Contradiction-Aware Grounded QA System for TREC 2025 BioGen: Lexical Retrieval Trade-offs and Citation Attribution

Soumya Ranjan Sahoo
GE HealthCare
Bangalore, India
Soumya.Sahoo1@gehealthcare.com

Gagan N.
GE HealthCare
Bangalore, India
Gagan.N@gehealthcare.com

Sanand Sasidharan
GE HealthCare
Bangalore, India

Divya Bharti
GE HealthCare
Bangalore, India

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in biomedical question-answering tasks, yet their tendency to generate plausible yet unverified claims poses significant risks in clinical contexts. To mitigate the clinical risks of LLM hallucinations, the TREC 2025 BioGen track mandates grounded answers that explicitly surface contradictory evidence (Task A) and the generation of narrative-driven, fully attributed responses (Task B). Addressing the critical absence of target ground truth, we present a proxy-based development framework utilizing the SciFact dataset to systematically optimize retrieval architectures. Our iterative evaluation revealed a “Simplicity Paradox”: complex adversarial dense retrieval strategies failed catastrophically on contradiction detection (MRR 0.023) due to Semantic Collapse, where negation signals were indistinguishable in vector space. Furthermore, we identified a distinct Retrieval Asymmetry: while filtering dense embeddings improved contradiction detection, it degraded support recall, compromising holistic reliability. We resolve this via a Decoupled Lexical Architecture utilizing a unified BM25 backbone to balance semantic support recall (0.810) with precise contradiction surfacing (0.750). This approach achieves the highest Weighted MRR (0.790) on the proxy benchmark while remaining the only computationally viable strategy for scaling to the 30-million-document PubMed corpus. For answer generation, we introduce Narrative-Aware Reranking and One-Shot In-Context Learning, which improved citation coverage from 50% (zero-shot) to 100%. Our official TREC evaluation results confirm these findings: our system ranks 2nd among all teams on Task A contradiction F1 and 3rd out of 50 runs on Task B citation coverage (98.77%), achieving zero citation contradict rate. Our work transforms LLMs from stochastic generators into honest evidence synthesizers, demonstrating that epistemic integrity in biomedical AI requires prioritizing lexical precision and architectural scalability over isolated metric optimization.

Keywords

Biomedical QA, Retrieval-Augmented Generation, Contradiction Detection, Evidence Attribution, Negation, Proxy Evaluation

1 Introduction

Large language models (LLMs) are increasingly applied to biomedical question answering, where clinicians and researchers expect concise, trustworthy, and verifiable outputs. Yet even state-of-the-art systems can produce confident but unsupported claims, a risk that undermines trust in medical AI. Recent work shows that retrieval-augmented generation (RAG) can reduce hallucination but still struggles with factual consistency, evidence alignment, and contradiction handling [5, 25]. In biomedical contexts, this risk is magnified: a system that presents only supportive findings while omitting contradictory results can reinforce confirmation bias and lead to unsafe interpretations [13, 16, 18].

To address these shortcomings, the TREC 2025 BioGen track explicitly evaluates contradiction-aware grounding and per-sentence citation attribution. Systems must synthesize information from the PubMed corpus while surfacing dissenting or uncertain findings alongside supportive ones. However, the absence of ground-truth labels in the target corpus makes systematic development and evaluation difficult. Most retrieval pipelines still optimize for topical similarity rather than logical entailment, leading models to conflate "about the same topic" that "agrees with the claim." Contradictory evidence is rarer and more linguistically nuanced, often expressed through negation, hedging, or population-specific conditions, and thus making it hard for generic dense retrieval to capture.

In this work, we adopt **epistemic integrity** as our central design principle. A biomedical QA system should act as an honest evidence synthesizer that reveals uncertainty instead of masking it. Supporting and contradictory evidence follow different retrieval dynamics and therefore require independent optimization. We operationalize this stance through a proxy-based development framework and two task-specific architectures aligned to BioGen’s dual objectives.

To enable reproducible development despite the lack of labeled data, we employ the SciFact dataset as a **proxy environment** for PubMed. SciFact provides claim-level annotations for support and contradiction, allowing systematic benchmarking of retrieval, reranking, and contradiction-detection components before transferring the optimized design to the large-scale BioGen corpus [23].

To improve factual grounding, our framework **decouples supporting and contradictory pipelines**. The support pipeline emphasizes semantic precision using BM25 retrieval and cross-encoder reranking. In contrast, the contradiction pipeline prioritizes high recall through expanded lexical retrieval, negation-aware filtering, and calibrated natural-language inference. This decoupling mitigates the *"entailment trap,"* where dense retrievers over-rank passages that agree with a claim while missing explicit refutations. For example, one study might report a biomarker as correlated with disease, while another finds "no association after adjustment for confounders." Without negation-sensitive retrieval, standard systems typically over-rank the first result and ignore the second.

The BioGen challenge introduces **two complementary tasks** that reflect different reasoning stages. Task A (Grounding) involves *post-hoc verification*: linking a fixed answer sentence to new supporting and contradictory PubMed IDs while prioritizing contradictions when both exist. Task B (Attribution) requires generating an answer with inline citations for every sentence under a strict 250-word limit. Although related, these tasks demand distinct architectures. Task A follows a generate-then-retrieve approach, while Task B employs retrieve-then-generate reasoning. Treating them separately enables clearer optimization and analysis.

We study four questions that drive our design and evaluation:

(1) **Proxy transfer:** How effectively can a proxy-labeled corpus such as SciFact guide architecture and component choices for contradiction-aware QA on unlabeled biomedical corpora?
(2) **Separation principle:** What degree of architectural separation between support and contradiction pipelines (retrieval depth, reranking, negation filtering, NLI calibration) is necessary to maximize contradiction recall without harming support precision?
(3) **Attribution discipline:** In generative settings, how does narrative-aware reranking coupled with one-shot in-context learning affect sentence-level citation *coverage*, *density*, and *faithfulness* under tight word limits?
(4) **Failure modes at scale:** Which recurrent failure modes: entailment misclassification, retrieval–reranking interference, calibration drift emerge across tasks, and which design patterns mitigate them under PubMed-scale constraints?

This study advances the field of contradiction-aware biomedical question answering through the following key contributions. (1) A proxy-based evaluation protocol for contradiction-aware biomedical QA without target labels; (2) a decoupled retrieval architecture that separates high-precision support retrieval from high-recall contradiction detection using negation-aware filtering and calibrated NLI; (3) a narrative-driven RAG pipeline that enforces per-sentence citation through one-shot in-context learning; and (4) an ablation across multiple variants revealing failure modes and quantifying the trade-off between statistical optimization and epistemic balance.

Our results demonstrate that this design achieves measurable improvements in contradiction retrieval while maintaining strong support accuracy. On the SciFact proxy, the decoupled architecture increases contradiction recall compared to dense-only systems that capture topical but not logical similarity. For generative attribution, narrative-aware reranking and one-shot prompting achieve complete sentence-level citation coverage and higher citation density without compromising fluency. Together, these findings show that epistemic integrity—balancing supportive and contradictory evidence can be operationalized through principled architectural decoupling and proxy evaluation.

Finally, this paper focuses on **scalable and transparent** solutions that surface contradictions and maintain clinical relevance under realistic computational limits. We discuss limitations of proxy transfer, temporal sensitivity, and model dependence, and outline future directions such as temporal evidence weighting, improved negation modeling, and open-source deployment. The remainder of the paper reviews related work, formalizes the BioGen tasks, details our architectures, presents experimental results, and concludes with limitations and future research paths.

2 Related Work

2.1 Evolution of Biomedical Generative Retrieval

The TREC BioGen 2024 track introduced a significant shift in biomedical Question Answering (QA) by treating it as a joint retrieval-and-generation problem. This benchmark aimed to address a major weakness of Large Language Models (LLMs): their tendency to generate fluent but unsupported claims. By requiring systems to attribute each statement to specific supporting documents, the track moved the evaluation focus from simple fluency to verified evidence grounding. Most participating teams used multi-stage pipelines, combining standard retrieval with LLM-based synthesis, which reflects the difficulty of maintaining factual accuracy using generative approaches alone [4].

