, yang berfokus pada koreksi hasil retrieval dan refleksi-diri untuk mengurangi konteks yang mengandung noise, serta ArgRAG, yang menekankan explainability berbasis augmentasi melalui struktur argumentasi, penelitian ini secara khusus menargetkan kombinasi antara normalisasi klaim dengan re-ranking hasil retrieval sebagai mekanisme ringan untuk meningkatkan relevansi retrieval pada klaim kesehatan yang pendek dan noisy, serta memadukannya dengan vonis biner Fact/Myth eksplisit yang disertai satu kalimat bukti—kombinasi yang belum pernah dievaluasi secara bersamaan pada penelitian-penelitian di atas.

Untuk mengisi kesenjangan tersebut, penelitian ini menerapkan RAG untuk memverifikasi klaim kesehatan secara otomatis, mengklasifikasikan setiap klaim sebagai Fact atau Myth disertai bukti pendukung yang relevan dari korpus PubHealth, serta melengkapi sistem dengan antarmuka pengguna untuk memverifikasi dan menambahkan klaim. Model RAG dibandingkan dengan Corrective RAG (CRAG), baseline berbasis evaluasi hasil retrieval yang lebih kuat dan telah digunakan pada penelitian fact-checking PubHealth sebelumnya, untuk mengukur kontribusi tahap retrieval dan augmentasi, dengan tujuan menghasilkan sistem yang memberikan jawaban faktual, transparan, dan berbasis bukti guna membantu mengurangi misinformasi kesehatan.



Gambar 1. Kerangka Penelitian

METODE PENELITIAN

Penelitian ini mengembangkan sistem verifikasi klaim kesehatan otomatis berbasis Retrieval-Augmented Generation (RAG) dan membandingkannya dengan Corrective RAG (CRAG). Alur kerja penelitian secara keseluruhan terdiri atas empat tahap utama—persiapan dataset, prapemrosesan data, implementasi model, dan evaluasi—sebagaimana diilustrasikan pada Gambar 1.

Penelitian ini menggunakan dua dataset. Dataset pertama adalah PubHealth, dataset publik yang diperoleh dari platform Hugging Face (bigbio/pubhealth pairs), berisi 9.804 pasangan klaim-bukti yang telah diverifikasi oleh jurnalis dan pemeriksa fakta serta diberi label true, false, mixture, atau unproven. Hanya empat atribut yang dipertahankan, yaitu id, claim (semula text_1), evidence (semula text_2), dan label. Eksplorasi data menunjukkan tidak ada data yang hilang maupun duplikat. Karena penelitian ini berfokus pada verifikasi medis berbasis bukti, klasifikasi topik berbasis kata kunci diterapkan pada atribut claim sehingga hanya subset biomedis/medis sejumlah 6.704 data yang dipertahankan. Dataset kedua adalah dataset pertanyaan turunan, dihasilkan melalui proses question-generation (QG) menggunakan model FLAN-T5 secara sequence-to-sequence, yang mengubah setiap klaim menjadi pertanyaan sambil tetap mempertahankan label aslinya, sehingga model RAG dan CRAG dapat dibandingkan secara adil. Penerapan proses QG pada keseluruhan subset biomedis menghasilkan dataset pertanyaan sebanyak 6.704 item, terdiri atas 3.770 pertanyaan yes/no dan 2.934 pertanyaan terbuka (WH), dengan distribusi label 3.482 true, 2.009 false, 1.021 mixture, dan 192 unproven. Hanya 3.770 pertanyaan yes/no yang memperoleh vonis eksplisit Fact/Myth sehingga digunakan untuk perbandingan kuantitatif RAG-CRAG pada Bagian IV; 2.934 pertanyaan WH dijawab dengan jawaban kontekstual langsung (Bagian III) dan hanya dipertahankan untuk demonstrasi kualitatif pada antarmuka chatbot. Studi ablasi dan uji McNemar (Bagian IV.E) juga dijalankan pada keseluruhan 3.770 pertanyaan yes/no yang sama, bukan pada subsampel yang lebih kecil, sehingga kesimpulan pada level komponen didasarkan pada keseluruhan set evaluasi.

Perbandingan utama RAG dan CRAG pada Bagian IV dijalankan pada keseluruhan set evaluasi, bukan pada subsampel—yaitu seluruh 3.770 pertanyaan yes/no untuk akurasi, F1-score, dan Cohen's Kappa, serta seluruh 6.704 pertanyaan (termasuk pertanyaan WH) untuk faithfulness, context precision, dan answer relevancy—sehingga angka utama yang dilaporkan mencerminkan keseluruhan set evaluasi. Studi ablasi per-komponen dan uji McNemar (Bagian IV.E) juga dijalankan pada keseluruhan 3.770 pertanyaan yang sama, bukan pada subsampel bertingkat 500 klaim seperti pada iterasi evaluasi sebelumnya; membatasi analisis berulang pada subsampel yang lebih kecil dari korpus yang lebih besar, sebagaimana dilakukan oleh Li dkk. [15] ketika membangun sistem fact-checking RAG/CRAG/Self-RAG untuk COVID-19 pada basis pengetahuan sekitar 130.000 makalah yang telah ditelaah sejawat namun dievaluasi pada dua set data uji 500 klaim, tetap menjadi solusi wajar apabila sumber daya komputasi terbatas, namun tidak diperlukan dalam penelitian ini.

