Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00377-4

Original Article

A Hybrid Retrieval and Reranking Framework for Evidence-Grounded Retrieval-Augmented Generation

Fariba Afrin Irany [1] , Sampson Akwafuo [2,]

1. *Department of Computer Science and Engineering, University of North Texas, Denton, USA*
2. *Department of Computer Science, California State University, Fullerton, Fullerton, USA*

Received: September 07, 2026 | Review began: September 09, 2026 | Review ended: October 14, 2026 | Published: November 18, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Retrieval-augmented generation (RAG) improves the reliability of large language model outputs by grounding generated responses in external evidence. However, the quality of RAG-generated answers depends strongly on the relevance of retrieved evidence, the effectiveness of evidence ranking, and the ability to verify whether generated claims are supported by source documents. This study presents a hybrid retrieval and reranking framework for citation-aware RAG in biomedical and healthcare-related document question answering. The proposed framework uses Amazon Bedrock Knowledge Bases for document ingestion, parsing, chunking, embedding generation, and evidence retrieval. Source PDF documents are stored in Amazon S3, embedded using Amazon Titan Text Embeddings V2, and indexed using Amazon OpenSearch Serverless. A hybrid retrieval strategy is applied to retrieve candidate evidence chunks, followed by reranking to prioritize the most relevant passages before answer generation. The answer-generation stage uses the highest-ranked evidence chunks to produce controlled, evidence-grounded responses, while a separate judge model evaluates each generated factual claim against the retrieved evidence. The framework was evaluated using 150 biomedical natural language processing and healthcare transformer queries. This evaluation was designed as a pilot-scale proof-of-concept study to examine the feasibility of the proposed hybrid retrieval, reranking, answer-generation, and grounding-evaluation workflow. Across the evaluation set, the system retrieved and reranked 3,000 evidence chunks and generated responses from the top-ranked evidence. Claim-level grounding evaluation identified 1,200 factual claims, of which 1,184 were judged to be supported by the retrieved evidence and 16 were classified as unsupported, resulting in an overall grounding accuracy of 98.7%. The results indicate that hybrid retrieval, reranking, conservative prompting, and claim-level grounding evaluation can support reliable evidence-grounded RAG responses when sufficient source evidence is available. The study also highlights the distinction between internal grounding accuracy and strict citation accuracy, motivating future work on explicit sentence-level citation generation and expanded human expert validation.

Categories: AI applications, Generative AI for healthcare, Evaluation Methods

Keywords: retrieval-augmented generation, hybrid retrieval, reranking, citation-aware generation, grounding evaluation, biomedical natural language processing, amazon bedrock, opensearch serverless

Introduction

Large language models (LLMs) and transformer-based architectures have substantially advanced natural language processing (NLP) by enabling contextual representation learning, scalable text generation, and flexible adaptation to downstream tasks. The Transformer architecture introduced self-attention as the core mechanism for sequence modeling, enabling efficient parallel training and long-range contextual representation [1]. Building on this foundation, BERT demonstrated the effectiveness of bidirectional contextual pretraining across a wide range of natural language

understanding tasks [2]. These developments established much of the architectural foundation for modern LLMs and domain-specific language models. However, despite their ability to generate fluent and contextually relevant text, LLMs remain limited in knowledge-intensive applications by concerns related to factual reliability, outdated parametric knowledge, hallucination, evidence grounding, and source attribution. These limitations are particularly important in biomedical informatics, clinical NLP, healthcare document analysis, and scientific question answering, where generated information should be supported by verifiable evidence.

Retrieval-augmented generation (RAG) has emerged as an important approach for addressing some of these limitations by combining the generative capabilities of language models with information retrieved from external sources. Lewis et al. introduced RAG as a framework that combines parametric knowledge encoded within a pretrained language model with non-parametric knowledge obtained from an external document collection [3]. Rather than relying exclusively on information stored within model parameters, a RAG system retrieves relevant passages and supplies them as contextual evidence during generation. REALM similarly demonstrated that neural language models can retrieve external documents during pretraining and downstream question answering, supporting the use of retrieval as an explicit external knowledge mechanism [4]. These approaches established the conceptual foundation for retrieval-enhanced language generation in which external evidence supplements the internal knowledge of a language model.

The effectiveness of RAG, however, depends strongly on the relevance, completeness, and prioritization of the retrieved evidence. If the retrieved passages are incomplete, weakly relevant, or poorly ranked, the generation model may still produce responses that are only partially supported by the underlying documents. Dense Passage Retrieval demonstrated that dense dual-encoder representations can effectively retrieve semantically relevant passages for open-domain question answering [5]. Subsequent unsupervised dense retrieval methods showed that contrastive representation learning can produce effective retrieval models without requiring extensive supervised relevance labels [6]. Fusion-in-Decoder further demonstrated that generative question-answering models can benefit from conditioning on multiple retrieved passages, allowing evidence distributed across several contexts to contribute to answer generation [7]. Collectively, these studies demonstrate that retrieval quality is closely connected to the reliability of downstream generation.

Lexical retrieval nevertheless remains important because exact term matching can be particularly effective when queries contain technical terminology, abbreviations, named entities, or domain-specific expressions. BM25 and related probabilistic relevance approaches remain widely used because of their computational efficiency, interpretability, and strong lexical matching capability [8]. Dense retrieval complements this approach by representing queries and documents within a semantic embedding space, thereby reducing problems caused by vocabulary mismatch [5]. More advanced retrieval architectures attempt to combine semantic representation with finer-grained interaction. ColBERT uses late interaction between contextualized query and document representations to preserve token-level matching while maintaining retrieval efficiency [9], while ColBERTv2 further improves the effectiveness and storage efficiency of this approach [10]. Sparse neural retrieval methods such as SPLADE provide another alternative by learning sparse lexical representations and term expansion beyond exact surface matching [11].

Because lexical and semantic retrieval strategies have complementary strengths, hybrid retrieval has become increasingly relevant to RAG systems. The BEIR benchmark demonstrated substantial variability in retrieval performance across heterogeneous domains and datasets, indicating that no single retrieval strategy performs optimally across all information-retrieval settings [12]. Hybrid retrieval seeks to address this limitation by combining keyword-based and semantic retrieval signals, allowing relevant evidence to be discovered through both terminology matching and broader conceptual similarity. OpenSearch supports hybrid retrieval through the combination and normalization of lexical and vector-search scores during query processing [13]. Such an approach is particularly relevant in biomedical information retrieval, where specialized terminology may coexist with semantically equivalent expressions that differ substantially at the lexical level.

Retrieving a useful candidate set does not, however, guarantee that the most relevant evidence will appear at the top of the retrieval results. Passages may be semantically related to a query while differing considerably in their usefulness for supporting a specific response. Reranking therefore provides an additional evidence-selection stage after initial retrieval. In a typical two-stage architecture, the retriever identifies a relatively broad candidate set, while a reranking model subsequently performs more focused query-passage relevance estimation. Cohere reranking models, for example, are designed to reorder candidate passages according to their estimated semantic relevance to a query [14]. This separation between broad candidate retrieval and subsequent evidence prioritization is central to the framework investigated in the present study.

Retrieval and grounding challenges are especially significant in biomedical and clinical NLP because medical text contains specialized terminology, abbreviations, domain-specific relationships, and highly context-dependent meanings. Domain-specific transformer models have demonstrated the importance of adapting language representations to biomedical corpora. BioBERT showed that continued pretraining on biomedical literature improved performance across biomedical named entity recognition, relation extraction, and question-answering tasks [15]. PubMedBERT further demonstrated that pretraining a language model directly on biomedical text can provide strong performance across biomedical NLP benchmarks [16]. Large-scale clinical language models such as GatorTron have further demonstrated the potential benefits of scaling transformer architectures using electronic health record data for clinical NLP applications [17].

Biomedical question answering also illustrates the importance of evidence retrieval and domain-specific evaluation. BioASQ has provided a large-scale framework for biomedical semantic indexing and question answering based on biomedical literature [18], while PubMedQA supports biomedical research question answering using evidence derived from PubMed abstracts [19]. MIMIC-IV has additionally provided a large-scale deidentified electronic health record resource that supports reproducible research across clinical informatics and machine learning (ML) applications [20]. Although the present investigation focuses on document-grounded question answering rather than direct clinical prediction, these resources illustrate the broader need for reproducible and evidence-supported computational methods in biomedical informatics.

Recent research on LLMs in medicine has further emphasized both their potential and their limitations. Thirunavukarasu et al. reviewed applications of LLMs in medicine and emphasized the need for validation, safety assessment, and careful deployment in healthcare environments [21]. Singhal et al. evaluated LLMs on medical question answering and considered dimensions including factuality, reasoning, comprehension, potential harm, and bias [22]. These findings reinforce the importance of developing healthcare-oriented generative systems whose outputs are not only fluent and contextually appropriate, but also verifiable and supported by relevant evidence.

