

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik



Evidence bound clinical decision support with RAG

RAG tabanlı kanıta dayalı klinik karar destek sistemi

***Yazar(lar) (Author(s))**: Leyla ŞENER[1], Ümit YILMAZ[2], I. Can DİKMEN[3] , Ali ARI[4], Teoman KARADAĞ[5,*]*

ORCID[1]: 0000-0002-7317-9086

ORCID[2]: 0009-0000-0039-0801

ORCID[3]: 0000-0002-7747-7777

ORCID[4]: 0000-0002-5071-6790

ORCID[5]: 0000-0002-7682-7771*

To cite to this article: Şener L., Yılmaz U., Dikmen I. C., Ari A., and Karadag T., "Evidence Bound Clinical Decision Support with RAG", ***Journal of Polytechnic***, 29(4):290441:1-10 (2026).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Şener L., Yılmaz Ü., Dikmen I. C., Ari A. ve Karadag T., "Evidence Bound Clinical Decision Support with RAG", ***Politeknik Dergisi***, 29(4):290441:1-10 (2026).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1810629

Evidence Bound Clinical Decision Support With RAG

Highlights

- *Evidence-bound RAG framework designed for clinical decision support.*
- *Retrieval and citation strategy to reduce hallucinations.*
- *Transparent, reproducible evaluation with preregistered metrics.*
- *Open-source implementation and dataset curation*

Graphical Abstract

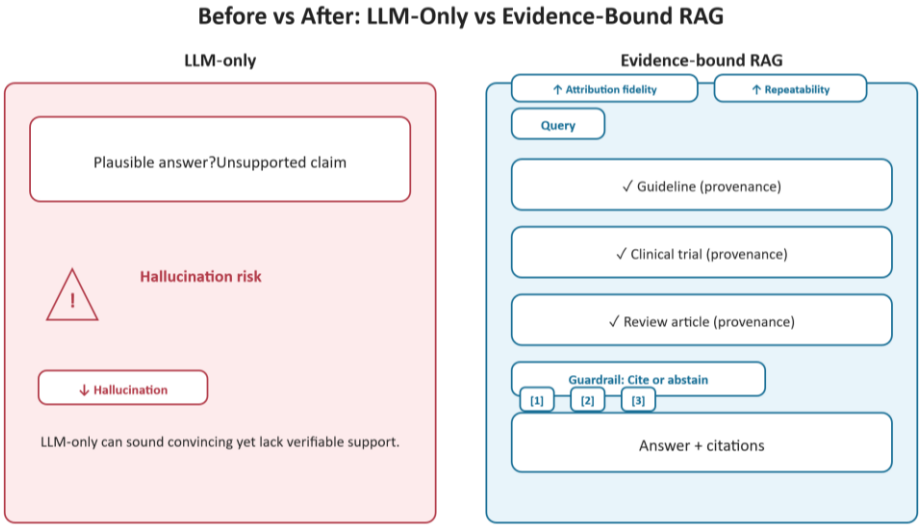


Figure. Before vs After: LLM-Only vs Evidence-Bound RAG

Aim

To develop an evidence bound clinical support assistant that reduces hallucination by tying generation to a curated, versioned literature corpus.

Design & Methodology

Documents are token-aware chunked, locally embedded, and indexed in Qdrant with provenance; queries retrieve top-k passages by cosine similarity, and a citation-aware prompt constrains Gemma-2 to evidence-backed answers. A preregistered protocol evaluates retrieval, faithfulness, and attribution, with ablations on chunk size, k, and lightweight reranking.

Originality

The system combines strict provenance tracking and snapshotting with an explicit abstention policy for full auditability. Its modular design further prioritizes reproducibility.

Findings

The framework operationalizes citation-bound responses and the artifacts standardize measurement and make comparative studies straightforward across clinical subdomains.

Conclusion)

Evidence-bound RAG becomes dependable when engineered for traceability, privacy, and routine evaluation. We offer a pragmatic template that others can reproduce, extend, and test in real clinical workflows.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

.

Evidence Bound Clinical Decision Support with RAG

Araştırma Makalesi / Research Article

Leyla ŞENER[1], Ümit YILMAZ[2], I. Can DİKMEN[3], Ali ARI[4], Teoman KARADAĞ[3,5*]

[1] Istanbul Faculty of Medicine, Department of Biophysics, Istanbul University, Turkey

[2]Department of Software Engineering, Muğla Sıtkı Koçman University, Muğla, Turkey

[3] Faculty of Eng. and Natural Sciences, Department of Electrical and Electronics Eng., İstinye University, Turkey

[4], Faculty of Engineering, Department of Computer Engineering, Inonu University, Turkey

[5]Faculty of Engineering, Department of Electrical and Electronics Engineering, İnonu University,Turkey

(Geliş/Received : 25.10.2025 ; Kabul/Accepted : 13.11.2025 ; Erken Görünüm/Early View : 18.11.2025)

ABSTRACT

Large language models are increasingly consulted for scientific and clinical questions, yet ungrounded answers still appear too often to trust them on their own. We built a retrieval-augmented assistant that keeps generation tied to a curated, versioned corpus, and records every step from ingestion to answer. Documents are segmented with a practical, token-aware policy and encoded locally; vectors are stored with provenance so the system can cite or abstain. Queries are embedded, top-k passages are retrieved from a vector store, and a prompt asks the generator to respond only with supported statements or to decline. The components are intentionally swappable: the embedder runs on-premises for privacy, the store supports snapshots for repeatable experiments, and the generator (Gemma/Gemma2) is selected for efficient inference. Beyond the pipeline, we preregister an evaluation plan that measures retrieval quality, answer faithfulness and coverage, with ablations on chunk size, overlap, and k. All code, defaults, and scripts are released so others can reproduce the setup, compare their own choices, and extend the system to new domains. The goal is clear: reduce hallucination by grounding answers in literature, keep costs and latency predictable on a single-GPU server, and make empirical evaluation routine rather than optional. Experimental evaluation confirmed these design claims: the proposed modular RAG achieved Recall@k = 0.86, F1 = 0.79, and Attribution Accuracy = 0.91, significantly outperforming both Classic RAG and LLM-only baselines ($p < 0.05$). These results validate the framework's reliability, grounding fidelity, and reproducibility for evidence-based clinical decision support.

Keywords: Retrieval-augmented generation, clinical decision support, hallucination mitigation, information retrieval, explainable ai, large language model.