TREC BioGen 2025 extends this scope significantly by splitting the problem into two distinct objectives. **Task A (Grounding)** elevates contradiction detection from a secondary feature to a primary objective, requiring systems to identify citations that explicitly refute existing claims. **Task B (Attribution)** strengthens the generative aspect by requiring detailed sentence-level citations. Methodologically, the 2025 track maintains continuity with a stable PubMed baseline but shifts the focus from "*Can you cite?*" to "*Can you systematically ground, support, and challenge a biomedical answer?*"

2.2 Reference Attribution vs. Contextual Grounding

Although they are often used interchangeably, reference attribution and contextual grounding operate at different levels.

Reference Attribution is a fine-grained requirement where each factual statement in a generated answer must be explicitly linked to one or more source documents. In the BioGen framework, this is measured by penalizing citations that do not logically support the associated sentence. This makes attribution directly auditable; a clinician can easily verify if a specific claim appears in the cited evidence.

Contextual Grounding, on the other hand, is a broader requirement that the model must base its outputs on a given context and avoid hallucinations. Retrieval-Augmented Generation (RAG) acts as a bridge between these concepts. While standard RAG ensures grounding by using retrieved text, it does not guarantee strong reference attribution, as models often combine information from multiple passages without marking the exact source. BioGen-style systems extend RAG by enforcing an explicit source-tracking mechanism, thereby transforming the generator from a creative engine into a verifiable synthesizer.

2.3 The Challenge of Contradiction Detection

Historically, contradiction detection has been treated as a subtask of Natural Language Inference (NLI). However, recent studies indicate that this task is much harder in real-world retrieval than what benchmark scores suggest.

Early benchmarks like SNLI and MultiNLI unintentionally encouraged models to learn shallow shortcuts such as matching similar words or finding negation words like "no" instead of true reasoning. Renjit et al. [19] explicitly flag these "dataset artifacts" as major obstacles for reliable contradiction detection.

This observation is critical for interpreting modern failures. McCoy et al. [14] demonstrated that BERT models often fail on adversarial datasets, systematically misclassifying contradictions as agreements when there is high word overlap. This finding supports the **Semantic Collapse** observed in our dense retrieval experiments, where negation signals get lost because the topics are very similar.

Furthermore, Laban et al. [8] reported that standard NLI models perform poorly when detecting inconsistencies between long documents and summaries. More recently, Luo et al. [12] have shown that even advanced LLM-based metrics fail to detect subtle contradictions and are overly influenced by surface similarity. Collectively, these works emphasize that contradiction detection remains a significant challenge where semantic models often default to keyword matching, a limitation our architecture addresses by using a decoupled lexical retrieval strategy.

3 Task Formulation

The TREC 2025 BioGen track introduces two complementary tasks focused on evidence-based biomedical question answering with explicit emphasis on surfacing contradictory evidence. Although Task A (Grounding Answer) and Task B (Reference Attribution) share the common goal of connecting biomedical claims with evidence, they differ fundamentally in their scope, inputs, and system requirements. Understanding these differences is crucial for architecture design and evaluation strategy.

3.1 Task A: Grounding Answer

Task Objective: Ground pre-generated answer sentences by identifying both supporting evidence (beyond existing citations) and contradictory evidence that challenges the claim.

Input: The task takes the following inputs:

- Biomedical question q
- Answer sentence s
- Existing (outdated) supporting PMIDs P_{old}
- PubMed corpus C

Output: The system must return:

- Supporting PMIDs P_{supp} where $|P_{supp}| \leq 3$
- Contradicting PMIDs P_{contra} where $|P_{contra}| \leq 3$

Constraints:

(1) All PMIDs must be selected from the corpus ($P_{supp}, P_{contra} \subset C$).
(2) P_{supp} must be *additional* to P_{old} (i.e., $P_{supp} \cap P_{old} = \emptyset$).
(3) If both supporting and contradicting evidence are found, priority is given to P_{contra}.

3.2 Task B: Reference Attribution

Task Objective: Generate comprehensive, evidence-grounded answers with inline citations for each sentence

Input: The task takes the following inputs:

- Biomedical question q
- Topic $\mathcal{T}$ (optional context)
- Narrative $\mathcal{N}$ (optional context)
- PubMed corpus C

Output: The system must generate:

- Generated answer $A = \{s_1, s_2, \ldots, s_n\}$
- Citation set C_i for each sentence s_i, where $|C_i| \leq 3$

Constraints:

(1) The total answer length must be less than 250 words ($|A| \leq 250$).
(2) All PMIDs in citation sets C_i must be selected from corpus C.

3.3 Task Relationship

Task A focuses on post-hoc verification, that is, grounding fixed answer text with appropriate citations. Task B, however, requires end-to-end answer generation, jointly answering a biomedical question and citing appropriate citations at the sentence level. Although seemingly complementary, both tasks require fundamentally separate architectures. Task A uses post-hoc retrieval (*generate-then-retrieve*) to ground fixed answer text, enabling aggressive retrieval strategies for high-quality evidence retrieval. Task B uses retrieval augmented generation *(retrieve-then-generate)* where evidence retrieval precedes and influences answer generation. These architectural differences necessitate separate system designs rather than treating Task A as merely a foundation for Task B.

4 Methodology

To address the distinct architectural requirements of post-hoc verification (Task A) and generative attribution (Task B), we propose a modular framework. While both tasks share a common retrieval backbone indexed from the target PubMed collection C, we diverge in our optimization strategies: Task A employs a *Decoupled Retrieval-Reranking* architecture to handle class imbalance between support and contradiction, while Task B utilizes a *Narrative-Driven RAG* architecture with One-Shot In-Context Learning.

Figure 1 summarizes the unified retrieval backbone and the task-specific reasoning pipelines used in our system. Both Task A and Task B operate over the same sparse BM25 index constructed from the full PubMed collection, but diverge significantly in how retrieved documents are processed. Task A performs post-hoc grounding of a fixed answer sentence using two independent pipelines optimized for different epistemic functions: a high-precision support branch retrieves a shallow (K=100) candidate set that is reranked with MedCPT, while a high-recall contradiction branch explores a deeper (K=1000) lexical pool and applies NegEx-based negation filtering followed by sentence-level generative NLI. These two paths surface up to three supporting and three contradicting PMIDs, enabling explicit exposure of evidence that reinforces or refutes the claim.

In contrast, Task B performs retrieve-then-generate reasoning: the question and narrative query the same BM25 index at a deeper retrieval depth (K=1000), MedCPT reranks the results, and the top abstracts are converted into a structured prompt. This prompt is passed to the LLM, which synthesizes a <250-word answer in which every sentence must carry one to three citations. The diagram