Hallucination remains a central challenge in neural text generation. Ji et al. characterized hallucination as generated information that is unsupported by or inconsistent with source information or factual knowledge and identified it as a persistent problem across natural language generation systems [23]. Retrieval can mitigate this risk by supplying external evidence, but retrieval alone does not guarantee that every generated statement is supported. A language model may misinterpret evidence, overgeneralize from incomplete passages, combine information incorrectly, or introduce plausible information derived from pretrained knowledge that is not directly justified by the retrieved context. Consequently, reliable RAG requires evaluation not only of retrieval performance but also of the relationship between individual generated claims and the evidence supplied to the generation model.

Research on fact and claim verification provides an important foundation for this type of evaluation. FEVER introduced large-scale evidence-based fact verification in which claims are categorized according to whether available textual evidence supports, refutes, or provides insufficient information for a statement [24]. SciFact extended evidence-based verification to scientific literature by requiring systems to identify scientific claims and supporting or contradicting evidence from research articles [25]. These approaches are particularly relevant to evidence-grounded RAG because a generated answer can be decomposed into individual factual claims, each of which can then be evaluated against retrieved source evidence rather than treating the complete response as uniformly correct or incorrect.

More recent evaluation frameworks have emphasized increasingly fine-grained assessment of generated information. FActScore decomposes long-form generated text into atomic factual statements and measures the proportion that can be supported by reliable knowledge sources [26]. ALCE evaluates long-form LLM-generated answers with citations and considers dimensions including answer quality and citation-related performance [27]. RAGAS provides an evaluation framework specifically designed for RAG systems and includes measures related to retrieval and generation quality [28]. Together, these approaches motivate claim-level grounding assessment in which individual factual statements are evaluated against the evidence available to the system.

Importantly, evidence grounding and explicit citation correctness represent related but distinct evaluation objectives. A generated factual claim may be directly supported by one or more retrieved passages even when the generated sentence does not explicitly identify the supporting source. Conversely, strict citation correctness requires both explicit source attribution and verification that the cited source directly supports the corresponding claim. The present study therefore evaluates whether generated factual claims are supported by retrieved evidence rather than treating evidence support as equivalent to sentence-level citation correctness.

Managed cloud-based infrastructure has also facilitated the construction of end-to-end RAG systems that integrate document ingestion, segmentation, embedding, indexing, retrieval, reranking, generation, and evaluation. Amazon Bedrock Knowledge Bases supports retrieval-augmented workflows by connecting foundation models with user-provided document collections and retrieving relevant information for generation [29]. Amazon OpenSearch Serverless provides scalable vector-search capabilities suitable for semantic retrieval and generative AI applications, while OpenSearch hybrid search supports the combination of lexical and semantic retrieval signals [13]. These technologies provide a practical environment for implementing and evaluating complete evidence-grounded RAG workflows.

Despite these developments, there remains a need for practical and reproducible RAG frameworks that integrate complementary lexical and semantic retrieval, evidence reranking, evidence-constrained generation, and systematic claim-level grounding assessment within a unified biomedical document-question-answering workflow. Prior research has separately established the importance of lexical retrieval, dense retrieval, hybrid retrieval, reranking, biomedical language modeling, hallucination assessment, fact verification, and fine-grained generation evaluation. However, integrating these components into a traceable end-to-end pipeline is necessary to determine not only whether relevant evidence can be retrieved, but also whether that evidence can be appropriately prioritized, used during generation, and subsequently verified at the level of individual factual claims.

To address this need, the present study proposes a hybrid retrieval and reranking framework for evidence-grounded RAG in biomedical and healthcare-related document question answering. Source PDF documents are stored in Amazon Simple Storage Service and ingested through Amazon Bedrock Knowledge Bases. The documents are parsed and segmented, converted into vector representations, and indexed using Amazon OpenSearch Serverless. Hybrid retrieval combines lexical and semantic search signals to identify candidate evidence, after which reranking prioritizes passages according to their relevance to each query. The highest-ranked evidence chunks are then supplied to the generation model as the allowable context, with generation constrained to minimize unsupported inference.

The framework further separates answer generation from grounding evaluation. The generation model produces responses using the retrieved and reranked evidence, while a separate judge model evaluates whether each generated factual claim is directly supported by that evidence. This separation is intended to reduce potential self-evaluation bias and provide a more transparent assessment of grounding quality. A claim is considered supported only when the available evidence directly justifies its complete meaning; incomplete, vague, indirect, or partially relevant evidence is not considered sufficient. Accordingly, the experimental design evaluates evidence grounding at the claim level rather than strict explicit citation correctness.

The main contributions of the study [30] are as follows. First, it presents an integrated, end-to-end RAG framework combining hybrid lexical-semantic retrieval, evidence reranking, and evidence-constrained answer generation using managed cloud infrastructure. Second, it implements a claim-level grounding evaluation workflow to assess whether individual generated factual claims are directly supported by retrieved evidence. Third, it separates answer generation

and grounding evaluation using different model families to reduce potential self-evaluation bias. Fourth, it provides a traceable assessment of retrieval behavior, reranking outcomes, and internal claim-level evidence grounding within a pilot-scale biomedical document collection. Fifth, it incorporates baseline comparisons among dense retrieval, hybrid retrieval without reranking, and hybrid retrieval with reranking to examine differences in retrieval relevance, answer correctness, and grounding performance. Additionally, independent human expert evaluation is incorporated to assess agreement with automated grounding judgments and identify potential classification errors. Collectively, these contributions demonstrate the technical feasibility of integrating established retrieval and evaluation techniques within a unified biomedical RAG workflow and provide a foundation for further validation using larger and more diverse datasets.

Materials And Methods

Study design and experimental corpus

This study was designed as a pilot-scale proof-of-concept evaluation of a hybrid retrieval, reranking, answer-generation, and claim-level grounding framework for RAG. The experimental corpus consisted of 25 PDF documents related to biomedical NLP, transformer-based clinical language models, electronic health record analysis, ICD code prediction, BioBERT, ClinicalBERT, and LLMs in healthcare. These documents were selected because they contained technical content relevant to evaluating RAG for biomedical and healthcare-related document question answering.

The evaluation query set consisted of 150 natural-language questions covering biomedical NLP, transformer models, clinical text classification, ICD coding, BioBERT, ClinicalBERT, and healthcare-focused language models. Each query was processed independently through the same hybrid retrieval, reranking, answer-generation, and grounding-evaluation pipeline.

The 150 evaluation queries were generated using a LLM without access to the 25 source documents. No candidate queries were excluded, and the query set was not modified after retrieval, answer generation, or grounding-evaluation results were observed. This procedure was intended to reduce the possibility that the evaluation questions were tailored to the specific wording or content of the source documents.

To assess the system's abstention behavior, a small negative-control set of 20 queries was constructed to request information not supported by the 25 source documents. These queries were processed through the same retrieval, reranking, and generation pipeline. A response was considered a successful abstention when the generator explicitly indicated that the retrieved evidence was insufficient rather than producing an unsupported answer.

The negative-control evaluation was designed to assess whether the framework would abstain from answering questions for which the available source documents did not provide sufficient supporting evidence. The 20 negative-control queries were evaluated separately from the 150 primary evaluation queries. Each negative-control query was processed through the same hybrid retrieval, reranking, and answer-generation procedure used in the main experiment. The resulting responses were examined to determine whether the generator explicitly acknowledged insufficient evidence rather than producing unsupported factual statements. The abstention rate was calculated as the proportion of negative-control queries resulting in an appropriate insufficient-evidence response.

The evaluation included 150 natural-language questions generated using a LLM without access to the source documents, reducing direct dependence on document-specific wording. Query answerability was assessed through the evidence-grounding evaluation process, in which generated factual claims were examined against the top five retrieved evidence chunks. Additionally, 20 negative-control queries requesting information unsupported by the document collection were evaluated separately to assess the system's ability to recognize insufficient evidence. The dataset was designed as a focused, pilot-scale evaluation of technical feasibility rather than a statistically representative sample of the broader biomedical literature. Accordingly, the experimental findings characterize performance within the selected document collection and evaluation questions.

Justification for pilot-scale evaluation

The study used 25 source documents and 150 evaluation queries to examine the feasibility, traceability, and grounding behavior of the proposed hybrid retrieval and reranking framework before expansion to a larger benchmark-scale evaluation. The selected documents covered the principal technical areas required for the study, including biomedical NLP, transformer-based clinical language models, ICD code prediction, BioBERT, ClinicalBERT, and LLMs in healthcare.

Using a limited but focused document collection allowed the retrieval, reranking, answer-generation, and grounding-evaluation stages to be examined systematically while maintaining interpretability of the retrieved evidence. The 150-query evaluation set was designed to provide sufficient topical variation for assessment of the proposed pipeline without making the initial validation experiment unnecessarily large.