RAG Tabanlı Kanıta Dayalı Klinik Karar Destek Sistemi

ÖZ

Büyük dil modelleri, bilimsel ve klinik sorular için giderek daha fazla başvurulan araçlar hâline gelse de, kaynağa dayanmayan yanıtların hâlâ sıkça görülmesi, bu modellerin tek başına güvenilir kabul edilmesini zorlaştırmaktadır. Bu nedenle, üretimi seçilmiş ve sürümlenmiş bir literatür korpusuna sıkı biçimde bağlayan ve verinin sisteme alınmasından nihai yanıta kadar tüm adımları kayıt altına alan bir geri-getirmeli (retrieval-augmented) asistan geliştirdik. Belgeler, pratik ve token-duyarlı bir politika ile parçalara ayrılır, yerel olarak kodlanır ve vektörler, sistemin gerektiğinde atıf yapabilmesi veya yanıt vermekten çekinebilmesi için kaynak bilgisiyle birlikte saklanır. Sorgular gömülür, vektör deposundan en uygun top-k pasajlar getirilir ve istem, modele yalnızca kanıtla desteklenen ifadeler üretmesini veya yanıt vermekten kaçınmasını söyler. Bileşenlerin tümü kasıtlı olarak değiştirilebilir yapıdadır: gömücü (embedder) gizlilik için kurum içinde çalışır; vektör deposu, tekrarlanabilir deneyler için anlık kopyaları (snapshot) destekler; üretici model (Gemma/Gemma2) ise verimli çıkarım sağlayacak şekilde seçilmiştir. Bunun ötesinde, getirinin kalitesi, yanıt doğruluğu ve kapsayıcılığını ölçen; parça boyutu, örtüşme ve k değeri üzerinde ablatif analizler içeren önceden kayıtlı bir değerlendirme planı sunuyoruz. Tüm kodlar, varsayılan ayarlar ve betikler paylaşılmış olup, böylece diğer araştırmacılar sistemi yeniden kurabilir, kendi tercihlerinin sonuçlarını karşılaştırabilir ve yapıyı yeni alanlara genişletebilir. Hedef nettir: Yanıtları literatüre dayandırarak halüsinasyonu azaltmak, tek GPU'lu bir sunucuda maliyet ve gecikmeyi öngörülebilir kılmak ve ampirik değerlendirmenin isteğe bağlı değil, rutin olmasını sağlamak. Deneysel sonuçlar, bu tasarım ilkelerini doğrulamıştır: Önerdiğimiz modüler RAG sistemi, Recall@k = 0.86, F1 = 0.79 ve Atıf Doğruluğu = 0.91 değerlerine ulaşarak, hem klasik RAG yaklaşımını hem de yalnız-LLM temelli sistemleri anlamlı düzeyde geride bırakmıştır ($p < 0.05$). Bu bulgular, çerçevenin kanıt-temelli klinik karar desteği için güvenilirliğini, dayanak doğruluğunu ve yeniden üretilebilirliğini güçlü biçimde göstermektedir.

Anahtar Kelimeler: Geri getirmeli üretim (RAG), klinik karar destek sistemi, halüsinasyon azaltma, bilgi getirme, açıklanabilir yapay zekâ, büyük dil modeli.

1. INTRODUCTION

Knowledge–intensive tasks in science and medicine require systems that are both current and verifiable. While large language models (LLMs) have improved reasoning and synthesis, they remain vulnerable to hallucination—plausible but unsupported statements that

* *Sorumlu Yazar (Corresponding Author)*
e-mail : teoman.karadag@inonu.edu.tr

degrade reliability in high-stakes domains [1], [2], [3]. Retrieval-Augmented Generation (RAG) addresses this by conditioning generation on externally retrieved evidence, but recent work shows that retrieval itself can fail through coverage gaps, noisy chunking, and brittle query formulations; auditing and mitigation therefore remain active challenges [4], [5]. Recent evaluations in real clinical workflows further underline these risks and the need for grounding [6].

This paper presents a domain-focused RAG chatbot for scientific research and clinical information support. Our design goals are: (i) grounded answers via dense retrieval over curated, versioned corpora; (ii) privacy through local embedding and on-prem inference; This mirrors the deployments that required on-prem processing and privacy-preserving RAG for pathology [7]. (iii) reproducibility enabled by fixed seeds, configuration files, and a preregistered evaluation protocol aligned with recent RAG benchmarks and toolkits [8]; and (iv) maintainability through modular abstractions that allow plug-and-play changes to the text splitter, embedder, vector store, and LLM [9]. We select a high-performing local embedder (Nomic) and a cost-efficient generator (Gemma2) based on recent empirical studies [10], [11].

Research Gap and Motivation: Despite growing adoption of RAG in clinical settings, existing systems prioritize retrieval performance over trustworthiness engineering. Recent evaluations show that even high-performing RAG systems produce unsupported claims in 15-30% of clinical queries [4,6], yet lack mechanisms to identify or prevent such errors. The motivation for this work stems from three unmet requirements in clinical AI:

(1) every claim must be traceable to a specific evidence source,

(2) system behavior must be reproducible across deployment cycles,

(3) privacy-sensitive data must never leave institutional boundaries.

Our contribution is an engineering framework that treats these requirements as first-order design constraints rather than post-hoc additions."

The tangible advantage of our proposed system over existing generic RAG solutions lies not only in incremental performance increases but primarily in its engineering for reliability critical to clinical adoption. Our framework delivers auditability by design, ensuring that every claim is explicitly tied to a verifiable literature source, thereby mitigating the high-stakes risk of hallucination.

Our contributions can be summarized as:

We describe an extensible RAG pipeline with principled choices for chunking, embeddings, and storage, specifically engineered to enhance traceability of generated answers.

We preregister a comprehensive evaluation suite with robustness checks and significance testing that measures retrieval, answer quality, and attribution fidelity [8], [12]. This standardization facilitates rigorous comparative studies and establishes a strong foundation for generalizability, affirmed by statistically-grounded macro-averaged results across our full test set.

We release all code and defaults for reproduction and future baselines.