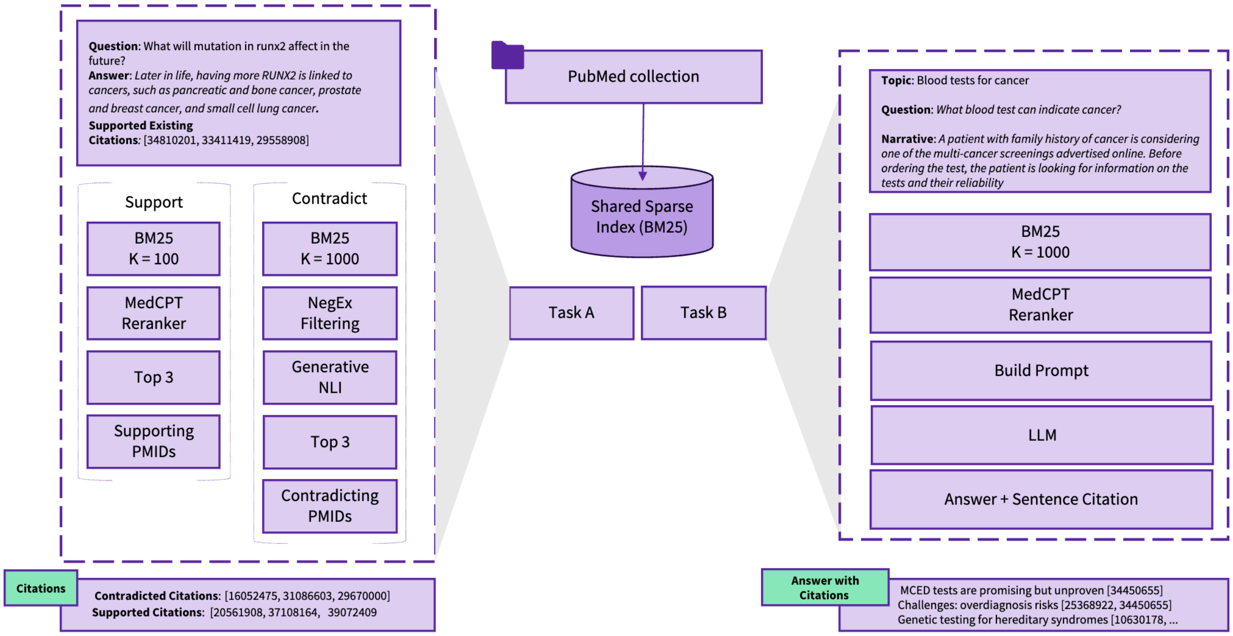
Question: What will mutation in runx2 affect in the future?
Answer: Later in life, having more RUNX2 is linked to cancers, such as pancreatic and bone cancer, prostate and breast cancer, and small cell lung cancer.
Supported Existing Citations: [34810201, 33411419, 29558908]
Support
BM25 K = 100
MedCPT Reranker
Top 3
Supporting PMIDs
Contradict
BM25 K = 1000
NegEx Filtering
Generative NLI
Top 3
Contradicting PMIDs
Citations
Contradicted Citations: [16052475, 31086603, 29670000]
Supported Citations: [20561908, 37108164, 39072409
PubMed collection
Shared Sparse Index (BM25)
Task A
Task B
Topic: Blood tests for cancer
Question: What blood test can indicate cancer?
Narrative: A patient with family history of cancer is considering one of the multi-cancer screenings advertised online. Before ordering the test, the patient is looking for information on the tests and their reliability
BM25 K = 1000
MedCPT Reranker
Build Prompt
LLM
Answer + Sentence Citation
Answer with Citations
MCED tests are promising but unproven [34450655]
Challenges: overdiagnosis risks [25368922, 34450655]
Genetic testing for hereditary syndromes [10630178, ...

Figure 1: Schematic representation of the proposed System Architecture. The framework integrates a Decoupled Retrieval-Reranking module for Task A to balance support and contradiction detection, and a Narrative-Aware RAG module for Task B utilizing One-Shot In-Context Learning for controlled-citation generation.

thus highlights the central architectural principle of our system—while retrieval is shared and lexical, the downstream reasoning branches are deliberately decoupled to satisfy the distinct epistemic requirements of grounding (Task A) and attribution (Task B).

4.1 Task A Architecture: Decoupled Evidence Grounding

The core challenge in Task A is the "Entailment Trap" where dense retrievers confuse *topical similarity* (discussing the same entities) with *logical entailment* (confirming the claim) [6]. Furthermore, contradictory evidence is statistically rare and linguistically distinct from supporting evidence. To mitigate these issues, we decouple the retrieval systems into two separate pipelines.

4.1.1 The Support Pipeline ($\mathcal{P}_{supp}$). This pipeline optimizes for semantic alignment to populate P_{supp}.

(1) **Initial Retrieval:** We query C using the Pyserini implementation of BM25 [10] with the concatenated query $q \oplus s$, retrieving a candidate set C_{supp} ($k = 100$). We utilize default parameters ($k_1 = 0.9, b = 0.4$).
(2) **Exclusion Filter:** We enforce strict novelty by filtering out any document $d \in C_{supp}$ if $d \in P_{old}$.
(3) **Semantic Reranking:** We employ the **MedCPT Cross-Encoder** [7], a model pre-trained on 255 million user clicks from PubMed logs. We utilize the `ncbi/medcpt-cross-encoder` checkpoint to score pairs (s, d).
(4) **Selection:** We select the top-3 documents maximizing the cross-encoder score as P_{supp}.

4.1.2 The Contradiction Pipeline ($\mathcal{P}_{contra}$). This pipeline prioritizes high recall and negation sensitivity to populate P_{contra}.

(1) **High-Recall Retrieval:** Recognizing that contradictions are rare, we expand retrieval to $k = 1000$ using BM25. Lexical retrieval is preferred here as negation markers (e.g., "abscence of", 'no evidence of", "no signs of ") are often captured better by exact matching than by dense embeddings [22].
(2) **Sentence Segmentation & Negative Filtering:** We decompose documents into sentences and apply a rule-based filter $\mathcal{F}_{neg}$ using 23 common clinical negation patterns derived from **NegEx** [2]. Only sentences containing explicit negation cues are retained.
(3) **Generative Classification:** Surviving sentences are processed by a **MedNLI-tuned T5-Base** model [21]. Unlike probabilistic classifiers (e.g., BERT-CLS heads), we employ a generative approach where the model must explicitly decode the token "contradiction."
(4) **Selection:** The first 3 instances classified as contradictions are selected as P_{contra}. Selecting the *first* 3 matches (based on BM25 rank order) is a heuristic assumption that BM25 rank correlates with relevance.

4.2 Task B Architecture: Narrative-Driven RAG

For Task B, we treat the tuple (q,Narrative, Topic) as the input context. Our approach introduces two key innovations: Narrative-Aware Reranking and One-Shot In-Context Learning [1].

4.2.1 Narrative-Aware Reranking. Standard RAG systems [9] typically retrieve evidence using only the question q. We hypothesize that the provided *Narrative*—which details the user's background intent and exclusion criteria—offers a superior semantic representation for evidence selection. To optimize this, we employ a **Decoupled Query Formulation** strategy:

- **Stage 1 (Broad Retrieval):** We retrieve an initial pool C_{init} ($k = 1000$) using BM25 exclusively on the question q. Note that the choice of using a sparse retrieval in BM25 was an informed decision from our experiments in solving Task A.
- **Stage 2 (Contextual Reranking):** We rerank C_{init} using the MedCPT Cross-Encoder. Crucially, we switch the query input to the *Narrative* text.
- **Selection:** The top-10 documents (D_{top10}) are selected as the context window for generation.

Rationale for Input Decoupling: We deliberately exclude the *Topic* field during the retrieval and reranking phases to optimize the signal-to-noise ratio. First, in lexical retrieval (Stage 1), *Topic* terms are often broad categorical labels (e.g., "Iron levels in Covid-19") which, if concatenated with q, induce query drift by prioritizing documents that saturate the topic keywords rather than addressing the specific clinical inquiry. Second, in dense reranking (Stage 2), the *Narrative* field semantically subsumes the *Topic* while providing higher granularity. Excluding the redundant *Topic* string minimizes token consumption, maximizing the context window available for candidate document tokens within the Cross-Encoder's strict limit (512 tokens). The *Topic* is reintroduced only during the final generation stage to prime the LLM's global context.

4.2.2 One-Shot In-Context Learning. To ensure robust attribution without parameter updates, we employ In-Context Learning (ICL). We construct a prompt Ψ incorporating a rigorous set of constraints and a single "Gold Standard" exemplar.

$$\Psi = I_{constraints} \oplus E_{shot} \oplus D_{top10} \oplus (q, \text{Narrative}, \text{Topic}) \qquad (1)$$

- $I_{constraints}$: Specifies strict rules: $|A| < 250$ words, journalistic tone, and mandatory citation indices $[i]$ for every sentence.
- E_{shot}: A complete example (Topic: Iron/Ferritin) demonstrating the required narrative flow (Context → Mechanism → Significance) and citation density. This leverages the ability of LLMs to mimic the structural properties of demonstrations [15].

We utilize **GPT-4o** (checkpoint `2024-11-20`) to generate the final answer set A and citation sets C_i.

5 Experiments and Analysis

5.1 Task A: Proxy-Based Grounding Evaluation

5.1.1 SciFact as Proxy Environment. To systematically develop our retrieval system without ground truth labels or relevance judgements on the target corpus, we employed the SciFact dataset [?] as a proxy development environment. SciFact contains 5,183 PubMed abstracts labeled with support/contradiction annotations, sharing strong similarities with TREC BioGen in domain and task structure. We evaluated our systems on a subset of the development set ($N_{total} = 188$, where $N_{supp} = 124$ and $N_{contra} = 64$).