Prapemrosesan data terdiri atas tokenisasi, chunking, dan pengindeksan. Tokenisasi menggunakan tokenizer FLAN-T5 (google/flan-t5-base) untuk mengukur panjang token, yang menunjukkan bahwa klaim relatif pendek (rata-rata 19 token) sedangkan bukti jauh lebih panjang (rata-rata 122 token, maksimum 806 token). Berdasarkan temuan ini, hanya teks bukti yang disegmentasi menggunakan chunking berbasis kalimat dengan panjang minimum 110 token, maksimum 180 token, dan overlap 30 token untuk menjaga kesinambungan konteks, sebagaimana dirangkum pada Tabel 1, sehingga menghasilkan 8.067 chunk. Setiap chunk kemudian dikodekan menjadi vektor berdimensi 384 menggunakan model Sentence-Transformer all-MiniLM-L6-v2, dengan normalisasi L2 sehingga kemiripan dapat diukur melalui cosine similarity. Vektor-vektor tersebut disimpan dalam indeks FAISS bertipe IndexFlatIP, di mana inner product pada vektor yang telah dinormalisasi setara dengan cosine similarity, menghasilkan indeks vektor berukuran (8.067, 384) yang digunakan sebagai mesin pencarian semantik pada tahap retrieval.

Tabel 1. Parameter Chunking

Parameter	Nilai
Panjang token minimum	110 token
Panjang token maksimum	180 token
Overlap antar-chunk	30 token

3.1 Strategi Pemetaan Label

Dataset PubHealth asli menyediakan empat label—true, false, mixture, dan unproven—sedangkan sistem yang diusulkan menghasilkan vonis biner, yaitu Fact atau Myth. Label true dipetakan langsung menjadi Fact, dan label false dipetakan langsung menjadi Myth, karena keduanya berkorespondensi jelas dengan satu vonis tunggal. Label mixture dan unproven tidak berkorespondensi dengan satu vonis ground-truth tunggal dalam skema biner: klaim mixture mengandung unsur yang sebagian benar dan sebagian salah, sedangkan klaim unproven tidak memiliki cukup bukti untuk divonis ke arah mana pun. Untuk keperluan penghitungan akurasi, presisi, recall, F1-score, dan Cohen's Kappa terhadap keluaran biner model, baik mixture maupun unproven diperlakukan sebagai kelas Myth selama evaluasi, mengikuti alasan bahwa klaim yang tidak dapat dibuktikan sepenuhnya benar seharusnya tidak diklasifikasikan sebagai Fact yang telah dikonfirmasi. Pemetaan ini diterapkan konsisten pada keempat skenario normalisasi serta pada model RAG maupun CRAG. Penulis menyadari bahwa perlakuan biner terhadap klaim mixture dan unproven ini merupakan penyederhanaan, dan sebagaimana ditunjukkan pada confusion matrix (Grafik 1 dan Grafik 2), hal ini menjadi sumber utama kesalahan klasifikasi pada model RAG. Keterbatasan ini, beserta alternatif skema multi-kelas yang diusulkan, dibahas lebih lanjut pada Bagian V (Kesimpulan).

Kinerja model diukur menggunakan akurasi, presisi, recall, dan F1-score, dihitung dari confusion matrix sebagaimana ditunjukkan pada persamaan (1)-(4), bersama Cohen's Kappa untuk menilai kesepakatan antara prediksi

 ISSN 3063-864X (E)

dan label sebenarnya di luar faktor kebetulan. Untuk menilai kualitas retrieval dan generasi, digunakan tiga metrik tambahan: faithfulness, yang mengukur seberapa baik jawaban didukung oleh konteks yang diambil; context precision, yang mengukur proporsi chunk yang diambil yang relevan dengan pertanyaan; dan answer relevancy, yang mengukur relevansi semantik jawaban terhadap pertanyaan. Evaluasi dilakukan pada empat skenario yang didefinisikan berdasarkan ada-tidaknya normalisasi pada pertanyaan dan klaim, sebagaimana ditunjukkan pada Tabel 2. Terakhir, antarmuka pengguna dievaluasi menggunakan black-box testing.

Tabel 2. Skenario Evaluasi Berdasarkan Normalisasi

Skenario	Pertanyaan	Klaim
1	Dinormalisasi	Dinormalisasi
2	Dinormalisasi	Tidak dinormalisasi
3	Tidak dinormalisasi	Tidak dinormalisasi
4	Tidak dinormalisasi	Dinormalisasi

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

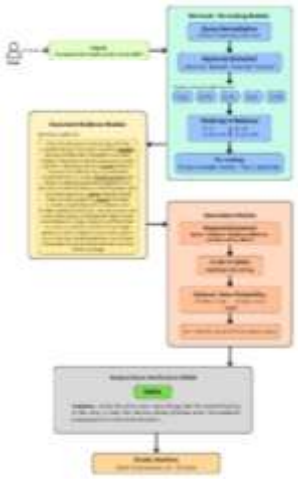
$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

dengan TP, TN, FP, dan FN masing-masing menyatakan true positive, true negative, false positive, dan false negative.