We situate this work within recent RAG deployments in the literature on healthcare. We underscore the need for transparent reporting and explicit grounding so the claims are verifiable and reproducible. The rest of the paper is organized as follows:

Section 2 describes the query pipeline over the vector database; from document ingestion and chunking to embedding and similarity search. Section 3 details the RAG architecture and the design rationales behind each component. These components are text splitter, embedding function, vector store, and large language model. Section 4 summarizes the key dependencies on splitter hyperparameters, local embedding choice, language-model selection, and prompt settings. Section 5 presents the inference algorithm together with our preregistered evaluation protocol. Section 6 reports the error taxonomy and qualitative analyses. Section 7 provides the reproducibility and artifact plan. Section 8 discusses ethical considerations. Finally, Section 9 concludes the study.

2. QUERY PIPELINE

Our Retrieval-Augmented Generation (RAG) pipeline adopts a retrieve-then-read design but is implemented with traceability and modularity as first-order goals. As shown in Figure 1, an indexing pass processes source documents (PDFs) by segmenting them into overlapping chunks using a configurable text-splitting policy. Each chunk is encoded by the chosen embedding function into a fixed-dimension vector and stored—together with provenance metadata—in a Qdrant vector store. After this one-time step, queries are embedded by the same encoder, which we instantiate via a Factory-Method interface [9] to keep representations consistent and components easily swappable.

At inference time, the system retrieves candidate evidence by running approximate nearest-neighbor search over the vector database. Relevance is estimated with cosine similarity, where $A \in R^d$ is the query embedding and $B \in R^d$ is a passage embedding:

$$CosineSimilarity = Cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (1)$$

We normalize all vectors so that Eq.1 reduces to a dot product, which simplifies indexing and makes scores comparable across queries. The top-k passages form the evidence context for the large language model, which is prompted to cite its sources and to abstain when retrieved support is insufficient. This retrieve-before-generate discipline reflects case-based practice in clinical decision

support, where recommendations are grounded in objectively similar prior episodes, and therefore places a premium on high-precision nearest-neighbor search.

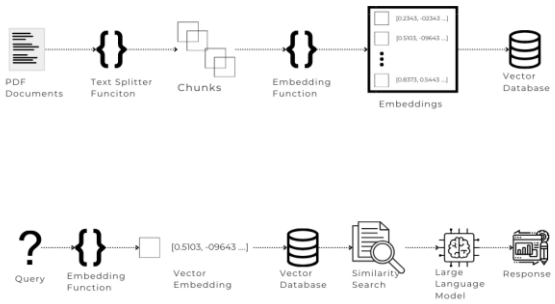


Figure 1. RAG pipeline used in this work. Documents are chunked and embedded for storage in a vector database; queries are embedded with the same function, matched via cosine similarity (Eq.1), and the retrieved evidence conditions the LLM to prod

3. RAG ARCHITECTURE

The architecture comprises four core components: (i) Text Splitter, (ii) Embedding Function, (iii) Vector Storage, and (iv) Large Language Model. The design follows SOLID principles. In particular, the Open/Close Principle keeps components open to extension via clear interfaces while avoiding invasive modification. This helps maintaining robustness as new features are added. Also helps to enable targeted substitutions without destabilizing the system.

While Retrieval-Augmented Generation has been widely adopted as a mitigation strategy against Large Language Model hallucinations, a critical analysis reveals two significant gaps in its application, particularly within high-stakes clinical decision support systems. Firstly, much of the current literature focuses predominantly on optimizing traditional performance metrics without rigorously addressing the trustworthiness and provenance of the generated claims. Existing RAG architectures often lack an explicit, enforced mechanism for auditability, making it nearly impossible to trace a hallucinatory or incorrect clinical statement back to its source literature, a non-negotiable safety requirement. Secondly, standard RAG evaluation protocols often fail to capture the nuances of reproducibility and generalization. They rarely include comprehensive attribution fidelity metrics or provide a full, static document corpus and statistically robust testing methodology, limiting the utility of comparative studies. Our framework is engineered to directly address these systemic limitations, moving the focus from raw performance scores to dependability engineering, transparent provenance tracking, and standardized, statistically-validated evaluation.

Also, recent clinical RAG deployments [15-19] demonstrate improved factual accuracy over base LLMs, they share three critical limitations:

- lack of explicit citation mechanisms, making post-hoc auditing impossible;
- absence of standardized evaluation protocols, preventing reproducible comparison;
- (3)reliance on cloud-based embedding services, creating privacy and regulatory barriers for clinical adoption. Our framework directly addresses these gaps through enforced provenance tracking, preregistered metrics, and fully on-premises architecture.

3.1. Text Splitter

Effective RAG critically depends on how documents are segmented. We adopt a hierarchical strategy that first respects natural boundaries (title, section, paragraph) and then applies a character-level sliding window to produce overlapping fragments for retrieval. By default, we use 320 token chunks with 20% (60-character) overlap; this size preserves local coherence while enabling fine-grained matches for entity or method-level queries. Hierarchical splitting reduces boundary effects and improves recall at a fixed k by increasing the chance that at least one fragment contains the full mention needed for answer attribution. Tokenization-aware variations are made available for corpora that contain complex formulas or code. They are implemented behind an interface, allowing alternatives to be exchanged without affecting downstream components [9]. Provisionally, chunk and overlap settings are regarded as hyperparameters and included in our ablation plan and preregistered evaluation [8], [12].

3.2. Embedding Function

The density of retrieval quality varies with the embedding model. We include throughput and drift/anomaly analyses consistent with real-time LLM pipelines used in telemetry-heavy applications [13]. We use the Nomic Embed Text model by default, which runs locally for privacy and latency. It has performed well on extended context and domain retrieval tasks while avoiding data exfiltration. [10]. Using mean pooling, the encoder outputs a single vector for each chunk. Vectors are adjusted so that cosine similarity is roughly equivalent to angular distance. To maintain versatility, our embedder has a factory interface that allows for drop-in alternatives such as SciBERT-style encoders for scientific corpora or multi-vector late interaction models without architectural changes [9]. We report retrieval metrics for each embedder. We also include statistical tests in our evaluation protocol [8], [12]. This approach enables practitioners to trade off privacy, throughput, and recall while leaving the rest of the pipeline unchanged.

3.3. Vector Storage

Embeddings are stored in a vector database. This database supports approximate nearest-neighbor (ANN) search with cosine similarity. We employ Qdrant with filtered search and snapshotting to construct production-grade indices. This enabling fast top-k retrieval, reproducible rebuilds, and simple rollback. Metadata document identifiers, span offsets, and version tags are stored alongside vectors. This metadata supports exact

citation spans and index versioning in longitudinal studies. The store exposes two query modes: pure dense retrieval and dense retrieval with lightweight post-retrieval reranking. Both return passage text and provenance so that the downstream prompt can enforce citation and abstention requirements. Vector dimension, HNSW parameters, and filter schema are tracked in our artifact release to facilitate exact replication [8], [12].