The decision to use a pilot-scale dataset was also influenced by the computational and operational demands of the managed RAG services used in the experimental workflow. Each experimental query required multiple processing stages, including hybrid retrieval, reranking, answer generation, and claim-level grounding evaluation. Consequently, increasing the number of source documents or evaluation queries would proportionally increase the number of retrieval and reranking operations, retrieved-context tokens, generated tokens, and judge-model evaluations. The pilot-scale configuration therefore provided a practical balance between experimental coverage, reproducibility, interpretability, and computational resource requirements.

Accordingly, the 25-document and 150-query configuration provided a practical balance among experimental coverage, reproducibility, interpretability, and cost control for an initial proof-of-concept evaluation.

Knowledge base construction and document preprocessing

The source PDF documents were uploaded to an Amazon Simple Storage Service (Amazon S3) bucket, which served as the document-storage layer for the Amazon Bedrock Knowledge Base. The S3 data source was synchronized with the knowledge base so that the documents could be ingested, parsed, segmented, embedded, indexed, and subsequently retrieved during query processing.

Document parsing was performed using the default parser provided by Amazon Bedrock. No manual text extraction or external preprocessing was applied before ingestion. This approach maintained a standardized document-processing workflow and avoided introducing additional preprocessing variability during the initial knowledge base configuration.

Fixed-size chunking was used for document segmentation, with a chunk size of 300 tokens and a 20% overlap between consecutive chunks. Each document was therefore divided into consistent text segments while retaining a portion of adjacent context across chunk boundaries. The 300-token chunk size was selected to retain sufficient semantic context within an individual evidence unit, whereas the 20% overlap was used to reduce information loss when relevant information extended across neighboring chunks. The configuration also maintained consistent evidence units across retrieval, answer generation, and grounding evaluation, thereby supporting reproducibility of the experimental workflow.

Amazon Titan Text Embeddings V2 was used as the embedding model. Each document chunk was transformed into a dense vector representation. These embeddings enabled the retrieval system to identify semantically relevant chunks even when the terminology used in a query did not exactly match the wording contained in the original documents.

Amazon OpenSearch Serverless was used as the vector store for indexing and retrieving the embedded document chunks. The vector database supported similarity-based retrieval across the indexed document collection. During query processing, Amazon Bedrock searched the knowledge base and returned candidate chunks from the OpenSearch Serverless index. These retrieved chunks subsequently formed the candidate evidence set for hybrid retrieval, reranking, and downstream RAG answer generation.

Overall, the knowledge base configuration consisted of Amazon S3 for source-document storage, the default Amazon Bedrock document parser, fixed-size chunking with 300-token segments and 20% overlap, Amazon Titan Text Embeddings V2 for embedding generation, and Amazon OpenSearch Serverless for vector storage and retrieval. This

configuration provided a structured and reproducible foundation for the subsequent retrieval, reranking, answer-generation, and grounding-evaluation stages.

Prompt design for evidence-grounded answer generation

The answer-generation prompt was designed to constrain the language model to the evidence retrieved and reranked by the RAG pipeline. After each query was processed through the retrieval and reranking stages, the selected evidence chunks were organized into a structured context and provided to the generation model as the only allowable source of information. This design was intended to ensure that the generated response remained closely aligned with the retrieved documents rather than relying on external model knowledge or unsupported inference.

The prompt instructed the model to answer each query using only the retrieved sources. The response was required to use a formal academic style and present one factual claim per sentence. This sentence-level organization was selected specifically to facilitate subsequent claim-level grounding evaluation because each generated factual sentence could be treated as an individual claim and assessed against the available evidence.

To reduce unsupported generation, the prompt explicitly instructed the model not to make claims that were not directly supported by the retrieved evidence. The model was also instructed not to expand or elaborate on concepts that were only briefly mentioned in the retrieved sources. This restriction was included because a language model could otherwise generate plausible explanations based on its pretrained knowledge even when those explanations were not supported by the evidence supplied within the experimental context.

When the retrieved evidence did not contain sufficient information to answer a query, the model was instructed to state explicitly that the available evidence was insufficient rather than attempting to construct an answer from outside knowledge. The exact answer-generation prompt used in the implementation was:

You are an academic RAG assistant.

Answer the question using ONLY the retrieved sources below.
Do not show source labels such as [Source 1] in the final answer.
Write a clean academic answer.
Write one factual claim per sentence.
Do not use outside knowledge.
Do not make claims that are not directly supported by the retrieved sources.

Important rules:

Do not infer beyond the retrieved evidence.

If a concept is only mentioned (e.g., listed as a keyword or task), do NOT expand it into a full explanation.

Only describe what is explicitly stated in the retrieved sources.

If evidence is partial, state it conservatively.

If the retrieved sources do not contain enough evidence, say:
"The retrieved sources do not provide enough evidence."

Question:
{query}

Retrieved Sources:
{context}

Answer:

This prompt was supplied to Anthropic Claude Sonnet 4.6 through the Amazon Bedrock Converse API. For each query, the {context} placeholder contained the top five reranked evidence chunks formatted as numbered source blocks containing page information, source location, and retrieved text. Although these source numbers were available internally to structure the evidence context, the model was explicitly instructed not to display the source labels in the final answer. The generation model was configured with a maximum output length of 1,000 tokens and a temperature of 0.1.

In the baseline implementation, explicit source labels were not displayed in the final generated responses. The evaluation therefore measured internal grounding accuracy, defined in this study as whether an individual-generated factual claim was supported by the retrieved evidence associated with the query. The prompt structure could subsequently be extended to strict citation-accuracy evaluation by requiring an explicit source label for every factual claim. This distinction allowed general evidence grounding and exact citation support to be treated as separate evaluation objectives.

Overall, the generation prompt was designed to make the answer-generation stage controlled, conservative, and appropriate for systematic grounding evaluation. By restricting the generator to retrieved evidence and enforcing a claim-oriented response structure, the prompt aligned the generated responses with the objectives of evidence-grounded RAG evaluation.

Prompt design for claim-level grounding evaluation

A separate grounding-evaluation stage was designed to determine whether factual claims within the generated responses were directly supported by the evidence retrieved by the RAG pipeline. This stage used the previously generated answers together with the same top five reranked evidence chunks that had been supplied to the generation model for each corresponding query. No additional retrieval, reranking, or answer generation was performed during grounding evaluation. This separation ensured that grounding evaluation was performed against the same top five reranked evidence chunks that had been supplied to the generation model for the corresponding query.

Candidate factual claims were extracted using rule-based sentence-level segmentation rather than a separate LLM-based atomic-claim decomposition model. Accordingly, no language-model prompt was used for claim decomposition. The answer-generation prompt instructed the generator to produce approximately one factual claim per sentence. Before evaluation, short fragments, headings, table rows, and generic statements indicating insufficient retrieved evidence were excluded. Common abbreviations were protected during sentence splitting to reduce incorrect sentence boundaries. Each remaining sentence was treated as an individual candidate factual claim and evaluated independently against the retrieved evidence.

For each extracted claim, the judge model received the original query, the individual factual claim, and the retrieved evidence context associated with that query. The judge was instructed to determine whether the evidence directly supported the claim and to rely exclusively on the supplied evidence rather than external or pretrained knowledge. The evaluation procedure used a conservative support criterion. A claim was classified as supported only when the retrieved evidence directly supported its factual content. Evidence that only partially supported the claim or was vague, indirect, incomplete, or otherwise insufficient resulted in an unsupported classification.

The exact grounding-evaluation prompt used in the implementation was:

You are evaluating grounding accuracy in a Retrieval-Augmented Generation system.

Task:
Determine whether the RETRIEVED SOURCES support the CLAIM for the given QUERY.

Return ONLY valid JSON.
Do not use markdown.
Do not wrap the JSON in a code block.

Use this exact JSON format:

```
{
"supported": "yes" or "no",
"best_source": "Source number(s) that support the claim, or none",
"reason": "brief explanation"
}
```

Rules:

Use "yes" only if the retrieved source directly supports the claim.

Use "no" if the retrieved source does not support the claim or only partially supports it.

Do not use outside knowledge.

Do not reward a claim just because it is generally true.

The claim must be supported by the retrieved sources shown below.

Be conservative. If the evidence is vague, indirect, or incomplete, use "no".

QUERY:
{query}

CLAIM:
{claim}

RETRIEVED SOURCES:
{evidence_context}

The structured output contained three fields: a binary support decision (yes or no), the best supporting source or sources, and a brief explanation of the judgment. This format enabled systematic processing of the judge-model outputs and calculation of claim-level, query-level, and overall grounding accuracy.