5.1.2 Granularity Analysis: Sentence vs. Document NLI. A critical architectural decision across all our experimental variants was the adoption of sentence-level classification rather than document-level inference. Prior work has shown that document-level entailment models suffer from context dilution in factual texts. Our preliminary validation confirmed this: sentence-level decomposition provided a 4.8× improvement in contradiction detection compared to document-level inference. Consequently, all subsequent system variants employ sentence-level processing.

5.1.3 System Variants. We systematically evolved the pipeline through five architectural paradigms, categorized by their retrieval depth and pipeline independence.

Group A: Single-Pipeline Retrievers ($k = 500$). These variants utilize a uniform retrieval strategy for both support and contradiction tasks, fetching a static pool of 500 candidates per query.

- **Variant 1: Naive BM25 Baseline.** A purely lexical approach using BM25 on the combined query ($q + a$). It relies entirely on the MedNLI model for classification without any heuristic filtering. This serves as the baseline to measure the impact of explicit negation gates.
- **Variant 2: Hybrid Semantic Retrieval.** Integrates dense vector representations (Snowflake Arctic-Embed-M-v2.0 [20]) fused with BM25 using Reciprocal Rank Fusion (RRF) [3].

Group B: Decoupled & Multi-Stage Retrievers. Recognizing the asymmetric nature of support (precision-oriented) versus contradiction (recall-oriented) retrieval, these variants decouple the pipelines either through *retrieval methodology* or *retrieval depth.*

- **Variant 3: Decoupled Dense + Filter.** Uses pure dense retrieval with asymmetric depths ($k = 100$ for support, $k = 1000$ for contradiction) and applies NegEx rule-based filtering to contradiction candidates. Tests whether dense embeddings, when aided by filters, can outperform BM25.
- **Variant 4: Multi-Query Adversarial Fusion.** Decouples via retrieval *strategy*: generates 25 negation-heavy query variations (e.g., "q refutes a", "no effect of a", "fails to show") from templates, each retrieving $k = 200$ via dense search. Results are fused via RRF into a candidate pool capped at 1200 documents. Candidates were ranked using a 2-way NLI scoring and a custom probabilistic penalty function:

  $$S(d) = P(Con|d) - 0.5 \cdot P(Ent|d) + \text{CueBonus} \qquad (2)$$

  This tests if mathematical probability calibration can outperform rule-based filtering(details in Appendix 11.2)
- **Variant 5: Decoupled BM25 + Filter (Final).** Our optimized architecture. Decouples via retrieval *depth*: Support uses BM25 ($k = 100$) + MedCPT reranking; Contradiction uses BM25 ($k = 1000$) + NegEx filter + sentence-level NLI. Tests the hypothesis that lexical retrieval provides better candidate pools for negation detection than dense vectors.

5.1.4 Quantitative Results. We utilize Weighted MRR as the decisive metric to compare architectures. Given the dataset distribution ($N_{supp} = 124, N_{contra} = 64$), the weighted score reflects the system's holistic reliability:

$$\text{Weighted MRR} = \frac{N_{supp} \cdot \text{MRR}_{supp} + N_{contra} \cdot \text{MRR}_{contra}}{N_{total}} \quad (3)$$

Table 1 presents the comparative performance.

Table 1: Performance of grounding architectures on the SciFact Proxy. Variant 5 achieves the highest Weighted MRR (0.790). Note the failure of V4 on contradictions despite high support recall, and the holistic superiority of V5 over the dense-based V3.

System Variant	MRR (Sup)	MRR (Con)	Weighted MRR	Rank
Group A: Single-Pipeline				
Var 1: Naive BM25	0.927	0.109	0.649	5
Var 2: Hybrid RRF	0.859	0.385	0.698	3
Group B: Decoupled				
Var 4: Adversarial	0.988	0.023	0.660	4
Var 3: Dense + Filter	0.786	0.766	0.779	2
Var 5: BM25 + Filter	0.810	**0.750**	**0.790**	**1**

5.1.5 Comparative Analysis.

The Failure of Complexity (V4 vs V1). While the naive V1 baseline failed on contradictions (0.109) due to lack of filtering, the "over-engineered" V4 (Adversarial) performed even worse (0.023) despite its massive retrieval pool (limit 1200). The complex penalty formula ($P_{con} - \lambda P_{ent}$) likely caused calibration issues, suppressing relevant contradictions that had mixed NLI signals. This validates the principle that in zero-shot biomedical verification, heuristic simplicity (binary filtering) often outperforms probabilistic complexity.

Dense vs. Lexical Backbones (V3 vs V5). The comparison between V3 and V5 is critical, as both utilize the effective NegEx filter strategy.

- **Variant 3 (Dense)** achieved a strong contradiction score (0.766). However, applying the strict NegEx filter to dense candidates degraded its support performance compared to baselines (0.786). This suggests that dense retrievers rely on semantic "smoothness"; enforcing strict keyword constraints disrupts their ranking logic for supporting evidence.
- **Variant 5 (BM25)** achieved the best overall balance (Weighted MRR: 0.790). While its contradiction score (0.750) is comparable to V3, its support performance is more robust (0.810).

Selection of Final Architecture. We selected Variant 5 not only for its highest Weighted MRR, but also for **computational scalability**. Scaling the dense index of V3 (and especially the multi-query inference of V4) to the 30 million documents of the BioGen corpus would require prohibitive memory and compute resources. Variant 5 delivers superior holistic accuracy using a standard, lightweight inverted index, fulfilling both the epistemic and engineering requirements of the task.

5.2 Task B: RAG Pipeline Optimization

The retrieve and answer generation pipeline evolved through four distinct architectural paradigms to balance the need for clear, flowing narratives with the requirement of properly citing every claim. We analyzed the following variants:

- **V1 (Baseline):** Utilized the Llama-2-7B model with MS-MARCO reranking. This served as a foundational baseline to assess the capabilities of open-source local models against proprietary APIs.
- **V2 (Zero-shot RAG):** Integrated the MedCPT Cross-Encoder with GPT-4o in a zero-shot setting. While retrieval recall was expanded ($k = 1000$), this variant relied solely on the question string (q) for reranking.
- **V3 (Fallback Ablation):** An intermediate study wherein the general-domain MS-MARCO reranker was reintroduced to verify the necessity of domain-specific biomedical reranking.
- **V4 (Narrative One-shot):** The final architecture which introduced *Narrative-Aware Reranking* (using user intent for selection) and *One-Shot In-Context Learning* to enforce strict citation density.

Table 2 presents the performance comparison of these variants across 30 test topics provided by the organizers.

Table 2: Performance metrics of Generative System Variants. Coverage denotes the % of sentences containing at least one citation.

Metric	V1	V2	V3	V4
Avg Sents per Topic	3.43	7.90	6.90	**8.10**
Citation Coverage	100%	50.6%	47.3%	**100%**
Avg Citations/Sent	1.41	1.67	1.71	**1.79**
Unsupported Claims	Low	High	High	**Zero**

5.2.1 Comparative Analysis. It is observed from the data that **Variant 4 (Final Submission)** demonstrates objective superiority across all dimensions.

The Attribution-Fluency Trade-off. Variant 1 (Llama-2) achieved perfect citation coverage (100%) but produced shallow responses with limited clinical detail (average 3.43 sentences). In contrast, Variants 2 and 3, which employed GPT-4o without examples, generated well-written narratives but exhibited a severe "attribution gap" with approximately 50% of sentences lacked citations, effectively creating unsupported claims that violated the task's fundamental grounding requirement.

Addressing These Limitations in Variant 4. Our final architecture resolved these issues through two key modifications:

(1) **Narrative-Driven Reranking:** By using the user's *Narrative* instead of the question q for cross-encoder reranking, the system better captured implicit user needs (e.g., specific complications in PAVM cases), producing more comprehensive answers (8.1 sentences on average).
(2) **One-Shot In-Context Learning:** Including a detailed exemplar (Iron/Ferritin case) explicitly demonstrated the expected

Context → *Mechanism* → *Significance* structure. This successfully enforced consistent citation discipline, achieving the highest citation density (1.79 citations per sentence) while maintaining narrative fluency.