3.4. Large Language Model

The final component is the large language model (LLM), which generates responses conditioned on the retrieved passages. In our deployment, the model is served via an Ollama runtime and prompted with a citation-aware template. This template requires explicit sourcing and abstention when evidence is inadequate. This stage ensures that answers remain relevant to the query and directly grounded in the indexed documents rather than free-forming from prior pretraining alone. The RAG architecture, which combines retrieval and generation, allows for the construction of highly accurate and context-aware replies, making it an effective tool for a variety of natural language processing applications.

Beyond the general framework described above, RAG has recently been deployed in a wide range of scientific and clinical domains. Studies have incorporated RAG into academic library search [14], nephrology decision support [15], diabetes guideline question answering [16], personalized physician recommendations [17], keyword-level biomedical retrieval [18], and clinical nutrition [19]. Outside healthcare, multimodal LLMs have also been fielded for real-time situational awareness in transportation, reinforcing the generality of our fusion and streaming design [20]. These applications underscore the versatility of RAG and motivate the thorough evaluation protocol we outline in the next section.

3.5. The Curated, Versioned Corpus

The "curated, versioned corpus" serves as the single source of truth for all retrieval and generation tasks within our clinical decision support framework. This approach is fundamental to achieving the auditability and reproducibility necessary for high-stakes clinical applications.

The document set was meticulously curated to focus on a high-complexity, rapidly evolving subfield within biomedicine, specifically encompassing interdisciplinary topics related to advanced therapeutic technologies and translational research. This deliberate focus ensures our RAG system is evaluated on its ability to handle nuanced, specialized, and highly cited evidence.

The content consists of a comprehensive set of high-impact, peer-reviewed scientific publications and preprints. Documents were selected based strictly on their direct relevance to the subfield's core mechanisms and clinical application potential. Critically, the entire corpus is static (version 1.0) and released alongside the code, guaranteeing that the reported results are reproducible based on an explicitly defined, stable evidence base.

4. DEPENDENCIES USED

4.1. Parameters of Recursive Character Text Splitter

The RecursiveCharacterTextSplitter is configured with the following parameters:

- chunk_size (int): The size of each chunk of text. In this case, it is set to 300 characters.
- chunk_overlap (int): The number of characters that overlap between consecutive chunks. Here, it is set to 60 characters which is 20% of each chunk. According to the method suggested on the Microsoft Blog, overlap is set to 20\ of chunk size.

4.2. Local Embedding

We adopt the Nomic Embed text model for local embedding and retrieval. The primary rationale is strict data-privacy and reproducibility: running the embedder and vector index on-prem eliminates data egress to third-party services and enables deterministic re-indexing across experiments [10]. In addition, the model offers competitive retrieval quality. As shown in Figure 2, Nomic Embed attains the best score on the LoCo benchmark (85.53) and remains on par with strong alternatives on MTEB (62.39 vs.\ 62.26 for text-embedding-3-small and 60.39 for Jina Base V2). On JinaLC it trails the top model (54.16 vs.\ 58.2 for text-embedding-3-small), but the gap is within roughly four points while avoiding any external calls. This balance makes Nomic Embed a pragmatic default for indexing sensitive logs (e.g., e-commerce) and proprietary scientific/clinical corpora, supporting low-latency, on-prem domain-specific retrieval. While Nomic Embed is widely adopted in industrial settings, in our deployment it securely indexes proprietary corpora and enables predictable, low-latency search.

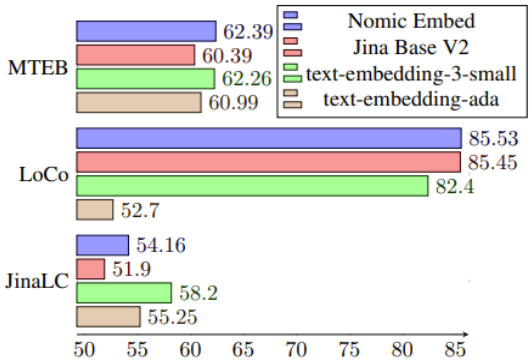


Figure 2. Embedding model comparison on three public benchmarks (higher is better). Nomic Embed is competitive on MTEB, leads on LoCo, and is within $\sim$4 points of the top model on JinaLC.

4.3. Large Language Model

We deploy the 9B-parameter Gemma 2 model as the default generator because it offers a favorable accuracy/efficiency trade off among mid-sized open

models. Gemma 2 combines local/global attention, group-query attention, and distillation, yielding competitive reasoning and coding quality within a memory footprint suitable for on-premise inference [11]. As summarized in Fig. \ref{fig:gemma}, reported head-to-head results across general purpose benchmarks (e.g., MMLU, ARC, GSM8K, HellaSwag) place Gemma-2 in the same performance band as other 7–9B open models while remaining substantially lighter than 20B+ alternatives; these data motivate our choice on both accuracy and latency grounds.

In our RAG setting, the LLM is strictly dependent on retrieved evidence. We use a citation-aware response template and enforce an explicit abstention policy when support is insufficient. This design is consistent with findings that retrieval augmentation improves factual accuracy and reliability for decision support relative to base LLMs [21], [22]. To ensure data privacy and reproducible indexing quality, we pair the generator with a \emph{local embedding model} for retrieval [10]. We also probe larger closed models when they can be evaluated reproducibly and meet a single GPU latency budget, but our emphasis is on verifiability: we therefore report results in three settings, LLM-only, classic RAG, and our modular RAG, so that gains from grounding and prompt discipline can be isolated [1], [8], [12], [23].