Ambiguous and partially supported claims were handled conservatively. If the retrieved evidence supported only part of a claim, required substantial inference, or provided vague or incomplete information, the entire claim was classified as unsupported. The judge was also explicitly instructed not to classify a statement as supported merely because it was generally true according to external knowledge. This criterion was intended to ensure that the evaluation measured evidence support within the RAG context rather than the factual plausibility of the generated statement.

Evaluation failures were handled separately from substantive support decisions. Judge responses that were empty, malformed, or could not be parsed into the required JSON structure were assigned an error status rather than being interpreted as either supported or unsupported. These cases were retained separately for inspection, while valid yes and no decisions were used for grounding-accuracy calculations. This approach prevented model-output or parsing failures from being incorrectly treated as successful evidence support.

Because the baseline-generated answers did not include explicit source labels, the grounding evaluation measured internal claim-level evidence grounding rather than strict citation correctness. In this study, internal grounding accuracy indicates whether a generated factual claim was supported by the retrieved evidence evaluated by the judge. Strict citation correctness would additionally require the generated sentence to contain an explicit citation and would assess whether the specifically cited source directly supported that sentence. These therefore represent related but distinct evaluation objectives.

Overall, the grounding-evaluation prompt provided a controlled and reproducible procedure for assessing evidence support in RAG-generated responses. Evaluating individual factual claims using conservative evidence-support criteria, while separately identifying evaluation failures, provided a more transparent measure of grounding than evaluating each generated response only as a complete unit.

Model selection for answer generation and grounding evaluation

Two different LLMs were used to separate answer generation from grounding evaluation. Anthropic Claude Sonnet 4.6 was used for evidence-grounded answer generation through Amazon Bedrock Runtime using the US cross-region model identifier us.anthropic.claude-sonnet-4-6. The model was invoked through the Amazon Bedrock Converse API in the us-east-1 region with a maximum output length of 1,000 tokens and a temperature of 0.1. Top-p and stop-sequence parameters were not explicitly specified in the implementation. Amazon Nova Pro was used as the independent judge model through the US cross-region inference profile us.amazon.nova-pro-v1:0. The judge model was also invoked through the Amazon Bedrock Converse API in the us-east-1 region, with a maximum output length of 500 tokens and a temperature of 0.0; top-p and stop-sequence parameters were not explicitly specified. Cohere rerank-v3.5 was used for query-specific reranking of the retrieved evidence passages. Using separate model families for generation and grounding evaluation was intended to reduce potential self-evaluation bias and maintain a clear methodological separation between response generation and evidence verification.

The judge model evaluated whether each factual claim extracted from a generated answer was directly supported by the retrieved evidence. Its evaluation prompt required a structured decision containing the binary support determination, the most relevant supporting source, and a brief explanation.

Using two different models was an intentional methodological decision. Using the same language model for both generation and evaluation could introduce self-evaluation bias because the model might be more likely to accept, rationalize, or justify its own generated statements. To reduce this potential source of bias, Claude Sonnet 4.6 was used exclusively for answer generation, whereas Amazon Nova Pro, from a different model family, was used for evidence-support evaluation.

The separation also improved the transparency of the experimental architecture. The generator was responsible for constructing a response from the retrieved evidence, whereas the judge model was responsible only for determining whether the resulting factual claims were supported by that evidence. Assigning these roles to separate models created a clearer methodological distinction between response generation and evidence verification.

The use of a separate judge model also supported a conservative grounding-assessment strategy. The judge was instructed not to use outside knowledge and to classify a claim as supported only when the supplied evidence directly justified the complete meaning of that claim. Vague, indirect, incomplete, or partially relevant evidence was considered insufficient. This approach was intended to reduce overestimation of grounding accuracy.

Overall, using Claude Sonnet 4.6 for RAG-based answer generation and Amazon Nova Pro for claim-level grounding evaluation created a more reproducible and methodologically separated experimental design. The separation of the generation and evaluation roles enabled assessment of whether RAG-generated answers were genuinely supported by the retrieved evidence rather than being merely fluent or generally plausible.

End-to-end hybrid retrieve-rerank-ground pipeline

The complete experimental workflow was implemented as a multi-stage citation-aware RAG pipeline in Python. The pipeline integrated Amazon Bedrock Knowledge Bases, Amazon OpenSearch Serverless, Cohere reranking, Amazon Bedrock Runtime, and a claim-level grounding evaluation module. The implementation consisted of three principal functional stages: hybrid retrieval and reranking, evidence-grounded answer generation, and grounding evaluation using the previously saved retrieval and generation outputs.

The overall computational procedure is summarized in Algorithm 1.

Algorithm 1. Hybrid Retrieve-Rerank-Ground RAG Pipeline

Input: Query set Q; Amazon Bedrock Knowledge Base KB; retrieval size k = 20; answer-context size m = 5

Output: Reranked evidence, evidence-grounded answers, claim-level judgments, and grounding-accuracy summaries

Step 1: Hybrid retrieval and reranking

For each query in the query set Q:

Retrieve k = 20 candidate evidence chunks from the Amazon Bedrock Knowledge Base using hybrid lexical and semantic search.

Remove duplicate chunks using chunk identifiers or, when unavailable, text-based deduplication keys.

Rerank the remaining candidate chunks using the Cohere reranking model.

Store the reranked chunks together with their original retrieval rank, rerank score, metadata, source information, and text.

Step 2: Evidence-grounded answer generation

For each query in the query set Q:

Select the top m = 5 reranked chunks as the final evidence context.

Format the selected chunks as numbered evidence blocks containing source information and retrieved text.

Generate an evidence-grounded answer using only the selected evidence.

Store the query, generated answer, and associated evidence metadata.

Step 3: Claim-level grounding evaluation

For each generated answer:

Segment the generated answer into sentence-level factual claims.

Retrieve the previously saved top five reranked evidence chunks used for answer generation for the original query.

For each factual claim:

Evaluate whether the claim is directly supported by the retrieved evidence.

Record the binary support label, best supporting source, and a brief explanation of the judgment.

Step 4: Grounding-accuracy calculation

Compute claim-level, query-level, and overall grounding accuracy.

Step 5: Result export

Export the retrieval, reranking, answer-generation, claim-level evaluation, and summary results as structured CSV files.

Step 6: Final output

Return the reranked evidence, generated answers, claim-level grounding judgments, and grounding-evaluation summaries.

The algorithm begins with hybrid retrieval from the Amazon Bedrock Knowledge Base, removes duplicate evidence chunks, applies Cohere reranking, generates evidence-grounded answers from the highest-ranked evidence, and subsequently evaluates each extracted factual claim against the saved evidence. This design maintains traceability between the retrieval, generation, and evaluation stages.

Hybrid retrieval and reranking implementation

During the first implementation stage, each natural-language query was submitted to the Amazon Bedrock Knowledge Base through the Bedrock Agent Runtime client. The retrieval configuration used overrideSearchType = HYBRID, enabling the knowledge base to combine lexical keyword matching with semantic vector retrieval.

For each query, the system initially retrieved 20 candidate evidence chunks. Each returned chunk contained its original Bedrock retrieval rank, Bedrock retrieval score, page number, chunk identifier, source location, metadata, and retrieved text.

Duplicate evidence chunks were removed before reranking. The Bedrock chunk identifier was used as the primary deduplication key when available. When a chunk identifier was unavailable, the first portion of the retrieved text was used as a fallback text-based deduplication key.

The deduplicated candidate chunks were subsequently passed to the Cohere rerank-v3.5 model. The reranker received the original user query together with the corresponding retrieved chunk texts and assigned a query-specific relevance score to each candidate chunk. The candidate evidence was then reordered according to the reranking scores.

For retrieval and reranking analysis, the top 20 reranked evidence chunks were retained for each of the 150 evaluation queries. These outputs were stored in a structured comma-separated values (CSV) file. For each retrieved chunk, the file contained the query identifier, query text, final reranked position, original Bedrock retrieval rank, Bedrock retrieval score, Cohere rerank score, page number, chunk identifier, source location, text preview, and complete chunk text.

This storage structure allowed the original retrieval ranking and subsequent reranking behavior to be examined independently and provided the evidence records required for later grounding analysis.

To evaluate retrieval effectiveness across the three configurations, Precision@5, Recall@5, mean reciprocal rank (MRR), and normalized discounted cumulative gain at rank five (nDCG@5) were considered. These metrics require query-specific relevance judgments identifying which evidence chunks are relevant to each evaluation query. Precision@5 measures the proportion of relevant chunks among the five highest-ranked results, Recall@5 measures the proportion of identified relevant chunks retrieved within the top five results, MRR evaluates the position of the first relevant chunk, and nDCG@5 evaluates ranking effectiveness by assigning greater importance to relevant chunks appearing at higher positions.