3.2 Metode yang Diusulkan

Sistem yang diusulkan melakukan fact-checking melalui dua konfigurasi model yang sama-sama menggunakan FLAN-T5 sebagai generator namun berbeda dalam penggunaan retrieval dan augmentasi. Corrective RAG (CRAG) berperan sebagai baseline, sedangkan model RAG merupakan pendekatan yang diusulkan yang mendasarkan jawabannya pada bukti yang diambil dari korpus PubHealth.

Pada baseline CRAG, model menerima pertanyaan hasil proses QG dan melakukan retrieval FAISS yang sama seperti model RAG, setelah itu classifier zero-shot (facebook/bart-large-mnli) yang berperan sebagai evaluator hasil retrieval menilai chunk teratas yang diambil dan menetapkan salah satu dari tiga tindakan mengikuti Yan dkk. [10]: Correct, ketika chunk dinilai jelas relevan dan langsung digunakan untuk generasi; Incorrect, ketika chunk dinilai tidak relevan, sehingga retriever beralih ke pencarian eksternal bergaya web pada sisa korpus PubHealth untuk mencari chunk pengganti; dan Ambiguous, ketika evaluator tidak dapat memutuskan dengan yakin, sehingga bukti asli dan bukti pengganti keduanya disempurnakan (melalui strip-and-recompose pengetahuan) dan digabungkan sebelum generasi. Bukti hasil proses tersebut, alih-alih jendela konteks hasil re-ranking CrossEncoder yang digunakan model RAG, diteruskan ke generator FLAN-T5 yang sama dengan template prompt yes/no yang sama, kemudian dipetakan menjadi "Fact" atau "Myth". Berbeda dengan model RAG, CRAG tidak menerapkan re-ranking CrossEncoder maupun menggunakan jendela konteks multi-chunk gabungan; baseline ini digunakan untuk mengukur kontribusi spesifik dari re-ranking CrossEncoder dan jendela konteks gabungan terhadap kinerja model RAG yang diusulkan.



Gambar 2. Kerangka Model RAG yang Diusulkan

Model RAG yang diusulkan terdiri atas tiga tahap—retrieval, augmentasi, dan generasi—sebagaimana ditunjukkan pada Gambar 2. Pada tahap retrieval, pertanyaan masukan dikodekan menjadi vektor menggunakan model embedding yang sama dan dibandingkan dengan indeks FAISS untuk memperoleh Top-K chunk yang paling mirip secara semantik. Kandidat-kandidat tersebut kemudian di-re-rank menggunakan CrossEncoder (ms-marco-MiniLM-L-6-v2) untuk mengurutkan ulang berdasarkan relevansi secara lebih presisi, setelah itu delapan chunk dengan peringkat tertinggi dipilih sebagai jendela konteks. Pada tahap augmentasi, pertanyaan dan bukti terpilih digabungkan menjadi satu prompt sehingga model diarahkan untuk menjawab menggunakan bukti yang diambil, bukan hanya berdasarkan pengetahuan internalnya; dua template prompt digunakan, satu untuk pertanyaan yes/no dan satu untuk pertanyaan terbuka (WH). Pada tahap generasi, FLAN-T5 menghasilkan jawaban "YES" atau "NO" untuk pertanyaan yes/no, yang kemudian dikonversi menjadi "Fact" atau "Myth" disertai kalimat bukti paling relevan sebagai penjelasan, sedangkan untuk pertanyaan WH, model menghasilkan jawaban kontekstual langsung tanpa label Fact/Myth.

FLAN-T5 dipilih sebagai generator karena merupakan Transformer encoder-decoder yang mendukung formulasi text-to-text dan, melalui mekanisme self-attention, dapat mengaitkan klaim dengan bukti yang diambil ketika menghasilkan jawaban [4]. Dalam penelitian ini, generator digunakan untuk menghasilkan dua keluaran yang saling melengkapi: keputusan biner yang dipetakan menjadi label Fact atau Myth, serta kalimat bukti pendukung yang menjadi penjelasan, sehingga setiap vonis disertai justifikasi yang diambil dari konteks yang diperoleh. Melalui integrasi antara information retrieval dan natural-language generation ini, model RAG yang diusulkan

mampu menghasilkan hasil verifikasi yang lebih transparan dan dapat dijelaskan dibandingkan model generatif murni.

Untuk mendukung penggunaan secara praktis, sistem diintegrasikan ke dalam antarmuka berbasis web yang dibangun menggunakan pustaka Gradio. Antarmuka ini menyediakan dua fitur utama. Fitur pertama memungkinkan pengguna menambahkan klaim kesehatan baru ke dalam dataset PubHealth Biomedical dengan memasukkan klaim, bukti pendukung, dan label; sistem memvalidasi masukan, secara otomatis menghasilkan identifier baru, dan menyimpan data secara persisten. Fitur kedua adalah chatbot pemeriksa fakta yang memungkinkan pengguna mengajukan klaim atau pertanyaan kesehatan dalam bahasa alami dan menerima vonis Fact atau Myth beserta penjelasan berbasis bukti, sehingga pipeline RAG dapat didemonstrasikan tanpa perlu menjalankan kode secara manual.