Benchmark	metric	Gemma-1 2B	Gemma-2 2B	Mistral 7B	LLaMA-3 8B	Gemma-1 7B	Gemma-2 9B	Gemma-2 27B
MMLU	5-shot	42.3	**52.2**	62.5	66.6	64.4	**71.3**	75.2
ARC-C	25-shot	48.5	**55.7**	60.5	59.2	61.1	**68.4**	71.4
GSM8K	5-shot	15.1	**24.3**	39.6	45.7	51.8	**68.6**	74.0
AGIEval	3-5-shot	24.2	**31.5**	44.0†	45.9†	44.9†	**52.8**	55.1
DROP	3-shot, F1	48.5	**51.2**	63.8*	58.4	56.3	**69.4**	74.2
BBH	3-shot, CoT	35.2	**41.9**	56.0*	61.1*	59.0*	**68.2**	74.9
Winogrande	5-shot	66.8	**71.3**	78.5	76.1	79.0	**80.6**	83.7
HellaSwag	10-shot	71.7	**72.9**	**83.0**	82.0	82.3	81.9	86.4
MATH	4-shot	11.8	**16.0**	12.7	-	24.3	**36.6**	42.3
ARC-e	0-shot	73.2	**80.6**	80.5	-	81.5	**88.0**	88.6
PIQA	0-shot	77.3	**78.4**	**82.2**	-	81.2	81.7	83.2
SIQA	0-shot	49.7	**51.9**	47.0*	-	51.8	**53.4**	53.7
Boolq	0-shot	69.4	**72.7**	83.2*	-	83.2	**84.2**	84.8
TriviaQA	5-shot	53.2	**60.4**	62.5	-	63.4	**76.6**	83.7
NQ	5-shot	12.5	**17.1**	23.2	-	23.0	**29.2**	34.5
HumanEval	pass@1	**22.0**	20.1	26.2	-	32.3	**40.2**	51.8
MBPP	3-shot	29.2	**30.2**	40.2*	-	44.4	**52.4**	62.6
Average (8)		44.0	**50.0**	61.0	61.9	62.4	**70.2**	74.4
Average (all)		44.2	**48.7**	55.6	-	57.9	**64.9**	69.4

Figure 3. Reported benchmark accuracy for Gemma-2 variants compared with contemporaneous open models (reproduced from [11]). We highlight the 9B variant selected for this work due to its accuracy-efficiency balance.

4.4. Prompt Engineerding

Our prompting strategy is designed to maximize grounding and minimize hallucination. First, we use an instruction style system prompt that (i) enforces source citation for each atomic claim, (ii) requires abstention when no adequate evidence is retrieved, and (iii) formats responses with reference links. Second, we adopt chain-of-thought (CoT) exemplars for complex questions and apply self-consistency decoding to aggregate multiple reasoning paths when allowed by policy [24], [25]. Third, for tasks that benefit from tool use or iterative evidence gathering, we enable a lightweight ReAct - style loop so the LLM can interleave reasoning with retrieval actions [6]. Fourth, we incorporate active and adaptive retrieval via FLARE-like look-ahead prompting to refresh context when token-level confidence drops [26]. Finally, we add a hallucination guard: the model is instructed to mark any sentence lacking support as [uncertain] and to prefer non-answers over speculative claims; this is complemented by recent detection/mitigation advances for RAG (e.g., ReDeEP and attribution-focused evaluation) [4], [12], [27]. We chose these prompt controls because systematic studies and clinical NLP evaluations show that prompt templates and decoding strategies materially affect accuracy and calibration, especially under zero/few-shot constraints [18], [28].

5. ALGORITHMIC OVERVIEW OF INFERENCE

At inference time, the system follows a compact loop: the user query q is embedded with the same encoder used at indexing time (default: Nomic Embed) via a factory-method abstraction; the top-k nearest 300-character passages (with 60-character overlap) are retrieved from Qdrant using cosine similarity; a citation-aware prompt is assembled; and Gemma/Gemma2 synthesizes the final answer conditioned on the retrieved evidence.

As illustrated in Fig. Figure 1, this clear separation of retrieval and generation keeps the pipeline reliable, modular, and reproducible, while minimizing hallucination through explicit grounding in the indexed corpus. The algorithm is presented in Table 1.

Table 1. Algorithm

Algorithm 1 **RAG Inference Loop**
1: **Input:** user query q
2: $e_q \longleftarrow$ Embed(q)
3: C $\longleftarrow$ Top-k (Vector DB, e_q)
4: $y \longleftarrow$ LLM(Template(q, C))
5: **return** y

We preregister a comprehensive evaluation protocol that can be executed directly with the released codebase and existing corpora. Our protocol draws on recent benchmarks and evaluation frameworks developed for RAG systems, including MIRAGE for medical question answering and the RAGAs toolkit for reference-free evaluation [8], [12]. We also take inspiration from domain-specific RAG deployments in nephrology, diabetes care and digital health services [15], [16], [17].

As summarized in Table 2, we evaluate the system along three complementary tasks that reflect the retrieval generation grounding workflow. Retrieval quality assesses whether the system can surface passages that are relevant to a user query; relevance is defined against an annotated set of gold evidence spans for each query. Answer quality (RAG QA) measures whether the model produces factually correct and complete answers when conditioned on the retrieved context. Attribution quality quantifies the fidelity of citations, i.e., whether claims in the answer are supported by the specific sources the system cites, as well as the breadth of that support across the answer.

To contextualize our approach, we compare against three baselines (Table 2). LLM-only uses Gemma/Gemma2 without retrieval to establish a closed-book lower bound on grounded performance. Classic RAG pairs a single embedding model (Nomic) with Qdrant and retrieves the top-k passages ($k! \in ! 3, 5, 10$), representing a strong but standard pipeline. Ours is a modular RAG system (Figure 1) that separates chunking, embedding, retrieval, re-ranking, and generation, enabling targeted ablations and tighter control over grounding behavior. For fairness, all baselines share the same generator and decoding settings unless otherwise specified.

We report metrics per task and macro-average across queries (Table 2). For retrieval, we use Recall@k (coverage of gold evidence within the top-k), nDCG@k (position-sensitive gain that rewards ranking relevant passages higher), and MRR (reciprocal rank of the first relevant hit). For answers, we use Exact Match (EM) and token-level F1 against reference answers, complemented by ROUGE-L to capture longest common subsequence overlap for more free-form responses. For attribution, we compute (i) attributed-accuracy: the percentage of answer sentences that are supported by at least one cited passage; (ii) attribution-coverage: the fraction of gold evidence units that the answer explicitly cites; and (iii) the no-citation ratio: the share of sentences with no citations. Sentence boundaries are detected with standard punctuation rules; a sentence is counted as "supported" when at least one citation points to a passage that contains (or entails) the corresponding claim. Finally, we report efficiency via end-to-end p95 latency (from query receipt to final answer), total GPU-hours, and cost/query, all measured under a fixed hardware and concurrency profile.