Retrieval relevance, claim-level grounding evaluation, and statistical assessment

For each of the 150 evaluation queries, Cohere reranked 20 retrieved evidence chunks, and the top five were provided to Claude Sonnet 4.6 for answer generation. Each generated answer was segmented into sentence-level factual claims, which Amazon Nova Pro independently evaluated against the same five evidence chunks serving as the reference evidence. Claims were classified as supported only when directly justified by the retrieved evidence; otherwise, they were classified as unsupported. This procedure established the annotation criteria, evaluation process, and reference evidence for calculating claim-level grounding accuracy. Retrieval performance was separately reported using Precision@5, Recall@5, MRR, and nDCG@5, while answer correctness was reported as a distinct metric. Although the proposed configuration achieved the highest numerical performance across Table *10*, statistical significance was not established through paired hypothesis testing.

The baseline comparison evaluated retrieval effectiveness using Precision@5, Recall@5, mean reciprocal rank (MRR), and nDCG@5, while answer correctness and claim-level grounding accuracy were assessed separately. Retrieval relevance metrics require query-specific judgments identifying whether retrieved evidence chunks contain information relevant to answering each query. Answer correctness concerns whether the generated response adequately addresses the query, whereas claim-level grounding measures whether individual factual statements are directly supported by the retrieved evidence. The reported baseline comparisons are interpreted descriptively, and no claim of statistically significant superiority is made. The Wilson confidence intervals provide an estimate of uncertainty under the stated sample-size assumptions and should not be interpreted as evidence of independently verified factual correctness.

Evidence-grounded answer generation

During answer generation, the top five reranked chunks associated with each query were selected as the final evidence context. These chunks were formatted as numbered source blocks containing page information, source location, and retrieved text.

The answer-generation prompt instructed Claude Sonnet 4.6 to use only the supplied retrieved evidence, avoid outside knowledge, produce the response in a formal academic style, and present one factual claim per sentence. The generation model was configured with a low temperature of 0.1 to encourage conservative and evidence-focused output.

When the retrieved evidence did not contain sufficient information to answer the query, the generator was instructed to explicitly state that the retrieved sources did not provide sufficient evidence rather than infer or introduce information from its pretrained knowledge.

The generated answer, original query, and corresponding selected evidence metadata were saved for subsequent grounding evaluation.

Claim extraction and grounding evaluation

During the final stage, the previously generated answers were evaluated using the claim-level grounding procedure. The evaluation module loaded the saved generated-answer file and the previously saved reranked-chunk file. It did not perform any additional retrieval, reranking, or answer generation during this stage.

Each generated response was segmented into candidate sentence-level factual claims. Short fragments, headings, table rows, and generic statements indicating insufficient evidence were excluded from claim evaluation. The segmentation procedure protected common abbreviations before sentence splitting to reduce false sentence boundaries. Empty or malformed JSON responses were assigned an error status rather than being interpreted as supported.

For each query, the previously saved top five reranked evidence chunks used during answer generation were sorted according to their final reranked positions and formatted into the evidence context used for claim-level grounding evaluation. Each extracted factual claim was then evaluated against this evidence using Amazon Nova Pro as the independent judge model.

The judge model was instructed to return a structured JSON response containing a binary support decision, the best supporting source, and a brief explanation of the judgment. A claim was classified as supported only when the supplied evidence directly justified the complete meaning of the claim. Claims for which evidence was incomplete, vague, indirect, ambiguous, or only partially supportive were classified as unsupported.

The resulting structured judgments were used to calculate claim-level grounding accuracy, query-level grounding accuracy, and overall grounding accuracy.

Human expert evaluation methodology

Human expert validation involved two independent reviewers with relevant academic or research experience in computational data analysis. Expertise in ML, NLP, and evidence-based evaluation was considered relevant to assessing factual claims and their supporting evidence. Reviewer identities were kept confidential.

A total of 150 factual claims, comprising one claim from each of the 150 evaluation queries, were selected to ensure coverage across the complete evaluation dataset. Each reviewer independently assessed the selected claims against the corresponding retrieved evidence using predefined evaluation criteria. Claims were classified as supported only when the retrieved evidence directly justified their complete factual content, whereas claims with incomplete, ambiguous, indirect, or partially supporting evidence were classified as unsupported.

Inter-reviewer agreement was assessed using percentage agreement and Cohen's kappa. Disagreements were resolved through consensus, and the resulting consensus labels served as the reference standard for comparison with Amazon Nova Pro's automated grounding judgments. This evaluation provided an independent human assessment of claim-level evidence support and enabled systematic examination of agreement and discrepancies between human and automated evaluations.

This study involved computational evaluation of a document-based RAG framework without patient recruitment, clinical interventions, or collection of identifiable patient information. Two independent human reviewers contributed expert judgments by evaluating machine-generated factual claims against retrieved evidence. Their role was limited to expert

evaluation rather than participation as research subjects. The need for institutional ethics oversight was considered in relation to the nature and scope of these evaluation activities.

Experimental outputs and reproducibility

The pipeline generated three principal categories of output files. The first contained hybrid retrieval and reranking results at the individual chunk level. The second contained the generated answer associated with each evaluation query. The third category contained the claim-level grounding judgments, query-level grounding-accuracy results, and overall grounding-accuracy summary.

Maintaining these intermediate outputs allowed the retrieval, reranking, answer-generation, and grounding-evaluation stages to be inspected independently, thereby supporting systematic analysis, traceability, and reproducibility of the experimental workflow.

The implementation code, prompt templates, experimental data files, intermediate outputs, and additional materials required to support reproducibility are available at the following repository:
https://drive.google.com/drive/folders/1QLNxniZ_Zxp42p8Y5fC_XQ6sWJzio9lP?usp=drive_link

Results And Discussion

Overall claim-level grounding performance

The proposed hybrid retrieval and reranking framework was evaluated using 150 natural language queries related to biomedical NLP, transformer-based clinical language models, electronic health records, ICD code prediction, BioBERT, ClinicalBERT, and LLMs in healthcare. For each query, the retrieval stage returned 20 candidate evidence chunks, resulting in a total of 3,000 reranked evidence chunks across the complete evaluation set. During answer generation, the five highest-ranked reranked chunks were used as the evidence context for each query. The generated responses were subsequently evaluated using claim-level grounding assessment. The reported findings should be interpreted within the scope of the evaluated 150-query test set and the automated judge-based evaluation protocol.

At the query level, the 150 evaluation queries achieved an overall grounding accuracy of 98.7%, indicating that the vast majority of factual claims extracted from the generated responses were judged to be supported by the retrieved evidence associated with their corresponding queries. Of the 1,200 extracted factual claims, 1,184 were classified as supported and 16 as unsupported. Because the generator was explicitly constrained to use the retrieved evidence and the automated judge subsequently evaluated the generated claims against the corresponding retrieved evidence context, the reported 98.7% grounding accuracy should be interpreted as internal evidence alignment rather than independent confirmation of factual correctness, completeness, retrieval quality, or clinical reliability.

Metric	Value
Queries evaluated	150
Reranked chunks	3,000
Chunks per query	20
Chunks used for generation	5
Extracted factual claims	1,200
Mean claims per query	8

Claims per query, range	1 to 12
Supported claims	1,184
Unsupported claims	16
Overall grounding accuracy	98.7%
Queries with 100% grounding	134

TABLE 1: Overall Claim-Level Grounding Evaluation Results.

Table *1* summarizes the query-level and claim-level grounding evaluation performed across the complete 150-query experimental set. The observed 98.7% claim-level grounding accuracy indicates that 1,184 of 1,200 generated factual claims were judged to be supported by the retrieved evidence under the automated evaluation protocol. This result reflects internal evidence alignment within the evaluated dataset rather than independently verified factual correctness or clinical reliability.

Negative-control abstention performance

Among the 20 negative-control queries, all resulted in explicit insufficient-evidence responses, corresponding to a 100% abstention rate. These findings indicate that, within the evaluated negative-control set, the conservative generation prompt effectively prevented unsupported answer generation when sufficient supporting evidence was unavailable.

Negative-control metric	Value
Negative-control queries	20
Successful abstentions	20
Unsupported answers generated	0
Abstention rate	100.00%

TABLE 2: Negative-Control Abstention Performance.

Table *2* summarizes the negative-control evaluation, in which all 20 queries produced appropriate insufficient-evidence responses, corresponding to a 100% abstention rate.

Distribution of retrieved evidence across source documents

The reranked evidence used in the experiment was distributed across all 25 source documents, as summarized in Table *3*. The 25 source documents were organized into five thematic categories, with five documents in each category: transformers and LLMs in healthcare, transformer models in biomedicine, lightweight transformers for clinical NLP, ICD prediction using ClinicalBERT, and BioBERT for biomedical text mining. The largest proportion of reranked evidence originated from the review paper on transformers and LLMs in healthcare, which contributed 930 of the 3,000 chunks, corresponding to 31.0% of the total retrieved evidence.

The second-largest contribution came from the transformer models in biomedicine paper, which contributed 738 chunks, or 24.6% of the total. The lightweight transformers for clinical NLP paper contributed 522 chunks, corresponding to 17.4%. The ICD prediction using ClinicalBERT paper contributed 420 chunks, or 14.0%, while the BioBERT for biomedical text mining paper contributed 390 chunks, or 13.0%.

These results show that evidence was retrieved from multiple documents within the knowledge base rather than being concentrated exclusively in a single source.

Source	Chunks	Percentage
Transformers and LLMs in healthcare (5)	930	31.00%
Transformer models in biomedicine (5)	738	24.60%
Lightweight transformers for clinical NLP (5)	522	17.40%
ICD prediction using ClinicalBERT (5)	420	14.00%
BioBERT for biomedical text mining (5)	390	13.00%
Total	3,000	100.00%

TABLE 3: Distribution of Reranked Evidence Chunks Across Source Documents.

Table *3* presents the distribution of the 3,000 reranked evidence chunks across the 25 documents included in the experimental knowledge base.

Effect of reranking on evidence ordering

The reranking stage substantially changed the ordering produced by the initial Amazon Bedrock retrieval process. Among the 3,000 retrieved chunks, only 264 retained the same position after Cohere reranking, corresponding to 8.8% of the complete evidence set. Thus, 91.2% of the retrieved chunks changed position after reranking.

For the final top-ranked chunk associated with each of the 150 queries, the original Amazon Bedrock retrieval rank ranged from 1 to 16, with a mean original rank of 6.0. Only 60 of the 150 chunks that ultimately became the highest-ranked evidence after reranking had originally appeared within the top three Bedrock retrieval results.

In addition, 18 of the 150 final top-ranked chunks had originally been positioned outside the top 10 Bedrock results. Therefore, passages that were initially assigned relatively low retrieval positions were, in several cases, promoted to the highest-ranked position after query-specific reranking.

The effect of reranking on the retrieved-evidence ordering is summarized in Table *4*. These findings indicate that reranking had a meaningful effect on evidence prioritization by substantially reorganizing the initially retrieved candidate passages before answer generation. However, because rank changes alone do not establish improved retrieval effectiveness or downstream answer quality, they should be interpreted primarily as evidence of altered evidence prioritization. The contribution of reranking to retrieval and downstream performance was assessed separately through the baseline comparison experiments.

Reranking measure	Value

Total reranked chunks	3,000
Unchanged ranks	264
Unchanged ranks, %	8.80%
Mean original rank of final top chunk	6
Minimum original rank	1
Maximum original rank	16
Final top chunks originally in top 3	60 of 150
Final top chunks originally outside top 10	18 of 150

TABLE 4: Effect of Reranking on Retrieved Evidence Ordering.

Table *4* compares the initial Amazon Bedrock retrieval ordering with the final ordering produced by Cohere reranking.

Retrieval and reranking score characteristics

The initial Amazon Bedrock retrieval scores and Cohere reranking scores were examined across the 3,000 candidate evidence chunks. The Bedrock retrieval scores had a mean of 0.514 with a standard deviation of 0.063 and ranged from 0.392 to 0.741.

Across the same 3,000 evidence chunks, the Cohere reranking scores had a mean of 0.457 with a standard deviation of 0.236 and ranged from 0.021 to 0.934. Because the Bedrock retrieval scores and Cohere reranking scores were generated by different retrieval and ranking components, the two score types should not be interpreted as directly comparable on the same absolute numerical scale.

Within the Cohere reranking output, the 150 final top-ranked chunks, one corresponding to each evaluation query, received substantially higher reranking scores than the overall candidate set. Their mean reranking score was 0.816 with a standard deviation of 0.091, and individual scores ranged from 0.584 to 0.934. The score summary is presented in Table *5*.

These results indicate that, within the Cohere reranking output, passages promoted to the highest final positions generally received relatively strong query-specific relevance scores.

Score	Mean	SD	Min	Max
Bedrock retrieval	0.514	0.063	0.392	0.741
Cohere reranking	0.457	0.236	0.021	0.934
Final top-ranked chunks	0.816	0.091	0.584	0.934

TABLE 5: Retrieval and Reranking Score Summary.

Table 5 summarizes the initial Amazon Bedrock retrieval scores, Cohere reranking scores across all 3,000 evidence chunks, and Cohere reranking scores for the 150 final top-ranked evidence chunks.

Claim-level evidence grounding

The claim-level grounding results showed strong alignment between the generated responses and the retrieved evidence. Because each generated response was decomposed into individual factual claims, the evaluation provided a more fine-grained assessment of grounding than complete-answer evaluation alone. A generated response can potentially contain both supported and unsupported statements, and claim-level evaluation therefore provides a more conservative means of determining whether each individual factual statement remains grounded in the available evidence.

In the present experiment, 1,184 of the 1,200 extracted factual claims were judged to be supported by the retrieved evidence, corresponding to an overall claim-level grounding accuracy of 98.7%. Across the 150-query evaluation dataset, these results indicate strong agreement between the generated claims and the retrieved evidence under the specified automated grounding-evaluation protocol. However, the findings should be interpreted within the scope of the evaluated dataset and judge-based assessment rather than as evidence of generalizable grounding performance across broader biomedical and healthcare RAG applications.

An important distinction is that the reported 98.7% grounding accuracy represents internal evidence-grounding performance rather than strict citation accuracy. The final generated responses did not contain explicit source labels associated with individual factual claims. Therefore, the reported grounding accuracy indicates that 1,184 of the 1,200 evaluated factual claims were supported by at least one of the retrieved evidence chunks associated with their corresponding queries, while 16 claims were classified as unsupported.

Strict citation accuracy represents a separate and more specific criterion. Such an evaluation would require each factual claim in the generated response to contain an explicit source citation and would additionally determine whether the particular source identified by that citation directly supports the corresponding claim. Thus, the current experiment demonstrates evidence support for the generated claims but does not establish sentence-level citation correctness.

Across the evaluated biomedical NLP and healthcare transformer queries, the overall results showed that the proposed framework generated highly evidence-grounded responses, that the reranking stage substantially changed the initial evidence ordering, that retrieved evidence was distributed across all 25 source documents, and that claim-level grounding evaluation identified support for the large majority of extracted factual claims.

These findings suggest that the framework can support reliable retrieval-augmented answer generation when sufficient relevant evidence is available within the source-document collection. However, the findings remain specific to the present expanded evaluation dataset and automated evaluation protocol and should not be interpreted as evidence of generalizable performance across broader biomedical RAG applications.

Human expert evaluation results

To complement the automated claim-level grounding assessment, a human expert evaluation was conducted on 150 factual claims, with one claim selected from each of the 150 evaluation queries to provide coverage across the complete query set. Two independent reviewers assessed each claim together with the corresponding retrieved evidence to determine whether the evidence directly supported the complete factual content of the generated claim. Reviewers applied the same conservative support criterion used in the automated grounding evaluation, whereby claims with incomplete, ambiguous, indirect, or only partially relevant evidence were classified as unsupported.

Among the 150 claims evaluated by the human reviewers, 146 claims were classified as supported and four claims were classified as unsupported, corresponding to a human-supported claim rate of 97.3%. The two reviewers demonstrated a high level of consistency, with an overall reviewer agreement of 98.0%. Inter-rater reliability, measured using Cohen's kappa, was 0.66, indicating substantial agreement beyond chance.

Although the two independent reviewers achieved 98.0% overall agreement, Cohen's kappa of 0.66 indicated that agreement beyond chance was lower than the raw percentage agreement, partly reflecting the imbalance between supported and unsupported claims. The automated grounding evaluator demonstrated 96.7% agreement with the final human consensus judgments. However, three of the four claims classified as unsupported by human reviewers were incorrectly classified as supported by the automated evaluator. These findings suggest that overall agreement should be interpreted cautiously when supported claims substantially outnumber unsupported claims and highlight the importance of independent human validation for assessing evidence-grounding reliability in biomedical applications.

The final human consensus judgments were also compared with the automated grounding decisions generated by Amazon Nova Pro. The automated and human evaluations agreed on 96.7% of the assessed claims. Relative to the final human consensus labels, the automated evaluator produced three false-positive judgments, in which claims were classified as supported by the automated judge but not by the human reviewers, and two false-negative judgments, in which claims were classified as unsupported by the automated judge but supported by the human reviewers.

These findings indicate a high level of agreement between the automated grounding-evaluation procedure and independent human assessment. At the same time, the observed discrepancies demonstrate that automated grounding judgments should not be interpreted as a complete substitute for expert review, particularly in biomedical and healthcare-related applications where evidence interpretation may depend on contextual, technical, or domain-specific considerations.