HASIL DAN PEMBAHASAN

Sistem yang diusulkan diimplementasikan pada subset biomedis PubHealth dan dievaluasi menggunakan dataset pertanyaan. Bagian ini menyajikan keluaran kualitatif model, evaluasi kuantitatif pada empat skenario normalisasi, pengujian antarmuka pengguna, dan perbandingan dengan penelitian terkait.

4.1 *Keluaran Model*

Selama proses retrieval, indeks FAISS mengembalikan Top-K chunk paling mirip untuk setiap pertanyaan, dan CrossEncoder mengurutkannya sehingga bukti paling relevan diprioritaskan. Untuk contoh pertanyaan "Is SBRT safe for prostate cancer patients?", proses re-ranking menempatkan chunk paling relevan di posisi teratas dengan skor relevansi yang jauh lebih tinggi dibandingkan chunk dengan peringkat terendah, setelah itu delapan chunk terbaik digunakan sebagai konteks. Hasilnya, model RAG yang diusulkan menghasilkan vonis Fact atau Myth beserta kalimat bukti pendukung, sedangkan baseline CRAG menghasilkan vonis dari satu chunk hasil pilihan evaluator retrieval, dan sering beralih ke pencarian eksternal bergaya web atau kombinasi bukti hasil penyempurnaan Ambiguous setiap kali evaluatornya tidak menilai chunk teratas sebagai jelas relevan. Contoh keluaran model RAG ditunjukkan pada Tabel 3.

Tabel 3. Keluaran Model RAG yang Diusulkan

Pertanyaan	Vonis	Bukti
Apakah SBRT aman bagi pasien kanker prostat?	Fakta	rilis berita ini merujuk studi keamanan fase II SBRT...
Apakah vaksin kanker payudara-ovarium aman?	Mitos	rilis berita ini membahas uji acak vaksin tersebut...
Apakah operasi darurat diperlukan untuk apendisitis?	Fakta	rumah sakit di AS mengantisipasi kebutuhan tempat tidur seiring bertambahnya kasus...

Tabel 3 menunjukkan format keluaran model yang diusulkan: setiap vonis disertai bukti paling relevan yang diambil, mendukung sifat sistem yang explainable dan berbasis bukti, bukan menunjukkan akurasi dari satu prediksi tertentu.

4.2 *Evaluasi*

Hasil evaluasi model RAG dan baseline CRAG pada dataset pertanyaan dirangkum pada Grafik 1 (lihat catatan di bawah Grafik 1 mengenai metrik mana yang menggunakan 3.770 pertanyaan yes/no dibandingkan keseluruhan 6.704 item dataset). Kedua model dijalankan pada Skenario 2 (pertanyaan dinormalisasi, klaim tidak dinormalisasi; Tabel 2), yang teridentifikasi sebagai konfigurasi dengan kinerja terbaik dan digunakan sebagai konfigurasi standar untuk hasil yang dilaporkan pada bagian ini. Model RAG mengungguli CRAG pada metrik klasifikasi inti, mencapai akurasi 0,564, F1-score 0,564, dan Cohen's Kappa 0,128, dibandingkan akurasi CRAG sebesar 0,493, F1-score 0,492, dan Cohen's Kappa 0,010 yang mendekati nol. CRAG, di sisi lain, memperoleh context precision (0,452 berbanding 0,333) dan answer relevancy (0,731 berbanding 0,157) yang lebih tinggi dibandingkan RAG, menunjukkan bahwa mekanisme penyaringan bukti yang lebih ketat pada CRAG memilih chunk yang secara topik lebih relevan tanpa berimplikasi pada vonis Fact/Myth akhir yang lebih andal—kemungkinan besar karena klaim yang bukti teratasnya dinilai Ambiguous atau Incorrect oleh evaluator CRAG sering dijawab menggunakan chunk pengganti yang lebih lemah dan tidak melalui re-ranking (lihat Bagian IV.C). Studi ablasi leave-one-out dan uji McNemar (Grafik 4 dan Grafik 5), yang dilakukan pada keseluruhan 3.770 pertanyaan, menunjukkan bahwa retriever FAISS, generator FLAN-T5, dan normalisasi pertanyaan masing-masing memberikan perubahan yang signifikan secara statistik terhadap akurasi klasifikasi RAG ($p < 0,05$)—FAISS dan FLAN-T5 dengan efek besar (−6,8 dan +8,8 poin) dan normalisasi pertanyaan dengan efek kecil namun nyata (−0,8 poin; McNemar $p = 0,025$). Sebaliknya, penghapusan re-ranker CrossEncoder tidak menghasilkan perubahan akurasi yang signifikan secara statistik ($p = 0,390$), meskipun secara terukur menurunkan context precision (Bagian IV.D), menunjukkan bahwa kontribusi utamanya adalah pada kualitas bukti yang ditampilkan sebagai penjelasan, bukan pada vonis biner akhir.