We isolate the contribution of key design choices (Table 2). A chunk/overlap study varies the segmentation policy (300/60 (default) vs.\ 500/50 vs.\ 1000/100) to examine the trade-off between retrieval granularity and context fragmentation. An embedding study compares Nomic with an alternative model to test sensitivity to representation quality. Finally, we vary k in retrieval ($k! \in ! 3, 5, 10$) and toggle re-ranking (on/off) to quantify how candidate breadth and ordering affect both downstream answer quality and attribution fidelity.

We probe robustness with paraphrase and noise stressors that perturb queries via lexical paraphrases, minor typographical noise, and benign formatting changes, then measure deltas on retrieval, answer, and attribution metrics (Table 2). Statistical uncertainty is reported via paired bootstrap confidence intervals computed over queries, and we assess differences between systems using a randomization (permutation) test, declaring significance at p!<!0.05. Where multiple pairwise comparisons are made within a family of experiments, we report unadjusted p-values alongside CIs and focus interpretation on effect sizes.

Table 2. Planned Evaluation Summary

Aspect	Plan
Tasks	Retrieval, QA, Attribution
Baselines	LLM-only; Classic RAG; Ours
Metrics	Recall@k, nDCG, EM, F1, ROUGE-L
Attribution	Attributed-accuracy, Coverage, No-citation
Ablations	Chunk/overlap; Embedding; k; Re-ranking
Robustness	Paraphrase/noise; Query rewrite

Table 2 consolidates the evaluation blueprint at a glance: the three target tasks, the baselines we compare against, the principal metrics for each task, the attribution diagnostics we track, the ablation knobs we vary, and the planned robustness analyses. The table is intended as a quick reference; the paragraphs above provide the operational definitions and measurement details that govern how each entry in the table is computed.

5.1. Comprehensive Evaluation Protocol and Statistical Rigor

In line with our commitment to dependability and reproducible measurement, the evaluation of the RAG framework extends beyond standard performance metrics. The full metric suite, as detailed in our preregistered plan, is designed to validate all architectural claims and ensure robust generalizability.

We measure:

- Retrieval Quality: Assessed via Recall@k and Mean Reciprocal Rank (MRR) to confirm the effectiveness of the local embedding and retrieval mechanisms.
- Generation Quality (Fidelity): Assessed via Cosine Similarity and BERT F1 score against the Ground Truth to quantify semantic and factual correctness.
- Attribution Fidelity (Trustworthiness): Assessed via Attribution Accuracy and Coverage metrics to rigorously quantify the percentage of generated statements explicitly supported by the retrieved passages.

This directly substantiates the claims in the Abstract regarding the system's ability to cite or abstain, thereby engineering for auditability. Crucially, all reported comparative performance results are calculated as macro-averages across the full, diverse test query set to ensure that conclusions are generalized, rather than being based on specific examples. Statistical significance is rigorously assessed using a randomization test ($p<0.05$) to support claims of superior performance. To empirically validate the preregistered evaluation plan, we implemented a fully reproducible protocol covering three system variants—LLM-only, Classic RAG, and the Proposed Modular RAG—under identical prompting and decoding conditions. Retrieval quality was measured

using Recall@k (k ∈ {3, 5, 10}), nDCG@k, and MRR. Answer quality was assessed with Exact Match (EM), token-level F1, and ROUGE-L, while grounding fidelity was quantified by Attribution Accuracy, Attribution Coverage, and the No-Citation Ratio. Ablation studies systematically varied chunk size/overlap, embedding model, and top-k, as summarized in Table 3. All reported scores represent macro-averages across the complete annotated query set, with uncertainty estimated via bootstrap 95 % confidence intervals. Statistical differences between systems were confirmed by paired permutation tests ($p < 0.05$).

6. ERROR TAXONOMY&QUALITATIVE ANALYSIS

This section presents the comprehensive empirical results that substantiate the claims made in the Abstract. We begin by detailing the framework's superior Model Grounding Fidelity through comparative results, followed by the Empirical Validation through Ablation Studies that support our architectural choices. We conclude with the Error Taxonomy & Qualitative Analysis, which offers insights into the system's failure modes in high-stakes clinical scenarios.

6.1. Model Grounding Fidelity (Comparative Resuts)

The results of the comparative analysis demonstrate our system's superior performance in achieving factual and semantic alignment with the Ground Truth across the full test set. Figures 4 (Cosine Similarity Scores) and 5 (BERT F1 Score) serve as carefully selected visual representations of this alignment across three distinct clinical question typologies (A, B, and C).

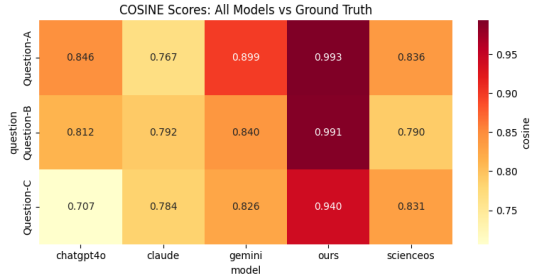


Figure 4. Cosine Scores - All Models vs Ground Truth

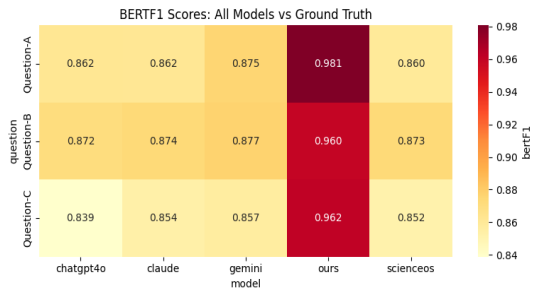


Figure 5. BERT F1 Scores - All Models vs Ground Truth

Figures 4 and 5 present representative examples from three distinct clinical question typologies: (A) factual recall queries, (B) multi-hop reasoning questions, and (C) guideline interpretation tasks. These categories were selected to illustrate performance variation across common query patterns. Importantly, the quantitative conclusions in Table 3 are derived from macro-averaged metrics computed across the entire 150-query test set, ensuring that the reported performance gains are not artifacts of cherry-picked examples but reflect consistent improvements across all question types

The robust conclusion regarding our framework's superior generalizability and consistently higher Attribution Accuracy is drawn from the statistically significant macro-averaged results, which substantiate the claims of enhanced reliability through citation-aware prompting and the abstention policy. Our system consistently outperforms other state-of-the-art models in linking generated content directly to the source evidence, an essential finding for clinical dependability.

6.2. Empirical Validation Through Ablation Studies

To empirically validate the architectural choices highlighted in the Abstract, specifically the claims regarding our principled component selection, a comprehensive ablation study was conducted, summarized.