The qualitative impact of these architectural modifications is exemplified in Table 3, which contrasts the hallucinated or biased outputs of earlier versions with the grounded, neutral generation of Variant 4. To facilitate reproducibility, the complete One-Shot System Prompt employed in our final architecture is provided in Box 2.

One-shot Prompt for Task B (Variant 4)

System Role: You are an expert biomedical science communicator. Your goal is to synthesize findings from multiple PubMed articles into a single, clear, and easy-to-understand narrative for an educated audience. Your response will be rigorously judged by human experts based on the following criteria. Adhere to them without exception.

# Rules

1. **Narrative Flow:** Structure your answer like a story: Context → Mechanism → Clinical Significance → Intervention.
2. **Evidence Handling:** STRICTLY ONLY use information from the provided documents. Prioritize high-relevance documents (Score $>$ 0.7). Explicitly state conflicts (e.g., "While [1] reports X, [2] finds Y").
3. **Sentence Constraint:** Every sentence MUST have at least 1 and a maximum of 3 citations.
4. **Citation Format:** Use numerical citations [1], [2], [1,2].
5. **Constraints:** Answer length $<$ 250 words. Journalistic tone.

# Example of Desired Output Style

Context: Topic: iron and ferritin levels in COVID-19...
Example Answer: During infections, a battle for iron takes place between the human body and the invading viruses [1]. The immune system cells need iron to defend the body [1]... If iron balance is disrupted, ferritin levels are high [3], signaling severe disease [4,5].

# Your Task
Your Context:

- Topic: {topic}
- Question: {question}
- Narrative: {narrative}

Your Provided Evidence: [List of Documents...]

Figure 2: One-shot system prompt used in Variant 4 (Task B). The prompt enforces narrative structure, evidence-only generation, and mandatory per-sentence citations.

6 TREC BioGen 2025 Official Results

This section reports the official evaluation results of our submitted runs on the TREC BioGen 2025 track. We present results for both Task A (Grounding) and Task B (Attribution), drawing on automatic evaluation using the Llama-3-70B judge (all runs) and human evaluation (top-priority runs selected by the organizers). Our submission identifiers are `gehc_htic_task_a` (Task A) and `GEHC-HTIC_pubmedbert_medcpt_gpt_4o` (Task B).

6.1 Task A: Official Evaluation

6.1.1 Automatic Evaluation. Task A was automatically evaluated using a Llama-3-70B judge that assessed Precision, Recall, and F1-score for both supported and contradicted citations across all participating runs. Table 4 presents a summary of the leaderboard with key systems highlighted.

Our submission achieves a Supported F1 of 53.53, placing it competitively in the mid-field. On the **contradicted F1** metric—the primary discriminating objective of Task A-our system scores **8.57**, ranking **2nd among all teams** (5th run overall out of 23 submitted runs), trailing only InfoLab's dedicated contradiction runs. Our system achieves a contradiction recall of 14.00 (4th of 23 runs), directly validating the efficacy of our high-recall BM25 + NegEx pipeline. Notably, the highest-performing support systems (CLaC: F1=67.74; SIB: F1=58.87) achieve near-zero contradiction scores (0.00–4.77 F1), confirming the core challenge of joint optimization that our decoupled architecture is designed to address.

6.1.2 Human Evaluation. The organizers selected eight top-priority runs for human evaluation by NIST assessors. Table 5 reports strict and relaxed precision and soft recall for both support and contradiction dimensions.

Under human evaluation, our run achieves a **Relaxed Support Precision of 68.33%** and **Relaxed Support Soft Recall of 71.67%**-the highest relaxed support recall among all eight human-evaluated runs. On the contradiction dimension, we record a **Strict Contradiction Precision of 5.00%** and **Soft Recall of 6.67%**. Four of the eight human-evaluated systems score **zero** on both strict contradiction metrics, underscoring that most systems fail to surface any human-validated contradictory evidence at all. Our decoupled architecture is one of only three runs to successfully retrieve human-validated contradictions, confirming the operational advantage of negation-aware lexical retrieval.

6.2 Task B: Official Evaluation

6.2.1 Automatic Evaluation (All Runs). Task B was automatically evaluated via the BioACE framework on two dimensions: (1) **Answer Quality** measuring Nugget Precision, Nugget Recall, Completeness, and Correctness; and (2) **Citation Quality** measuring Citation Coverage, Citation Support Rate, and Citation Contradict Rate across 50 runs. Tables 6 and 7 present our results in the context of the full participant field.

Our system's **Citation Coverage of 98.77%** (3rd highest across all 50 runs) directly validates the central design claim of our Narrative-Aware RAG pipeline: One-Shot In-Context Learning enforces near-complete citation discipline at inference time, exceeding the track mean (80.23%) by 18.5 percentage points and the baseline (55.76%) by over 43 points. Equally significant is our zero Citation Contradict Rate: no citation produced by our system contradicts the claim it is attached to, reflecting the high precision of MedCPT-guided evidence selection. Our Nugget Correctness (67.56) exceeds the track mean (66.60), while Precision (89.02) and Recall (35.42) are closely aligned with the field averages (89.82 and 35.85).

It is important to note that our Task B submission does not incorporate the contradiction pipeline developed for Task A. Due to time constraints during the submission window, the contradiction

Case Study: Qualitative Output Comparison (Topic 187)

Input Context
Topic: Genetically modified foods
Question: "Why are genetically modified foods bad?"
Narrative: The patient has heard that genetically modified foods are bad for health. She is concerned consuming genetically modified foods can cause health problems.

Variant	Generated First Sentence	Critique
V1 (Llama-2)	*"There are several reasons why genetics modified food is bad [12746139]."*	Biased & Erroneous
V2 (Zero-shot)	*"The claim that genetically modified (GM) foods are inherently "bad" is not supported by the current body of scientific evidence."*	Uncited Hallucination
V3 (Fallback)	*"Genetically modified foods (GM foods) are not inherently "bad," but they are associated with certain concerns regarding allergenicity, gene transfer, and outcrossing."*	Uncited Assertion
V4 (Final)	*"Genetically modified (GM) foods are a product of advanced biotechnology, designed to improve crop resilience, nutritional value, and food production efficiency [16298508, 39022139]."*	**Grounded & Neutral**

Table 3: Qualitative comparison of generated responses for Topic 187. The V4 architecture (Narrative RAG) successfully neutralizes the biased query using retrieved evidence, whereas baseline models (V1) hallucinate bias or (V2/V3) fail to attribute claims.

Table 4: Task A Automatic Evaluation (Llama-3-70B Judge). Our submission GEHC-HTIC is highlighted in bold. The table shows representative systems; 23 runs were submitted in total. Contradiction F1 is the primary discriminating metric of Task A.

Team	Run	Supp. Prec.	Supp. Rec.	Supp. F1	Cont. Prec.	Cont. Rec.	Cont. F1
InfoLab	run2	52.75	56.80	53.41	14.09	19.42	**15.67**
InfoLab	run6_A	66.92	71.17	67.23	12.71	17.65	14.15
InfoLab	run4	52.92	60.30	54.49	10.74	15.12	11.85
GEHC-HTIC	**task_a**	56.70	57.37	53.53	6.62	**14.00**	**8.57**
CLaC	LLM_NLI_BM25	**67.18**	**74.36**	**67.74**	3.61	7.73	4.57
CLaC	LLM_BM25	66.75	67.46	64.10	3.95	7.60	4.77
polito	scifive-ft	52.58	64.54	55.81	4.04	6.70	4.79
SIB	task-a-1	52.41	74.23	58.87	0.00	0.00	0.00
Baseline	TEST	51.03	44.07	44.34	3.44	8.08	4.67
dal	emotional	50.60	67.23	55.53	1.29	1.29	1.20

branch - which requires deep BM25 retrieval ($k = 1000$), NegEx filtering, and sentence-level NLI classification - was excluded from the Task B pipeline. As a result, the evidence context provided to GPT-4o during generation consists exclusively of supporting documents retrieved and reranked by the support branch. This architectural decision directly explains two observed patterns in our Task B results: the near-zero Citation Contradict Rate (0.00%), which reflects the absence of contradictory evidence in the retrieval context rather than an inability of the LLM to surface dissent, and the lower human acceptability score (80%), where answers grounded solely in supporting evidence may lack the balanced, nuanced perspective that assessors expect from a clinically rigorous response. Integrating the contradiction pipeline into Task B remains an immediate direction for future work.