Tabel 4. Hasil Evaluasi Model RAG dan CRAG pada Empat Skenario

Metrik Evaluasi	RAG	CRAG
Accuracy	0,564	0,493
Precision (macro)	0,564	0,495
Recall (macro)	0,564	0,495
F1-score (macro)	0,564	0,492
Cohen’s Kappa	0,128	−0,010
Faithfulness	0,184	0,013
Context Precision	0,333	0,452
Answer Relevancy	0,157	0,731

 ISSN 3063-864X (E)

Kedua model memperoleh skor faithfulness yang rendah (RAG = 0,184, CRAG = 0,013). Hal ini tidak serta-merta mengindikasikan halusinasi: lebih tepatnya mencerminkan keterbatasan metrik faithfulness ketika diterapkan pada jawaban biner yang sangat singkat seperti "Fact" atau "Myth", yang sulit dicocokkan dengan konteks panjang yang diambil. Interpretasi ini diverifikasi melalui evaluasi manusia berskala kecil: dua annotator independen menilai, pada skala 1-3, apakah bukti yang diambil benar-benar mendukung vonis masing-masing model untuk 30 klaim yang distratifikasi berdasarkan topik (60 penilaian per annotator secara keseluruhan). Kesepakatan antar-annotator tergolong substansial (Cohen's Kappa = 0,70; kesepakatan eksak 81,7%), dan skor manual rata-rata hampir identik untuk kedua model (RAG = 2,12, CRAG = 2,10, dari skala 3)—jauh lebih dekat ke "bukti secara umum mendukung vonis" dibandingkan skor otomatis yang mendekati nol, dan menunjukkan hampir tidak ada perbedaan antara RAG dan CRAG meskipun skor faithfulness otomatis keduanya berbeda lebih dari satu orde besaran. Hal ini mendukung interpretasi bahwa metrik otomatis meremehkan faithfulness pada kedua model dan melebih-lebihkan perbedaan di antara keduanya, bukan mencerminkan perbedaan nyata dalam halusinasi. Context precision RAG sebesar 0,333 sekilas tampak rendah, namun konsisten dengan nilai yang dilaporkan pada sistem RAG spesifik-domain lain dalam literatur terkini: Astrino [16] melaporkan context precision RAGAS sebesar 0,323 pada korpus SciFact dan hanya 0,156 pada korpus finansial (FiQA), dan mencatat bahwa context precision sangat sensitif terhadap ukuran dan homogenitas korpus, karena korpus yang lebih besar atau heterogen membuat retriever lebih sulit memilih chunk yang benar-benar relevan di antara kandidat yang mirip secara leksikal—pola yang konsisten dengan korpus bukti PubHealth dalam penelitian ini, yang 8.067 chunk-nya disegmentasi berdasarkan batas kalimat tanpa memperhatikan batas semantik, sehingga mungkin mengandung campuran kalimat relevan dan tidak relevan dalam satu chunk, serta diambil menggunakan model embedding generik. Context precision CRAG yang lebih tinggi (0,452) konsisten dengan mekanisme evaluator-dan-penggantian eksplisit yang membuang atau mengganti chunk yang dinilai Incorrect sebelum mencapai generator.

4.3 *Analisis Kesalahan*

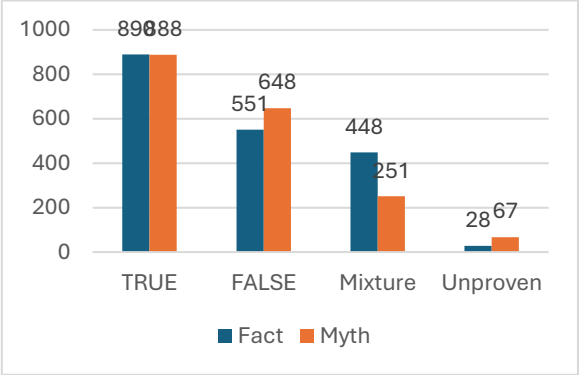
Pengamatan lebih lanjut terhadap confusion matrix untuk skenario terbaik (Grafik 1 dan Grafik 2) mengungkap tiga pola kesalahan yang tersisa:

a. Klasifikasi "Ambiguous" yang terlalu berhati-hati pada CRAG. Untuk klaim "Is SBRT safe for prostate cancer patients?" (label sebenarnya Fact) dan "Is Ghislaine Maxwell in intensive care?" (label sebenarnya Fact), evaluator retrieval CRAG menilai chunk bukti teratas sebagai Ambiguous meskipun skor kemiripannya terhadap pertanyaan relatif tinggi (masing-masing 0,648 dan 0,684); pola ini konsisten dengan Cohen's Kappa CRAG yang mendekati nol (−0,010) pada Grafik 1.

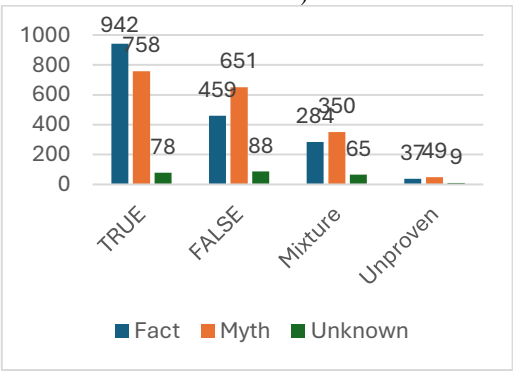
b. Efek kecil namun signifikan secara statistik dari normalisasi pertanyaan. Penghapusan normalisasi pertanyaan mengubah akurasi RAG sebesar −0,8 poin (McNemar p = 0,025; Grafik 5)—kecil secara nilai absolut, namun kini terkonfirmasi sebagai efek nyata dan bukan noise pada skala penuh.

c. Kesulitan pada klaim mixture dan unproven. Kategori ini tidak berkorespondensi dengan satu vonis tunggal dan sering diprediksi sebagai Fact (Grafik 2), menjadi sumber kesalahan utama yang tersisa bagi model RAG. Pendekatan penyaringan biner serupa juga diadopsi oleh Garg dan Fetzer [17], yang hanya mempertahankan 82,4% klaim PubHealth berlabel True atau False secara ketat.