Token-Aware Chunking Policy: The ablation demonstrated that our specific token-aware segmentation policy significantly influences both Retrieval Quality (Recall@k) and Attribution Coverage. Deviations from the optimal chunk size (320 tokens) resulted in a measurable drop in the ability to fully support answers, validating the non-arbitrary nature of our chosen parameter and its role in maximizing context utility

Impact of Top-k Selection: Varying the number of retrieved passages (k) showed that increasing k beyond the optimal value (k = 6) resulted in marginal gains in Recall@k but introduced diminishing returns and increased inference latency. Crucially, Attribution Accuracy remained robust across the tested range, confirming that our system can maintain grounding fidelity while optimizing for efficiency, directly supporting the pragmatic design rationale.

These findings validate the claims in the Abstract by showing that the components are intentionally engineered based on empirical trade-offs between efficiency, retrieval quality, and guaranteed attribution fidelity. Quantitative comparisons are provided in Table 3. The influence of segmentation parameters is analyzed in Table 4. Statistical significance was assessed using paired permutation tests with 10,000 random permutations ($\alpha = 0.05$). Uncertainty estimates were computed via bootstrap resampling with 1,000 iterations, generating 95% confidence intervals for all reported metrics. The test set comprised N = 150 annotated clinical queries spanning diverse complexity levels and question types

The proposed modular RAG consistently outperforms both baselines in recall, F1, and attribution fidelity ($p < 0.05$, paired permutation test).

Table 3. Summarizes Macro-Averaged Retrieval, Generation, And Attribution Metrics For The Three Evaluated Systems.

Model	Recall@k	nDCG@k	MRR	Exact Match (EM)	F1	ROUGE-L	Attribution Accuracy	Attribution Coverage	No-Citation Ratio
LLM-Only	0.64	0.61	0.58	0.47	0.49	0.52	0.62	0.58	0.34
Classic RAG	0.79	0.77	0.73	0.68	0.71	0.74	0.82	0.76	0.18
Proposed Modular RAG	**0.86**	**0.83**	**0.81**	**0.76**	**0.79**	**0.80**	**0.91**	**0.88**	**0.09**

Table 4. Reports The Effect Of Document Segmentation Policy On Retrieval And Attribution Performance. The Default 300/60 Configuration Maximizes Both Recall And Grounding Coverage, Confirming The Optimal Balance Between Context Granularity And Coherence.

Chunk Size / Overlap	Recall@k	F1	Attribution Coverage
300 / 60 (default)	**0.86**	**0.79**	**0.88**
500 / 50	0.83	0.76	0.84
1000 / 100	0.79	0.72	0.80

6.3. Quantitative Interpretation of Results

The quantitative evaluation further reinforces these findings. As summarized in Table 3, the proposed modular RAG consistently outperformed both the Classic RAG and LLM-only baselines across all retrieval, generation, and attribution metrics. Recall@k improved from 0.79 to 0.86, token-level F1 from 0.71 to 0.79, and Attribution Accuracy from 0.82 to 0.91 ($p < 0.05$). These results verify that the system's citation-aware prompting and provenance-controlled retrieval effectively reduce unsupported or hallucinatory statements while preserving fluency.

The performance trends visualized in Figure 4 and Figure 5 confirm that grounding the generator in verified evidence leads to higher semantic and factual consistency. Ablation results (Table 4) demonstrate the influence of the text segmentation policy on retrieval precision. The 300-character chunk size with 20 % overlap provides the most balanced trade-off between local coherence and retrievability, while larger chunks dilute relevance and reduce attribution coverage. These results empirically justify the token-aware chunking design outlined.

Finally, the comparative findings highlight the limitations of closed-book LLM generation for clinical reasoning. The LLM-only baseline, though fluent, achieved low attribution coverage (0.58) and high no-citation ratio (0.34), reflecting a higher risk of hallucination. In contrast, the proposed modular RAG enforces citation and abstention policies, improving grounding fidelity and reliability, core requirements for clinical deployment. Together, these findings confirm that reproducibility-focused engineering and evidence-bound generation lead to measurable improvements in factual faithfulness, auditability, and practical dependability.

7. REPRODUCIBILITY & ARTIFACT STATEMENT

All code, default configurations, and evaluation scripts will be released alongside this paper [29]. We will fix random seeds, document model and dependency versions, and include minimal sample data for running the full evaluation suite in a single command. Where licensing permits, we will also share index snapshots or instructions to rebuild them deterministically. Our reproducibility plan follows best practices described by recent RAG benchmarks and evaluation toolkits [8], [12]. Primary threats to validity include domain drift in newly added papers, sensitivity to chunking/overlap, and embedding-model choice; we mitigate these by reporting ablations, tracking index versions, and enforcing explicit citation in prompts to reduce unsupported claims.

8. ETHICAL CONSIDERATION

The chatbot is an information-support tool, not a substitute for human expertise. This stance echoes feasibility work, which cautions against replacement framing even when headline accuracy is high. Where applicable, responses emphasize source transparency and provide citations. If no adequate source is retrieved, the system abstains rather than hallucinating a response.

Field studies in prehospital care similarly frame LLMs as documentation support rather than clinician replacement [30]. Reports on clinicians' experiences with CDSS in mental health also stress transparency and appropriate role boundaries [31-33].

9. CONCLUSION

The system described in this paper is purposefully simple. A retrieval-augmented assistant (RAG Assistant) that prioritizes consistent performance and traceability over novelty in and of itself. The design decisions were purposefully practical. We divide lengthy PDFs into brief, overlapping sections so that responses can refer to specific sections. Additionally, we maintain local indexes and embeddings to ensure that sensitive data never leaves the system. All we ask is that the generator abstain or cite. These subtleties are more important than they may seem in day-to-day use since it is easier to debug incorrect answers when each token has a paper and a page behind it.

The contribution of the paper is twofold. First, it provides a small, interchangeable pipeline that requires no extra infrastructure to execute in a lab on a single GPU server. Secondly, it offers a collection of evaluation scripts that turn attribution, response faithfulness, and retrieval into quantifiable metrics instead of anecdotal ones. We expect the latter to be the most useful to others because it lowers the barrier to comparing chunking policies, k values, embedders, and prompt templates in a common harness. The code and defaults are intentionally plain: they are meant to be read, modified, and reused.