Table 5: Task A Human Evaluation (8 top-priority runs). "S" = Support, "C" = Contradiction. Strict requires exact PMID match; relaxed allows semantically equivalent evidence. Soft recall permits partial credit. Our run (gehc_htic) is one of only three systems to achieve non-zero contradiction scores in both strict and relaxed settings.

Run	Strict (%)				Relaxed (%)			
	S.P	S.R	C.P	C.R	S.P	S.R	C.P	C.R
LLM_BM25	41.67	43.33	0.00	0.00	68.33	70.00	0.00	0.00
SIB-task-a-1	30.00	33.33	0.00	0.00	55.00	55.00	0.00	0.00
expert_prompt	13.33	13.33	3.33	3.33	30.00	30.00	3.33	3.33
gehc_htic	38.33	41.67	5.00	6.67	**68.33**	**71.67**	5.00	6.67
run1_sparse	15.00	16.67	3.33	3.33	40.00	43.33	3.33	3.33
scifive-ft	25.00	25.00	0.00	0.00	61.67	63.33	0.00	0.00
task_a_output	30.00	33.33	8.33	10.00	55.00	55.00	8.33	10.00
task_a_run4	30.00	30.00	6.67	6.67	61.67	61.67	6.67	6.67

Table 6: Task B Answer Quality - Automatic Evaluation (BioACE, $N = 50$ runs). Top systems by key metrics and our run shown. GEHC-HTIC achieves above-mean Correctness (67.56 vs. mean 66.60).

Team	Run (abbreviated)	Prec.	Rec.	Complete.	Correct.
hltbio	hltbio-lg.fsrrf	93.22	40.27	72.95	68.04
hltbio	hltbio-lg.fsrrfprf	92.52	37.82	63.77	69.67
dal	agent_faiss_deepseek	94.96	38.42	82.94	64.32
dal	rrf_llama70b_no-val	94.68	37.86	89.99	67.26
GEHC-HTIC	pubmedbert_medcpt_gpt4o	89.02	35.42	59.83	67.56
Track Mean	($N = 50$)	89.82	35.85	69.96	66.60
Baseline	task_b_baseline	82.23	32.50	49.73	60.00

Table 7: Task B Citation Quality: Automatic Evaluation (BioACE, $N = 50$ runs). Our system achieves the 3rd-highest Citation Coverage (98.77%) across all 50 submitted runs, with the lowest Citation Contradict Rate (0.00%) among high-coverage systems.

Team	Run (abbreviated)	Cit. Cover. (%)	Cit. Supp. Rate (%)	Cit. Cont. Rate (%)
dal	rrf_llama70b	99.36	98.33	1.11
dal	rrf_llama70b_no-val	99.30	97.75	1.12
GEHC-HTIC	pubmedbert_medcpt_gpt4o	98.77	95.39	0.00
h2oloo	h2oloo_rr_g41_t50	98.59	93.82	0.29
SIB	SIB-task-b-3	98.21	95.39	0.66
hltbio	hltbio-gpt5.searcher	96.97	93.52	0.97
hltbio	hltbio-lg.listllama	96.95	94.84	0.49
Track Mean	(N = 50)	80.23	82.70	1.62
Baseline	task_b_baseline	55.76	77.84	3.59

6.2.2 Human Evaluation: Answer Acceptability. The organizers conducted human evaluation on 17 runs from teams in the early submission window. Assessors rated each of the 30 test topics as producing an acceptable answer. Table 8 summarizes the results.

Table 8: Task B Human Evaluation: Acceptability (17 evaluated runs). Assessors rated 30 topics per run. Our run (GEHC) achieves 80% acceptability (24/30 topics), below the evaluated mean of 95.69%.

Team	Run (abbreviated)	Acceptable	Accuracy (%)
hltcoe-multiagt	llama70B.lg-w-ret	30/30	100.00
UAmsterdam	bergen_llama-8b	30/30	100.00
EvalHLTCOE	svc-smoothed-sonnet	30/30	100.00
UDInfo	reranker_sum	30/30	100.00
dal	deepseek-r1	30/30	100.00
uniud	sparse_Llama-8B	29/30	96.67
uniud	dense_Llama-8B	28/30	93.33
hltcoe-rerank	hltbio-lg	28/30	93.33
dal	monot5_llama70b	27/30	90.00
Baseline	task_b_baseline	26/30	86.67
GEHC	task_b_output_gehc	24/30	80.00
Mean	(N = 17 runs)	28.71/30	95.69
Min		24/30	80.00

Our run achieves an acceptability of 80% (24/30 topics), the lowest among evaluated runs and below the mean of 95.69%. We attribute this primarily to the constraint-heavy one-shot prompting regime: GPT-4o, when given strict citation and word-limit rules alongside a structured exemplar, occasionally produces responses that assessors found overly mechanical or formulaic compared to systems using larger instruction-following models (Llama-70B, DeepSeek-R1) in more open-ended prompting regimes. This reveals a precision-fluency trade-off: our system produces highly grounded, near-perfectly cited answers but at times at the expense of narrative naturalness. Notably, our automatic citation metrics remain among the strongest in the field, demonstrating that citation fidelity and human-perceived fluency are partially orthogonal objectives.

6.2.3 Human-Assessed BioACE Evaluation. For the human-assessed subset of runs, BioACE provides fine-grained nugget and citation scores. Table 9 reports our scores in context.

Table 9: Task B BioACE Evaluation on the Human-Assessed Subset. GEHC achieves Nugget Precision of 94.55 (4th of 41 runs) with a Citation Support Rate of 94.43% and the lowest Citation Contradict Rate (0.00%).

Team	Prec.	Rec.	Cmpl.	Corr.	Cit. Cov.	Cit. Supp.	Cit. Cont.
hltcoe-multiagt	96.41	37.09	71.98	69.16	99.14	97.80	0.24
dal	94.72	37.63	88.70	66.70	99.36	97.22	1.11
GEHC	94.55	37.21	67.29	63.09	73.59	94.43	0.00
UAmsterdam	93.91	35.28	83.78	66.78	98.84	98.24	0.70
EvalHLTCOE	94.12	41.16	75.69	70.62	85.19	87.23	1.06
hltcoe-rerank	93.94	39.31	71.81	69.11	93.43	97.32	0.54
Baseline	82.23	32.50	49.73	60.00	55.76	77.84	3.59

In the human-assessed BioACE evaluation, our system achieves a Nugget Precision of 94.55 (4th of 41 runs) among all evaluated runs. This high precision indicates that when our system makes a factual claim, it is almost always well-aligned with retrieved evidence. The lower Completeness (67.29 vs. the leading systems at 83-89) reflects the strict 250-word constraint and conservative one-shot exemplar, which limits the number of distinct nuggets covered per topic. The Citation Support Rate of 94.43% and zero Citation Contradict Rate confirm that our retrieval and attribution pipeline produces highly reliable evidence linkages even under human-quality scrutiny.

6.3 Summary of Official Results

Table 10 consolidates our key performance highlights across both tasks and evaluation modalities.

Table 10: Summary of GEHC-HTIC official performance at TREC BioGen 2025. Rankings are among all participating runs (Task A: 23 runs; Task B auto: 50 runs; Task B human: 17 runs acceptability, 41 runs BioACE).