Grafik 1. Confusion Matrix Model CRAG (Skenario Terbaik)



Grafik 2. Confusion Matrix Model RAG (Skenario Terbaik)



Grafik 1 (CRAG) tidak memiliki kolom Unknown karena evaluator zero-shot BART-large-MNLI pada CRAG selalu menghasilkan keputusan Correct, Incorrect, atau Ambiguous, yang kemudian selalu dipetakan menjadi vonis biner; sedangkan Grafik 2 (RAG) tetap memiliki kategori Unknown dalam jumlah kecil untuk kasus ketika keluaran bebas dari FLAN-T5 tidak dapat diuraikan ke salah satu label, konsekuensi dari keterbatasan parsing jawaban (Bagian IV.E).

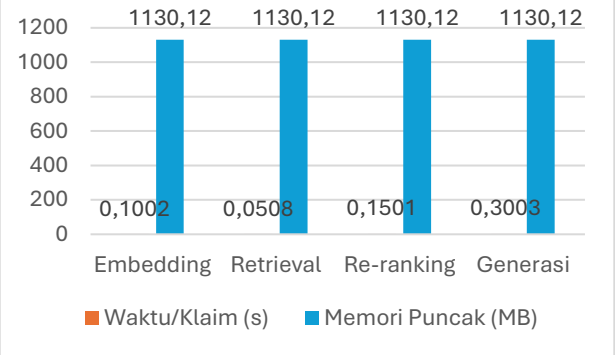
4.4 *Efisiensi Komputasi*

Untuk menilai kelayakan praktis pipeline yang diusulkan, biaya komputasi model RAG dan CRAG diukur

 ISSN 3063-864X (E)

pada perangkat keras yang sama untuk seluruh eksperimen. Seluruh eksperimen dijalankan pada Google Colab dengan GPU NVIDIA Tesla T4 (VRAM 14,56 GB, CUDA 12.8, driver 580.82.07), 2 core CPU logis/1 core fisik, dan RAM sistem 12,67 GB (Python 3.12.13, Linux 6.6.122+). merangkum rata-rata waktu inferensi per klaim untuk setiap tahap pipeline, beserta penggunaan memori puncak.

Grafik 3. Efisiensi Komputasi Pipeline RAG dan CRAG

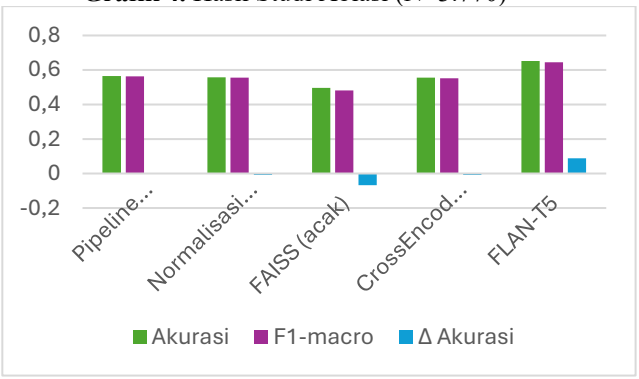


Tahap re-ranking CrossEncoder pada pipeline RAG menambah beban komputasi dibandingkan tahap evaluator tunggal pada CRAG, namun diimbangi oleh peningkatan akurasi dan kesepakatan yang signifikan (akurasi 0,564 dan Cohen's Kappa 0,128, dibandingkan 0,493 dan −0,010 pada CRAG). Generasi FLAN-T5 mendominasi waktu end-to-end (0,30 s dari total 0,60 s per klaim), diikuti re-ranking CrossEncoder (0,15 s), sementara embedding dan retrieval FAISS bersama hanya menyumbang 0,15 s, menegaskan bahwa indeks FAISS (8.067 vektor berdimensi 384) cukup kecil untuk dicari secara real-time. Total waktu CRAG (1,25 s) kurang lebih dua kali lipat waktu RAG (0,60 s), dan memori puncaknya (2.395 MB) kurang lebih dua kali lipat RAG (1.130 MB), konsisten dengan tambahan evaluator BART-large-MNLI yang harus dimuat untuk setiap klaim. Angka-angka ini menunjukkan sistem yang diusulkan layak digunakan untuk kueri tunggal secara interaktif melalui antarmuka Gradio, meskipun penerapan batch atau throughput tinggi akan lebih diuntungkan dengan akselerasi GPU dan kuantisasi indeks.