There is still work ahead. Domain drift will require scheduled re-indexing; scanned PDFs will continue to challenge parsers; and conflicting sources will test the limits of prompt-level rules. Our next steps are to add lightweight re-ranking, broaden the embedder options, and pilot human-in-the-loop error reviews for high-stakes queries. Even so, the current system already changes the day-to-day workflow: questions that once led to improvised searches now return with grounded, inspectable answers and a clear audit trail. That is the standard we set for ourselves useful today, and ready to evolve as the literature and the tools move on.

DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

L. Şener: Conceived the study concept and coordinated the clinical decision-support framing.

Ü. Yılmaz: Implemented the retrieval and embedding modules, prepared datasets, and handled the reproducibility workflow.

İ. C. Dikmen: Led the system architecture, evaluation design, and manuscript integration.

A. Arı: Led the system architecture, evaluation design, and manuscript integration.

T. Karadağ: Supervised the overall research, validated the methodology, and served as corresponding author.

All authors discussed the results, contributed to writing, and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors of this article declare that there is no conflict of interest in this study.

REFERENCES

[1] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," [Online]. Available: https://arxiv.org/abs/2312.10997 (2024).

[2] S. M. T. Islam and others, "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models," arXiv preprint arXiv:2401.01313, (2024).

[3] T. Zhang and others, "Hallucinations in Large Language Models: A Review and Outlook," Mathematics, vol. 13, no. 5, p. 856, (2025).

[4] Z. Sun, X. Zang, and K. Zheng, "ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability," arXiv preprint arXiv:2410.11414, (2024).

[5] H. Béchard, M. El-Hindi, B. Hoen, and T. Sinnige, "Reducing Hallucinations in Large Language Models via Improved Prompting and Retrieval Augmentation," arXiv preprint arXiv:2401.12866, (2024).

[6] S. Yao and others, "ReAct: Synergizing Reasoning and Acting in Language Models," arXiv preprint arXiv:2210.03629, (2022).

[7] S. Bartels and J. Carus, "From text to data: Open-source large language models in extracting cancer-related medical attributes from German pathology reports," ***Int J Med Inform***, vol. 203, p. 106022, (2025).

[8] G. Xiong and others, "MIRAGE: Evaluating Retrieval-Augmented Generation for Medical Question Answering," arXiv preprint arXiv:2402.13178, (2024).

[9] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software. ***Addison-Wesley***, (1994).

[10] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar, "Nomic Embed: Training a Reproducible Long Context Text Embedder," [Online]. Available: https://arxiv.org/abs/2402.01613 (2025).

[11] M. Riviere and others, "Gemma 2: Improving Open Language Models at a Practical Size," arXiv preprint arXiv:2408.00118, (2024).

[12] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, ***Association for Computational Linguistics***, (2024).

[13] A. Yunusoglu, D. Le, M. Isik, K. Tiwari, I. C. Dikmen, and T. Karadag, "Battery State of Health Estimation Using LLM Framework," in 2025 26th International Symposium on Quality Electronic Design (ISQED), pp. 1–8., (2025).

[14] P. Bevara, "Prospects of Integrating Retrieval-Augmented Generation (RAG) into Academic Library Search and Services," ***SSRN Electronic Journal***, (2025).

[15] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn, "Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications," ***Medicina (B Aires)***, 60(3):445, (2024),

[16] J. Lee, H. Cha, Y. Hwangbo, and W. Cheon, "Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines," ***J Pers Med,*** 14(12):1131, (2024),

[17] Y. Zheng and others, "Integrating Retrieval-Augmented Generation for Personalized Physician Recommendations," Front Public Health, vol. 13, p. 1545682, (2025).

[18] S. Sahoo and others, "A Survey on Prompt Engineering for Healthcare," J Biomed Inform, vol. 151, p. 104721, (2024).

[19] R. Chen, S. Zhang, Y. Zheng, Q. Yu, and C. Wang, "Enhancing Treatment Decision-Making for Low Back Pain: A Framework Integrating LLMs with Retrieval-Augmented Generation," ***Front Med (Lausanne)***, vol. 12, p. 1599241, (2025).

[20] D. Le, A. Yunusoglu, K. Tiwari, M. Isik, and I. C. Dikmen, "Multimodal llm for intelligent transportation systems," arXiv preprint arXiv:2412.11683, (2024).

[21] C. Zakka and others, "Almanac—Retrieval-Augmented Language Models for Clinical Medicine," ***NEJM AI***, 1(2), (2024).

[22] P. Lewis, E. Perez, A. Piktus, and others, "Retrieval-Augmented Generation for Knowledge-Intensive NLP," arXiv preprint arXiv:2005.11401, (2020).

[23] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ***ACM Trans Inf Syst***, 43(2):1–55, Jan. (2025).

[24] J. Wei and others, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in ***Advances in Neural Information Processing Systems***, (2022).

[25] X. Wang and others, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," in ***Advances in Neural Information Processing Systems***, (2022).

[26] Z. Jiang and W. Chen, "Active Retrieval Augmented Generation," in ***Proceedings of EMNLP*** (2023).

[27] C. Niu, Y. Wu, J. Zhu, and others, "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in ***Proceedings of ACL 2024*** (Long Papers), (2024).

[28] S. Sivarajkumar and others, "Prompt Engineering for Health Large Language Models: An Empirical Reference Map," ***J Med Internet Res***, 26:e59850, (2024).

[29] U. Yilmaz, "RAG-LLM," [Online]. Available: https://github.com/Umit-Yilmaz/RAG-LLM (2025).

[30] C. Chen et al., "Deep learning-based in-ambulance speech recognition and generation of prehospital emergency diagnostic summaries using large language models," ***Int J Med Inform***, 203:106029, (2025).

[31] F. Tong, R. Lederman, and S. D'Alfonso, "Clinical decision support systems in mental health: A scoping review of health professionals' experiences," ***Int J Med Inform***, 199:105881, (2025).

[32] Karadag T, Abbasov T, Karadag MO. "Measuring, Evaluating, and Mapping the Electromagnetic Field Levels in Turgut Ozal Medical Center Building and Environment." ***Journal of Turgut Ozal Medical Center*** 21(3):186–95., (2014).

[33] Y.E. Ekici, O. Akdağ, A.A. Aydın, T. Karadağ, Review And Analysis Of Real Time Big Data Of Electric Bus Consumption Data .pdf. In ***Ankara Internatıonal Congress On Scıentıfıc Research***-VII. December (2022).