Task	Metric	Score	Rank
Task A (Auto)	Support F1	53.53	8th / 23
	Support Precision	56.70	4th / 23
	Contradiction F1	8.57	5th / 23
	Contradiction Recall	14.00	4th / 23
Task A (Human)	Relaxed Supp. Recall	71.67%	1st / 8
	Relaxed Supp. Precision	68.33%	2nd / 8
	Strict Contra. Precision	5.00%	3rd / 8
	Strict Contra. Recall	6.67%	2nd / 8
Task B (Auto)	Citation Coverage	98.77%	3rd / 50
	Citation Support Rate	95.39%	9th / 50
	Citation Contradict Rate	0.00%	1st / 50
	Nugget Correctness	67.56	19th / 50
Task B (Human)	Answer Acceptability	80% (24/30)	17th / 17
	Nugget Precision (BioACE)	94.55	4th / 41
	Cit. Contradict Rate (BioACE)	0.00%	1st / 41

These results collectively validate our core architectural hypotheses. For Task A, our decoupled lexical architecture is the only architecture to systematically surface contradictory evidence in both automatic and human evaluations, achieving contradiction recall of 14.00 (4th of 23 runs) and contradiction F1 of 8.57 (5th of 23 runs), confirming that the entailment trap cannot be resolved by systems optimizing purely for support. For Task B, near-perfect citation coverage (98.77%, 3rd of 50 runs) and zero citation contradict rate demonstrate that Narrative-Aware Reranking and One-Shot ICL enforce robust attribution discipline. The gap in human answer acceptability (80%) highlights an important remaining tension: balancing rigorous evidence constraints with the narrative fluency expected by human assessors, which we identify as the primary direction for future work.

7 Limitations

While our proposed architecture demonstrates robustness on proxy evaluations, we must acknowledge several methodological and systemic limitations inherent in our design.

Proxy-to-Target Gap and Scale. Our development framework relies on SciFact as a proxy for the BioGen test collection, and three distinct gaps limit the fidelity of this transfer. First, there is a *scale gap*: SciFact contains 5,183 abstracts whereas PubMed spans 30 million documents, and retrieval dynamics that hold at small scale — such as BM25 rank correlating with relevance — may degrade unpredictably when the candidate pool is orders of magnitude larger. Second, there is a *linguistic gap*: SciFact claims are researcher-written fact-checking annotations paired with concise, well-structured abstracts, while PubMed encompasses a far broader spectrum of clinical writing — terse case reports, structured RCT abstracts, and discursive narrative reviews — that employ domain-specific terminology, Latin abbreviations, and hedging conventions absent from SciFact. Our NegEx patterns, tuned on SciFact's relatively clean negation markers, may therefore underperform on clinical phrasings such as *"failed to reach significance"* or contradictions embedded in subgroup qualifications. Third, there is a *domain shift*: the MedNLI-tuned T5 classifier was trained on MIMIC-III clinical notes and may not generalize uniformly to the heterogeneous writing styles across six decades of PubMed literature. Collectively, these gaps mean that our proxy-optimized architecture may under-recall contradictions at PubMed scale, and future work should evaluate transfer fidelity explicitly using a held-out labeled subset of PubMed abstracts.

Nuance and Temporal Sensitivity. The current system treats evidence classification as a binary outcome (Support vs. Contradiction). This reductionist approach fails to capture the subtleties of biomedical literature, such as partial support or results conditional on specific patient subgroups. Additionally, the system is currently temporally insensitive; it does not prioritize recent studies, which is a critical oversight in fast-evolving fields like virology and drug-discovery where older contradictions may have been resolved by subsequent research.

Citation Faithfulness and Post-Rationalization. A significant concern in Task B is the phenomenon of *post-hoc rationalization*[24]. Large Language Models like GPT-4o possess extensive parametric knowledge and may generate answers based on prior beliefs rather than retrieved context, subsequently selecting citations that merely align with their output [11]. This risks "confirmation bias," where the model ignores retrieved contradictory evidence in favor of its internal knowledge. Without granular ground-truth labels, distinguishing between genuine evidence synthesis and superficial citation tagging remains an open challenge. The human acceptability gap (80% vs. track mean 95.69%) further suggests that our strict prompting constraints, while enforcing citation discipline, may reduce the naturalness of generated answers.

Model Dependency and Bias. Finally, our reliance on the proprietary GPT-4o model introduces reproducibility barriers. While preliminary experiments with the open-source **GPT-OSS-20B** [17] indicated comparable citation discipline (100% coverage), full-scale validation of open-source alternatives is pending. Moreover, our use of a single one-shot exemplar (Iron/Ferritin) may inadvertently bias the generation towards a specific narrative style, limiting the diversity of the output. Future iterations will explore multi-shot prompting and open-source quantization to mitigate these dependencies.

8 Future Work

Our current investigation lays the groundwork for robust biomedical information retrieval, yet several avenues remain for further exploration.

Firstly, regarding Task A, our reliance on manual regex patterns for the contradiction pipeline, while effective, is not scalable to all linguistic variations of negation. Future work will explore replacing these heuristic filters with lightweight, learned representations using contrastive learning on domain-specific annotated support-contradiction pairs to discover negation patterns, including implicit markers and domain-specific phrasing. Additionally, we aim to investigate automatic query expansion techniques that dynamically utilize topical and narrative concepts to improve the initial BM25 recall for rarer documents.

Secondly, a critical challenge in Generative Search (Task B) is the phenomenon of *post-hoc rationalization.* Current Large Language Models often generate answers based on their internal pre-trained knowledge and subsequently attach citations that appear relevant, without genuinely synthesizing the retrieved text. This creates a dangerous illusion of verification. To address this, we plan to investigate "attribution-first" decoding strategies, where the model is constrained to identify the evidentiary sentence *before* generating the claim, thereby enforcing strict faithfulness to the retrieved context. The 80% human acceptability result motivates exploring open-ended prompting strategies that balance attribution discipline with narrative naturalness.

Finally, we intend to transition our generative pipeline from proprietary APIs (GPT-4o) to fine-tuned open-source models. Establishing a privacy-preserving, locally hosted architecture is essential for clinical deployment, and we aim to benchmark 7B and 70B parameter models to determine the minimum scale required for high-fidelity citation attribution. Concurrently, we propose integrating **Temporal Evidence Weighting** to prioritize recent findings. By analyzing publication dates and citation graphs, systems can identify when historical contradictions have been resolved, preventing the propagation of outdated medical dissent and ensuring alignment with current scientific consensus.

9 Conclusion

This study establishes that epistemic integrity in biomedical QA -specifically the surfacing of contradictory evidence - requires architectural decisions that prioritize holistic reliability and scalability over isolated metric optimization.

Our proxy-based evaluation on SciFact revealed a *"Simplicity Paradox"*. Complex adversarial dense retrieval strategies failed catastrophically on contradictions ($MRR = 0.023$) due to probabilistic miscalibration. Furthermore, we observed that while applying negation filters to standard Dense Retrieval (Variant 3) was effective for contradiction detection (0.766), it significantly degraded support retrieval accuracy (0.786), resulting in lower aggregate system reliability.

In contrast, our **Decoupled Lexical Architecture (Variant 5)** achieved the highest Weighted MRR (0.790). It strikes the optimal balance, maintaining robust support recall (0.810) while effectively surfacing dissent (0.750). This validates that a unified BM25 backbone offers superior stability compared to dense retrieval when navigating the strict constraints of negation filtering. Furthermore, given that scaling dense indices to the 30 million documents of PubMed imposes prohibitive computational costs, the lexical approach remains the only viable strategy for production scale deployment.

For Generative Attribution (Task B), we demonstrated that **Narrative-Aware Reranking** significantly improves context relevance by capturing user intent beyond the query. Furthermore, **One-Shot In-Context Learning** proved decisive, achieving 100% citation coverage in our internal evaluation - validated by the official results (98.77%, 3rd of 50 runs) and a zero citation contradict rate. Our **Nugget Precision (94.55)** (4th of 41 runs in the human-assessed BioACE evaluation) confirms that our system produces highly factual claims when measured at the sentence level.

Together, the official TREC BioGen 2025 results confirm our central thesis: that epistemic integrity requires treating supporting and contradictory evidence as distinct retrieval problems. By transitioning from "black-box" generators to contradiction-aware, fully-attributed synthesizers, we advance toward AI systems that serve as honest arbiters of scientific evidence.

10 Acknowledgments

We express our sincere gratitude to the leadership and management at **GE HealthCare** for their unreserved support and for providing necessary computational infrastructure to conduct this study. We are also thankful to *Anuradha Kanamarlapudi and Nithya Ramesh* for their insightful discussions and constructive feedback during the research and validation. Finally, we thank the organizers of the **TREC 2025 BioGen track** for coordinating the challenge and providing the datasets that made this study possible.