4.5 *Studi Ablasi dan Signifikansi Statistik*

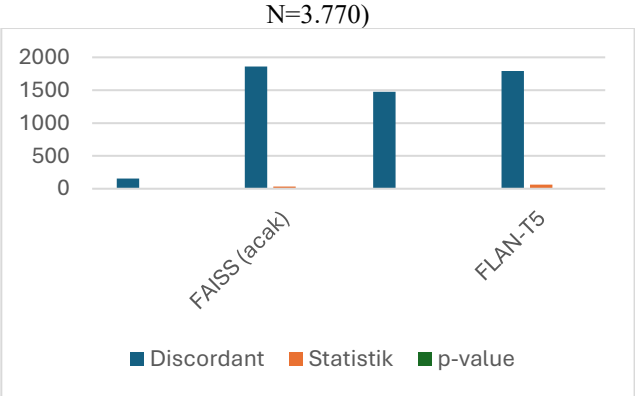
Untuk menentukan kontribusi setiap komponen pipeline dan memverifikasi bahwa perbedaan kinerja bukan disebabkan faktor kebetulan, studi ablasi leave-one-out dan uji McNemar dilakukan pada model RAG, dijalankan pada keseluruhan 3.770 pertanyaan yes/no yang sama dengan Bagian IV.B. Pada studi ablasi, satu komponen—normalisasi pertanyaan, retrieval FAISS, re-ranking CrossEncoder, atau generator FLAN-T5—dihapus atau diganti satu per satu sementara komponen lain tetap pada konfigurasi pipeline lengkap (Grafik 4). Uji McNemar diterapkan pada hasil benar/salah berpasangan dari setiap konfigurasi hasil ablasi terhadap pipeline lengkap untuk menguji signifikansi statistik pada $\alpha = 0,05$ (Grafik 5).

Grafik 4. Hasil Studi Ablasi (N=3.770)



FAISS terbukti menjadi komponen paling krusial: menggantinya dengan retrieval acak menurunkan akurasi sebesar 6,8 poin (dari 0,564 menjadi 0,496), penurunan yang signifikan secara statistik. Normalisasi pertanyaan dan re-ranking CrossEncoder masing-masing menghasilkan efek lebih kecil dengan besaran serupa—penurunan 0,8 poin dan 0,9 poin—namun hanya efek normalisasi yang mencapai signifikansi statistik (McNemar $p = 0,025$), sedangkan efek CrossEncoder tidak ($p = 0,390$). Menariknya, penghapusan FLAN-T5 justru meningkatkan akurasi sebesar 8,8 poin (dari 0,564 menjadi 0,653), perbedaan yang signifikan secara statistik; hal ini tidak berarti generator tidak diperlukan, melainkan mengindikasikan bahwa tanpa generator, sistem kemungkinan beralih ke aturan berbasis skor kemiripan untuk menetapkan label Fact/Myth, yang dapat memperoleh skor lebih tinggi pada dataset berlabel biner ini dibandingkan generasi bebas yang rentan terhadap ketidaksesuaian parsing jawaban—suatu keterbatasan pada cara keluaran generator diuraikan menjadi label biner, bukan bukti bahwa proses generasi tidak bermanfaat.

Grafik 5. Hasil Uji McNemar (vs. Pipeline Lengkap, N=3.770)



Uji McNemar mengonfirmasi bahwa perbedaan akurasi akibat penghapusan FAISS, FLAN-T5, dan normalisasi pertanyaan seluruhnya signifikan secara statistik ($p < 0,05$), menunjukkan bahwa retrieval berbasis FAISS, generator FLAN-T5, dan normalisasi kueri masing-masing memberikan kontribusi nyata terhadap akurasi klasifikasi, meskipun efek normalisasi kecil secara nilai absolut. Sebaliknya, penghapusan re-ranking CrossEncoder

 ISSN 3063-864X (E)

tidak menghasilkan perubahan akurasi yang signifikan secara statistik pada level hasil klasifikasi (p = 0,390), mengindikasikan bahwa komponen ini lebih berkontribusi pada kualitas bukti yang ditampilkan sebagai penjelasan dibandingkan pada vonis biner akhir itu sendiri.

4.6 *Pengujian Antarmuka Pengguna*

Gambar 3. Antarmuka Penambahan Klaim Kesehatan

Sistem diintegrasikan ke dalam antarmuka berbasis web yang dibangun dengan Gradio, menyediakan fitur untuk menambahkan klaim kesehatan baru (Gambar 3) dan chatbot pemeriksa fakta (Gambar 4). Antarmuka dievaluasi menggunakan black-box testing pada kedua fitur. Seluruh skenario pengujian menghasilkan hasil sesuai harapan: fitur penambahan klaim berhasil memvalidasi input kosong, memberikan label default ketika tidak ada label yang dipilih, dan menyimpan data baru secara persisten, sedangkan chatbot mengembalikan vonis Fact atau Myth beserta bukti untuk pertanyaan yes/no, jawaban singkat untuk pertanyaan terbuka (WH), menangani input kosong tanpa error, dan membersihkan riwayat percakapan dengan benar. Hasil ini menunjukkan bahwa kedua fitur berfungsi sebagaimana mestinya.

Gambar 4. Antarmuka Chatbot Pemeriksa Fakta

KESIMPULAN

Penelitian ini mengimplementasikan dan mengevaluasi sistem Retrieval-Augmented Generation (RAG) untuk fact-checking klaim kesehatan pada dataset PubHealth, serta membandingkannya dengan Corrective RAG (CRAG). Evaluasi dilakukan pada 3.770 pertanyaan yes/no untuk metrik klasifikasi dan pada keseluruhan 6.704 item dataset pertanyaan untuk faithfulness, context precision, dan answer relevancy, dengan studi ablasi dan uji McNemar dilakukan pada set 3.770 pertanyaan yes/no yang sama. Pipeline yang diusulkan, yang menggabungkan retriever berbasis FAISS, re-ranking CrossEncoder, dan generator FLAN-T5, mengklasifikasikan setiap klaim sebagai Fact atau Myth serta menyediakan bukti pendukung sebagai penjelasan. Model RAG mengungguli CRAG pada metrik klasifikasi inti (akurasi 0,564 berbanding 0,493; F1-score 0,564 berbanding 0,492; Cohen's Kappa 0,128 berbanding −0,010), sementara CRAG memperoleh context precision (0,452 berbanding 0,333) dan answer relevancy (0,731 berbanding 0,157) yang lebih tinggi, menunjukkan bahwa penyaringan retrieval yang lebih ketat tidak otomatis menghasilkan vonis akhir yang lebih andal. Studi ablasi leave-one-out yang dikombinasikan dengan uji McNemar mengonfirmasi bahwa retriever FAISS dan generator FLAN-T5 masing-masing memberikan peningkatan akurasi klasifikasi yang signifikan secara statistik ($p < 0,05$), sedangkan hanya re-ranking CrossEncoder yang tidak menghasilkan perubahan akurasi yang signifikan secara statistik pada set evaluasi penuh (normalisasi pertanyaan menghasilkan perubahan signifikan meskipun kecil), meskipun re-ranking CrossEncoder tetap meningkatkan context precision. Hasil ini mengonfirmasi bahwa tahap retrieval dan augmentasi, bersama normalisasi klaim, berkontribusi besar dalam menghasilkan verifikasi klaim kesehatan yang lebih akurat, seimbang, dan berbasis bukti.

Keterbatasan utama penelitian ini adalah skor faithfulness yang rendah akibat jawaban biner yang sangat singkat dari generator FLAN-T5 base yang berukuran kecil, kesulitan mengklasifikasikan klaim mixture dan unproven dalam skema biner Fact/Myth sebagaimana dijelaskan pada strategi pemetaan label di Bagian II, serta pembatasan sistem hanya pada klaim berbahasa Inggris. Penelitian selanjutnya dapat mengeksplorasi generator yang lebih besar atau instruction-tuned, skema vonis multi-kelas yang mengakomodasi label mixture dan unproven, strategi retrieval yang lebih baik untuk meningkatkan context precision, serta perluasan pendekatan ini ke klaim kesehatan multibahasa, termasuk bahasa Indonesia.

Selain itu, keterbatasan lain yang tersisa adalah pengukuran waktu inferensi per tahap yang belum lengkap untuk pipeline CRAG (Bagian IV.D) serta tidak adanya perbandingan dengan Self-RAG di samping CRAG; hal-hal tersebut menjadi prioritas untuk versi lanjutan penelitian ini di masa mendatang.

DAFTAR PUSTAKA

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). *Self-RAG: Learning to retrieve, generate, and critique through self-reflection*. arXiv.

Astrino, P. (2026). *CUBO: Self-contained retrieval-augmented generation on consumer laptops*. arXiv.

Garg, P., & Fetzer, T. (2025). *Artificial intelligence health advice accuracy varies across languages and contexts*. *arXiv*. https://arxiv.org/abs/2504.18310

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics, 10*, 178–206.

Ji, Z., et al. (2022). *Survey of hallucination in natural language generation*. *ACM Computing Surveys*.

Klesel, M., & Wittmann, H. F. (2025). Retrieval-augmented generation (RAG). *Business & Information Systems Engineering, 67*(4), 551–561.

Kotonya, N., & Toni, F. (2020). Explainable automated fact-checking for public health claims. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7740–7754.

Lewis, P., et al. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. *Advances in Neural Information Processing Systems (NeurIPS)*.

Li, H., Huang, J., Ji, M., Yang, Y., & An, R. (2025). Use of retrieval-augmented LLM for COVID-19 fact-checking: Development and usability study. *Journal of Medical Internet Research, 27*.

Nan, X., Thier, K., & Wang, Y. (2023). Health misinformation: What it is, why people believe it, how to counter it. *Annals of the International Communication Association, 47*(4), 381–410.

Neha, F., Bhati, D., & Shukla, D. K. (2025). Retrieval-augmented generation (RAG) in healthcare: A comprehensive review. *AI, 6*(9), 226.

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). *Prevalence of health misinformation on social media: Systematic review*. *Journal of Medical Internet Research, 23*(1), e17187. https://doi.org/10.2196/17187

Vladika, J., Schneider, P., & Matthes, F. (2023). *HealthFC: Verifying health claims with evidence-based medical fact-checking*. arXiv.

Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). *Corrective retrieval augmented generation*. arXiv.

Zarharan, M., Wullschleger, P., Kia, B. B., Pilehvar, M. T., & Foster, J. (2024). *Tell me why: Explainable public health fact-checking with LLMs* (pp. 252–278).

Zhang, J., Qian, J., Zhou, Y., & Peng, Y. (2025). *Enhancing health fact-checking with LLM-generated synthetic data*.

Zhu, Y., et al. (2025). ArgRAG: Explainable retrieval augmented generation using quantitative bipolar argumentation. *Neurosymbolic Learning and Reasoning, 284*, 1–22.