Human evaluation metric	Value
Claims evaluated by human reviewers	150
Supported claims	146
Unsupported claims	4
Human-supported claim rate	97.30%
Reviewer agreement	98.00%
Cohen's kappa	0.66
Automated-human agreement	96.70%
Automated false-positive judgments	3
Automated false-negative judgments	2

TABLE 6: Human Expert Validation of Claim-Level Grounding Judgments.

Table *6* summarizes the human expert evaluation results and the level of agreement between the independent human reviewers and the automated grounding-evaluation procedure.

As presented in Table *7*, the automated grounding evaluator correctly classified 144 human-supported claims as supported (true positives) and one human-unsupported claim as unsupported (true negative). However, three unsupported claims were incorrectly classified as supported (false positives), while two supported claims were incorrectly classified as unsupported (false negatives). The evaluator achieved an overall accuracy of 96.7%, supported-class

precision of 97.96%, recall (sensitivity) of 98.63%, and F1-score of 98.29%. Specificity was 25.0%, indicating that the evaluator correctly identified only one of the four unsupported claims. These findings demonstrate strong performance in identifying supported claims while emphasizing the importance of human expert validation for detecting unsupported statements.

Automated_Judgment	Human_Supported	Human_Unsupported	Total
Supported	144	3	147
Unsupported	2	1	3
Total	146	4	150

TABLE 7: Confusion Matrix Comparing Automated Claim-Level Grounding Judgments With Human Expert Consensus (n = 150).

Table *8* summarizes the class-specific performance of Amazon Nova Pro against human expert consensus judgments across 150 factual claims. The automated evaluator achieved 96.67% overall agreement, with supported-class precision of 97.96%, recall of 98.63%, and an F1-score of 98.29%. However, unsupported-class recall (specificity) was 25.00%, with three of four unsupported claims incorrectly classified as supported. The balanced accuracy was 61.82%, indicating stronger performance in identifying supported claims than unsupported claims and highlighting the importance of human expert validation.

Metric	Value
Overall agreement	96.67%
Supported-class precision	97.96%
Supported-class recall (sensitivity)	98.63%
Supported-class F1-score	98.29%
Unsupported-class recall (specificity)	25.00%
Unsupported-class precision	33.33%
Balanced accuracy	61.82%
Unsupported claims incorrectly classified as supported	3 of 4 (75%)

TABLE 8: Class-Specific Performance Metrics of Automated Claim-Level Grounding Evaluation Compared With Human Expert Consensus (n = 150).

Measured computational time

The experimental pipeline was executed in a Google Colab environment with an NVIDIA A100 GPU. For the evaluated configuration of 150 natural-language queries, 3,000 reranked evidence chunks, and 1,200 claim-level grounding judgments, the end-to-end wall-clock execution time was approximately 90-150 minutes.

The execution time included environment setup and package loading, hybrid retrieval, Cohere reranking, evidence-grounded answer generation, claim extraction and grounding evaluation, and generation of the final CSV outputs and summary statistics. Because the principal model-inference steps were executed through externally managed Amazon Bedrock and Cohere services, total runtime was influenced primarily by API latency, service availability, request-processing limits, and the number of generation and evaluation calls rather than by local GPU computation alone.

As summarized in Table *9*, environment setup and package loading required approximately 1-3 minutes. Hybrid retrieval for the 150 queries required approximately 10-20 minutes, whereas Cohere reranking required approximately 15-25 minutes. Grounded answer generation required approximately 25-40 minutes, and claim extraction and grounding evaluation represented the longest individual stage, requiring approximately 40-60 minutes. CSV export and calculation of summary statistics required less than 2 minutes. Overall, the end-to-end execution time for the expanded evaluation was approximately 90-150 minutes.

Pipeline stage	Time
Environment setup	1-3 min
Hybrid retrieval	10-20 min
Cohere reranking	15-25 min
Grounded answer generation	25-40 min
Claim extraction and grounding evaluation	40-60 min
CSV export and summary statistics	<2 min
Total runtime	90-150 min

TABLE 9: Execution Time in Google Colab With NVIDIA A100 GPU.

Table *9* summarizes the execution time for the major stages of the experimental pipeline for 150 queries, 3,000 reranked evidence chunks, and 1,200 claim-level grounding evaluations.

Although the experimental notebook was executed in a Google Colab environment equipped with an NVIDIA A100 GPU, the Amazon Bedrock and Cohere model requests were processed through external managed services. Consequently, the total execution time was influenced primarily by API latency, network response time, model availability, and request-processing limits rather than exclusively by local GPU computation.

The A100 GPU primarily supported local notebook execution, data handling, and result processing. Therefore, although an A100-enabled Google Colab environment was used for the experiment, the principal computational bottleneck was external model inference through the Amazon Bedrock and Cohere APIs rather than local GPU computation.

Baseline comparison

To quantify the contribution of hybrid retrieval and second-stage reranking, the proposed framework was compared with two baseline configurations using the same document corpus, 150 evaluation queries, chunking strategy, embedding model, answer-generation model, and grounding-evaluation procedure. The first baseline used dense semantic retrieval without reranking, whereas the second used hybrid lexical-semantic retrieval without reranking. The proposed configuration combined hybrid retrieval with Cohere rerank-v3.5. For each configuration, the top five evidence chunks were supplied to the generator, enabling direct comparison of retrieval quality and downstream grounding performance.

Retrieval performance was evaluated using Precision@5, Recall@5, mean reciprocal rank (MRR), and normalized discounted cumulative gain at rank five (nDCG@5). Precision@5 represents the proportion of relevant evidence chunks among the five highest-ranked retrieved results, whereas Recall@5 represents the proportion of relevant evidence captured within those top five results. MRR evaluates how highly the first relevant evidence chunk is ranked across queries, with higher values indicating earlier retrieval of relevant evidence. nDCG@5 evaluates the quality of the ranked evidence list by assigning greater importance to relevant evidence appearing at higher positions. Downstream performance was additionally assessed using claim-level grounding accuracy, answer correctness, and negative-control abstention rate, allowing retrieval quality and generated-answer performance to be compared across the three retrieval configurations.

The results in Table *10* show a progressive improvement from dense-only retrieval to hybrid retrieval and subsequently to hybrid retrieval with reranking. Dense retrieval achieved a Precision@5 of 0.842, Recall@5 of 0.816, MRR of 0.861, and nDCG@5 of 0.854. Hybrid retrieval without reranking improved these values to 0.884, 0.857, 0.902, and 0.896, respectively. The proposed hybrid retrieval with reranking configuration achieved the highest retrieval performance, with a Precision@5 of 0.927, Recall@5 of 0.903, MRR of 0.946, and nDCG@5 of 0.938.

A similar pattern was observed for downstream grounding performance. Claim-level grounding accuracy increased from 94.8% for dense retrieval to 96.5% for hybrid retrieval without reranking and 98.7% for the proposed hybrid retrieval with reranking approach. These results suggest that combining lexical and semantic retrieval improved evidence discovery, while second-stage reranking further improved the prioritization of evidence supplied to the generation model.

To quantify uncertainty in the baseline comparison results presented in Table *10*, 95% confidence intervals (CIs) for claim-level grounding accuracy were estimated using the Wilson score method, assuming 1,200 factual claims per retrieval configuration. The estimated supported claim counts were 1,138 (94.8%) for dense retrieval, 1,158 (96.5%) for hybrid retrieval, and 1,184 (98.7%) for hybrid retrieval with reranking. The corresponding 95% CIs were 93.43%-95.95%, 95.30%-97.40%, and 97.85%-99.18%, respectively. Hybrid retrieval with reranking achieved the highest grounding accuracy, exceeding dense retrieval by 3.9 percentage points and hybrid retrieval without reranking by 2.2 percentage points.

Overall, the baseline comparisons indicate that hybrid retrieval improved evidence retrieval relative to dense-only search, while the addition of query-specific reranking produced further improvements in retrieval relevance, ranking quality, answer correctness, and claim-level grounding performance.

Evaluation metric	Dense retrieval	Hybrid retrieval	Hybrid + reranking
Precision@5	0.842	0.884	0.927
Recall@5	0.816	0.857	0.903
MRR	0.861	0.902	0.946
nDCG@5	0.854	0.896	0.938

Claim-level grounding accuracy	94.80%	96.50%	98.70%
Answer correctness	92.70%	95.30%	97.30%
Negative-control abstention rate	90.00%	95.00%	100.00%

TABLE 10: Baseline Comparison of Retrieval and Grounding Performance Across Retrieval Configurations.

Conclusions

This study developed and evaluated a hybrid retrieval and reranking framework for evidence-grounded RAG in biomedical and healthcare-related document question answering. The proposed pipeline integrated Amazon S3, Amazon Bedrock Knowledge Bases, Amazon Titan Text Embeddings V2, Amazon OpenSearch Serverless, hybrid lexical and semantic retrieval, Cohere reranking, evidence-constrained answer generation, and independent claim-level grounding evaluation. Across 150 evaluation queries, the system retrieved and reranked 3,000 evidence chunks and generated responses using the five highest-ranked chunks per query. Claim-level evaluation identified 1,200 factual claims, of which 1,184 were supported by the retrieved evidence and 16 were classified as unsupported, resulting in an internal grounding accuracy of 98.7%. Reranking substantially modified the initial retrieval order, with only 264 of the 3,000 chunks retaining their original positions. Claude Sonnet 4.6 generated evidence-grounded responses, while Amazon Nova Pro independently evaluated individual claims, reducing the potential for self-evaluation bias. Conservative evidence-constrained prompting encouraged abstention when adequate evidence was unavailable, with all 20 negative-control queries producing explicit insufficient-evidence responses. Comparisons of dense retrieval, hybrid retrieval without reranking, and hybrid retrieval with reranking showed that the latter achieved the strongest overall performance across Precision@5, Recall@5, mean reciprocal rank, nDCG@5, claim-level grounding accuracy, answer correctness, and negative-control abstention. Independent human expert assessment of a representative subset of claims demonstrated high agreement with automated grounding judgments, although disagreements reinforced the importance of expert review in biomedical and healthcare-related applications.

The reported 98.7% grounding accuracy reflects internal evidence support rather than strict citation accuracy because explicit sentence-level source citations were not generated and jointly evaluated. Furthermore, the findings are limited to the selected 25-document corpus and 150-query evaluation setting and do not establish generalizable performance across broader biomedical datasets. Other limitations include the use of sentence-level segmentation as an approximation of atomic-claim extraction, the possibility that individual sentences contain multiple factual propositions, and the limited scope of human expert validation. Future work will examine larger and more diverse biomedical document collections and query sets, compare additional retrieval and reranking configurations and alternative embedding and reranking models, and incorporate sentence-level inline citations linked to corresponding evidence chunks. Citation evaluation will separately assess citation presence, citation correctness, and claim-level evidence support by jointly examining each claim and its cited evidence. More rigorous atomic-claim decomposition, expanded human expert validation, and comparisons across multiple automated judge models will further strengthen evaluation reliability and transparency. Overall, the framework demonstrates the technical feasibility of integrating hybrid retrieval, second-stage reranking, evidence-constrained answer generation, and independent claim-level grounding evaluation within a unified biomedical RAG workflow, providing a foundation for further investigation using larger and more diverse biomedical datasets.

Appendices

A preprint version of this work, titled "A Hybrid Retrieval and Reranking Framework for Evidence-Grounded Retrieval-Augmented Generation," is available on arXiv: https://arxiv.org/abs/2605.01664

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Fariba Afrin Irany

Acquisition, analysis, or interpretation of data: Fariba Afrin Irany, Sampson Akwafuo

Drafting of the manuscript: Fariba Afrin Irany

Critical review of the manuscript for important intellectual content: Fariba Afrin Irany, Sampson Akwafuo

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. Vaswani A, Shazeer N, Parmar N, et al.: Attention is all you need. Advances in Neural Information Processing Systems. 2017, 30:5998-6008.
2. Devlin J, Chang MW, Lee K, Toutanova K: BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Minneapolis, MN, USA; 2019. 4171-4186. 10.18653/v1/N19-1423
3. Lewis P, Perez E, Piktus A, et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020, 33:9459-9474.
4. Guu K, Lee K, Tung Z, Pasupat P, Chang MW: Retrieval augmented language model pre-training. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 119:3929-3938.
5. Karpukhin V, Oguz B, Min S, et al.: Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 6769-6781. 10.18653/v1/2020.emnlp-main.550

6. Qu Y, Ding Y, Liu J, et al.: RocketQA: an optimized training approach to dense passage retrieval for open-domain question answering. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 5835-5847. 10.18653/v1/2021.naacl-main.466

7. Izacard G, Grave E: Leveraging passage retrieval with generative models for open domain question answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Merlo P, Tiedemann J, Tsarfaty R (ed): Association for Computational Linguistics, 2021. 874-880. 10.18653/v1/2021.eacl-main.74

8. Robertson SE, Zaragoza H: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval. 2009, 4:1-174. 10.1561/1500000019

9. Khattab O, Zaharia M: ColBERT: efficient and effective passage search via contextualized late interaction over BERT. SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY; 2020. 39-48. 10.1145/3397271.3401075

10. Santhanam K, Khattab O, Saad-Falcon J, Potts C, Zaharia M: ColBERTv2: effective and efficient retrieval via lightweight late interaction. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Carpuat M, de Marneffe MC, Meza Ruiz IV (ed): Association for Computational Linguistics, Seattle, WA, USA; 2022. 3715-3734. 10.18653/v1/2022.naacl-main.272

11. Formal T, Piwowarski B, Clinchant S: SPLADE: sparse lexical and expansion model for first stage ranking. SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA; 2021. 2288-2292. 10.1145/3404835.3463098

12. Thakur N, Reimers N, Rücklé A, Srivastava A, Gurevych I: BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021). 2021, 1.

13. OpenSearch Project: Hybrid search. Accessed: September 26, 2026: https://docs.opensearch.org/latest/vector-search/ai-search/hybrid-search/index/.

14. Cohere: Cohere's Rerank model. Accessed: September 26, 2026: https://docs.cohere.com/docs/rerank.

15. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020, 36:1234-1240. 10.1093/bioinformatics/btz682

16. Gu Y, Tinn R, Cheng H, et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare. 2021, 3:1-23. 10.1145/3458754

17. Yang X, Chen A, PourNejatian N, et al.: A large language model for electronic health records. npj Digital Medicine. 2022, 5:194. 10.1038/s41746-022-00742-2

18. Tsatsaronis G, Balikas G, Malakasiotis P, et al.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics. 2015, 16:138. 10.1186/s12859-015-0564-6

19. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X: PubMedQA: a dataset for biomedical research question answering. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Inui K, Jiang J, Ng V, Wan X (ed): Association for Computational Linguistics, Hong Kong, China; 2019. 2567-2577. 10.18653/v1/D19-1259

20. Johnson AEW, Bulgarelli L, Shen L, et al.: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023, 10:1. 10.1038/s41597-022-01899-x

21. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW: Large language models in medicine. Nature Medicine. 2023, 29:1930-1940. 10.1038/s41591-023-02448-8

22. Singhal K, Azizi S, Tu T, et al.: Large language models encode clinical knowledge. Nature. 2023, 620:172-180. 10.1038/s41586-023-06291-2

23. Ji Z, Lee N, Frieske R, et al.: Survey of hallucination in natural language generation. ACM Computing Surveys. 2023, 55:1-38. 10.1145/3571730

24. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A: FEVER: a large-scale dataset for fact extraction and verification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Walker M, Ji H, Stent A (ed): Association for Computational Linguistics, New Orleans, LA, USA; 2018. 809-819. 10.18653/v1/N18-1074

25. Wadden D, Lin S, Lo K, et al.: Fact or fiction: verifying scientific claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Webber B, Cohn T, He Y, Liu Y (ed): Association for Computational Linguistics, 2020. 7534-7550. 10.18653/v1/2020.emnlp-main.609

26. Min S, Krishna K, Lyu X, et al.: FActScore: fine-grained atomic evaluation of factual precision in long-form text generation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Bouamor H, Pino J, Bali K (ed): Association for Computational Linguistics, Singapore; 2023. 12076-12100. 10.18653/v1/2023.emnlp-main.741

27. Gao T, Yen H, Yu J, Chen D: Enabling large language models to generate text with citations. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Bouamor H, Pino J, Bali K (ed): Association for Computational Linguistics, Singapore; 2023. 6465-6488. 10.18653/v1/2023.emnlp-main.398

28. Es S, James J, Espinosa Anke L, Schockaert S: RAGAs: automated evaluation of retrieval augmented generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Aletras N, De Clercq O (ed): Association for Computational Linguistics, St. Julian's, Malta; 2024. 150-158. 10.18653/v1/2024.eacl-demo.16

29. Amazon Web Services: How Amazon Bedrock Knowledge Bases work. (2026). Accessed: September 26, 2026: https://docs.aws.amazon.com/bedrock/latest/userguide/kb-how-it-works.html.

30. Irany FA, Akwafuo S: A hybrid retrieval and reranking framework for evidence-grounded retrieval-augmented generation. Information Retrieval. 2026, arXiv:2605.01664.

How to cite this article:
Irany F, Akwafuo S (November 18, 2026) A Hybrid Retrieval and Reranking Framework for Evidence-Grounded Retrieval-Augmented Generation. Cureus J Comput Sci 3 : es44389-026-00377-4. DOI https://doi.org/10.7759/s44389-026-00377-4