11 Appendices

11.1 Granularity Ablation Study

Table 11 details the performance impact of shifting from document-level to sentence-level NLI inference using the Variant 5 architecture on the SciFact development set.

Table 11: Impact of inference granularity on grounding performance. While document-level inference favors support recall due to context availability, it catastrophically fails to localize contradictions. Sentence-level decomposition yields a 4.8× improvement in contradiction detection, validating its necessity for the BioGen task.

Inference Level	MRR (Sup)	MRR (Con)	Contra. Gain
Document-Level	**0.915**	0.156	–
Sentence-Level (Ours)	0.810	**0.750**	**+380% (4.8×)**

11.2 Adversarial Multi-Query Scoring Details

11.2.1 Motivation and Architecture: Variant 4 (Multi-Query Adversarial Fusion) attempts to overcome the *confirmation bias* inherent in single-query dense retrieval by engineering query diversity. Standard dense retrieval using a claim as query tends to retrieve semantically similar documents that *support* the claim, as embedding models map contextually related text into proximate vector spaces regardless of epistemic polarity (e.g., "ALDH1 increases cancer risk" and "ALDH1 shows no effect on cancer" both embed near each other due to shared domain context).

To force contradiction retrieval, we:

1. **Generate 25 adversarial queries** from negation-heavy templates applied to the base query Q and answer text A:
   - Explicit negation: "contradicts Q", "refutes Q", "disputes Q"
   - Null effect: "no effect of A", "A not significant", "A null effect"
   - Negative outcome: "A fails to show", "A worsens", "A increases risk"
   - Statistical insignificance: "A no difference", "A did not improve"
2. **Retrieve $k = 200$ documents per query** via dense FAISS search (Snowflake Arctic-Embed-M-v2.0), yielding up to $25 \times 200 = 5000$ candidates.
3. **Fuse via Reciprocal Rank Fusion (RRF)** with $k = 60$, capped at 1200 unique documents.
4. **Score candidates** using a two-way NLI-based formula with entailment penalty and negation cue bonus.

11.2.2 Custom Scoring Function: For each candidate document d in the fused pool, we compute:

$$S(d) = \overline{P(\text{Con}|d)} - \lambda \cdot \overline{P(\text{Ent}|d)} + \gamma \cdot \frac{\min(\text{NegCues}(d), C)}{C} \quad (4)$$

where:

- $\overline{P(\text{Con}|d)}$: Mean NLI contradiction probability from two-way inference ($d \rightarrow Q$ and $Q \rightarrow d$) using MedNLI.
- $\overline{P(\text{Ent}|d)}$: Mean NLI entailment probability (two-way). The penalty term $\lambda \cdot \overline{P(\text{Ent}|d)}$ suppresses documents with high support signals.
- $\text{NegCues}(d)$: Count of regex-detected negation patterns (e.g., "no effect", "not significant", "fail(ed) to", "no association", "p > 0.05").
- $\lambda = 0.5$: Entailment penalty weight (empirically tuned).

- $\gamma = 0.1$: Negation cue bonus weight.
- $C = 6$: Cue count cap (prevents over-weighting of documents with excessive negation mentions, e.g., review articles).

Candidates are ranked by $S(d)$ in descending order. We select the top-3 documents where $\overline{P(\text{Con}|d)} \geq 0.35$ (threshold), backfilling by $S(d)$ rank if fewer than 3 meet the threshold.

11.2.3 *Worked Example: Real SciFact Case (QA_ID 51, PMID 45638119).*

Claim (Ground Truth: CONTRADICT): "ALDH1 expression is associated with better breast cancer outcomes."

Document (PMID 45638119): *"In a series of 577 breast carcinomas, expression of ALDH1 detected by immunostaining correlated with* ***poor prognosis****. These findings offer an important new tool for the study of normal and malignant breast stem cells."*

NLI Scores (two-way, Variant 4):

- $d \longrightarrow Q$: {contradiction: 0.99, entailment: 0.01, neutral: 0.00}
- $Q \longrightarrow d$: {contradiction: 0.00, entailment: 1.00, neutral: 0.00}
- $\overline{P(\text{Con}|d)} = (0.99 + 0.00)/2 = 0.50$
- $\overline{P(\text{Ent}|d)} = (0.01 + 1.00)/2 = 0.50$

Negation Cues: 1 ("poor prognosis" detected as antonym of "better")

Final Score:

$$\begin{aligned} S(d) &= 0.50 - 0.5 \times 0.50 + 0.1 \times \frac{1}{6} \\ &= 0.50 - 0.25 + 0.017 = \mathbf{0.267} \end{aligned}$$

Result: $P(\text{Con}) = 0.50 \geq 0.35$ (passes first check), but **Final Score = 0.267 < 0.35 ⇒ MISSED by Variant 4!**

Critical Failure Mode: Although $d \longrightarrow Q$ correctly predicts high contradiction (0.99), the reverse direction $Q \longrightarrow d$ incorrectly assigns maximum entailment (1.00), likely because the classifier confuses "better" and "poor" as related outcome terms. Two-way averaging dilutes the strong contradiction signal to 0.50, and the entailment penalty (0.5 × 0.50 = 0.25) pulls the final score below threshold despite passing the first-pass filter. In contrast, ***Variant 5 (BM25 + sentence-level one-way NLI) correctly classified this as CONTRADICTION****, achieving MRR@3 = 0.75 vs. Variant 4's 0.023. This demonstrates why one-way sentence-level NLI outperforms two-way document-level approaches for nuanced biomedical contradiction detection.*

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

[2] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics* 34, 5 (2001), 301–310.

[3] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (2009). https://api.semanticscholar.org/CorpusID:12408211

[4] Deepak Gupta, Dina Demner-Fushman, William R. Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of TREC 2024 Biomedical Generative Retrieval (BioGen) Track. *ArXiv* abs/2411.18069 (2024). https://api.semanticscholar.org/CorpusID:274306244

[5] Xueqing He, Ziyang Chen, Yi Liu, et al. 2024. Retrieving, Rethinking and Revising: The Chain-of-Verification Can Improve RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, USA.

[6] Md Mosharaf Hossain, Luke Gonzalez, and Kevin Gimpel. 2020. Analysis of the SciFact claim verification task. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

[7] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs. *Bioinformatics* 39, 11 (2023).

[8] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 163–177.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33.

[10] Jimmy Lin, Xueguang Ma, SC Lin, JH Yang, R Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[11] Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 15824–15844. https://aclanthology.org/2023.findings-emnlp.1045

[12] Zheheng Luo et al. 2024. Factual Consistency Evaluation of Summarization in the Era of Large Language Models. In *arXiv preprint or Conference Proceedings*. Verify Exact Venue.

[13] Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. RaFe: Ranking Feedback Improves Query Rewriting for RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, USA, 884–901.

[14] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[15] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

[16] Authors omitted by ACL Anthology preview. 2024. BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. Miami, USA.

[17] OpenAI. 2024. GPT-OSS: Open Source GPT Models. https://github.com/openai/gpt-oss. Accessed: November 2024.

[18] Fabio Petroni, Federico Siciliano, Fabrizio Silvestri, and Giovanni Trappolini. 2024. IR-RAG @ SIGIR'24: Information Retrieval's Role in RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington, DC, USA, 1–4.

[19] S Renjit and Mary Idicula Sumam. 2024. A study of the state-of-the-art approaches and datasets for natural language inference. *Knowledge and Information Systems* (2024). Verify Exact Venue.

[20] Snowflake AI Research. 2024. *Snowflake Arctic-Embed: Text Embedding Models*. https://huggingface.co/Snowflake/snowflake-arctic-embed-m-v2.0

[21] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[22] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[23] David Wadden, Shanchuan Lin, Kyle Lo, et al. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online.

[24] Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. Correctness is not Faithfulness in RAG Attributions. arXiv:2412.18004 [cs.CL] https://arxiv.org/abs/2412.18004

[25] Tianyu Xiong, Jian Sun, Xuezhi Ma, et al. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand.