Dal@TREC25: Improving Biomedical QA with Adaptive Retrieval and Multi-Stage RAG

Jitansh Arora
Dalhousie University
jt390105@dal.ca

Aman Jaiswal
Dalhousie University
am630836@dal.ca

Dr. Juan Ramirez-Orta
Dalhousie University
ramirez.orta.juan@gmail.com

Dr. Evangelos Milios
Dalhousie University
emilios@dal.ca

***Abstract*—This paper presents the design and implementation of Retrieval-Augmented Generation (RAG) pipelines developed for the TREC BioGen 2025 Challenge by the Dalhousie team, which focuses on grounding and attribution in Biomedical Question Answering. Our approach integrates Hybrid Retrieval, Re-Ranking, and Large Language Model (LLM) reasoning to improve Factual Accuracy and Citation Fidelity.**

For Task A (Grounding Answer), we introduce a pipeline that reformulates questions and answer sentences into supportive and contradictory queries, retrieves relevant PubMed articles, and classifies their stance using LLM-based reasoning. For Task B (Reference Attribution), we extend this framework to generate concise, evidence-grounded biomedical answers through a combination of Retrieval–Generation and Generate–then–Retrieve architectures. Multiple prompting strategies and validation mechanisms were explored to balance retrieval coverage, precision, and logical consistency.

The proposed systems emphasize modularity, reproducibility, and adaptability to evolving biomedical corpora, providing a robust foundation for advancing trustworthy, citation-grounded question answering in the biomedical domain.

***Index Terms*—Biomedical Question Answering, Retrieval-Augmented Generation (RAG), Large Language Models, Information Retrieval, Citation Attribution, Evidence Grounding, Hybrid Retrieval, TREC BioGen Challenge**

I. Introduction

Biomedical Question Answering (QA) is a rapidly growing area within Natural Language Processing (NLP) that aims to provide accurate, concise, and contextually appropriate responses to complex biomedical queries. Unlike general-domain QA, Biomedical QA presents unique challenges due to the high-stakes nature of the domain and the need for explicit evidence grounding. Errors in this space are not only academic; they carry significant risks for clinical decision making, research interpretation, and public health guidance. Thus, systems designed for biomedical QA must prioritize factual accuracy, reliability, and transparency above stylistic fluency alone.

Recent advances in Large Language Models (LLMs) have demonstrated remarkable capabilities in generating natural language responses. However, these models remain prone to hallucinations, weak citation grounding, and difficulties in distinguishing between contradictory or irrelevant evidence [1]. This has spurred increasing research into retrieval-augmented generation (RAG) pipelines [2], where external knowledge sources such as PubMed are tightly integrated into the generation process. RAG-based methods aim to mitigate hallucinations by constraining answers to retrieved, domain-relevant content, while ensuring that every generated claim is linked to verifiable sources.

The *TREC BioGen 2025 Challenge*, organized by the National Institute of Standards and Technology (NIST) and the Text REtrieval Conference (TREC), introduces two complementary tasks that directly address these challenges [3]. Task A (*Grounding Answer*) requires systems to identify supporting or contradicting PubMed references for each sentence in a provided biomedical answer. Task B (*Reference Attribution*) tasks participants with generating concise answers to biomedical questions, ensuring that each sentence cites up to three supporting PubMed articles. Together, these tasks highlight the dual demands of retrieval precision and attribution fidelity, making them an ideal benchmark for the next generation of biomedical QA pipelines.

Our systems build upon the methodological foundations established in the 2024 BioGen pilot track [5], which demonstrated the effectiveness of retrieval–reranking–generation pipelines and citation-grounded answer generation. We integrate and extend key strengths from several prior approaches into a unified framework designed for broader evidence coverage and improved attribution fidelity. Specifically, we draw on UR-IW's use of LLM-driven query expansion with Elasticsearch BM25, snippet extraction and reranking, and LLM-based generation [19]; Webis's integration of BM25 with MonoT5/duoT5 reranking and iterative retrieval↔generation refinement [6]; and ielab's fusion of lexical and dense retrieval signals followed by sentence-level attribution [5]. Our design combines these elements within a consistent BM25 + reranking scaffold—incorporating BM25 with Rocchio expansion as noted for H2oloo and MonoT5-based reranking as used by iiresearch [5]—and extends them in three key ways: (i) introducing a dual-branch hybrid retriever with a FAISS/BioBERT dense branch and RRF fusion, (ii) adding a stance-aware mechanism in Task A that explicitly searches for both supporting and contradicting evidence per sentence, and (iii) applying controlled prompting variants (emotional vs. expert) across reformulation and stance checking to isolate prompt-framing effects.

II. BioGen 2025 Challenge Tasks

The BioGen 2025 Challenge consists of two closely related tasks focused on biomedical question answering. These tasks evaluate a system's ability to ground answers in scientific literature and ensure citation accuracy. Below, we summarize the goals and requirements of each task.

A. Task A: Grounding Answer

Task A requires grounding a provided biomedical answer by identifying PubMed references that support or contradict each sentence. Each input consists of a biomedical question, a multi-sentence answer, and existing citations. The system must return up to three new supporting PMIDs and up to three contradicting PMIDs per sentence, restricted to the provided PubMed corpus [3]. Output is submitted in JSON format, ensuring structured attribution.

B. Task B: Reference Attribution

Task B requires generating a 250-word answer to a biomedical question, with each sentence citing up to three PubMed references. The key constraints are brevity ($\leq$ 250 words), citation density ($\leq$ 3 PMIDs per sentence), and corpus restriction (only provided PubMed articles) [3]. Unlike Task A, which validates existing answers, Task B requires constructing new answers with strong evidence attribution.

III. Methodology

A. Task A: Grounding Answer

In Task A, systems are given a biomedical question and a multi-sentence answer. For each sentence, the system must identify up to three new PubMed articles that support the claim (excluding any supporting citations already provided) and up to three that contradict it.

For this task, we submitted two runs: *emotional_prompt* and *expert_prompt*. Both are based on the same Retrieval–Classification–Selection pipeline, with the only difference being the prompting style applied during Query Reformulation and Stance Classification.

1) System Architecture: The pipeline is designed to identify both supporting and contradicting evidence for each answer sentence. It operates in four main stages: Query Reformulation, Retrieval, Stance Classification, and Citation Selection. The overall architecture is illustrated in Fig. 1.

1) **Query Reformulation:** The question and sentence are combined and rewritten into two queries: one targeted at retrieving supportive evidence and the other at retrieving contradictory evidence. This reformulation is performed using the `deepseek-r1` model [7], which is prompted separately for a support rewrite and a contradict rewrite. *Illustrative Example.* For the question *"What will mutation in runx2 affect in the future?"* and the answer sentence *"Mutations can cause bone deformities, height lower than expected, extra teeth and other dental problems,"* the system generates two reformulations:
    - **Support rewrite:** *"RUNX2 gene mutations are associated with skeletal development abnormalities, craniofacial anomalies, dental anomalies including supernumerary teeth, and impaired bone growth leading to short stature."*
    - **Contradict rewrite:** *"Research indicates RUNX2 mutations show no direct association with skeletal malformations, typical stature maintenance, or normative dental development patterns, contradicting previous findings on bone density abnormalities and hyperdontia."*

    The supportive rewrite restates the original claim with clearer biomedical terminology, helping retrieve articles that describe links between RUNX2 mutations and skeletal or dental abnormalities. The contradictory rewrite instead negates the claim, directing retrieval toward studies that challenge these associations. For each sentence, the pipeline generates one supportive query and one contradictory query, ensuring that both supporting and contradictory evidence can be found.
2) **Retrieval:** Each rewritten query is embedded using a BioBERT-based SentenceTransformer (BioBERT-STSb) [8], [9]. The same model was used to precompute embeddings for the PubMed corpus stored in a FAISS index [10] (PMID-mapped). The query embedding is searched against this index and the top-k candidates (k=30) are returned for Stance Classification.
3) **Stance Classification:** Each candidate article is compared to the original answer sentence (not the rewrite) using the `Mistral-7B-v0.3` model [16]. The model is prompted to decide whether the candidate article *supports*, *contradicts*, is *neutral*, or is *irrelevant* to the original answer sentence. Documents labeled as supporting are routed to a support pool; those labeled as contradicting are routed to a contradiction pool. This step verifies whether the documents retrieved by FAISS truly align with the intended stance.
4) **Citation Selection:** From the support and contradiction pools, up to three new citations are selected for each sentence, following task constraints (e.g., no overlap with provided supports). If more than three eligible documents are available, we retain the first three according to the FAISS similarity ranking, since earlier results are more likely to be relevant to the rewritten query.

2) Design Rationale: The pipeline is designed to directly address the requirements of Task A: for each sentence in a biomedical answer, the system must surface new supporting and contradicting evidence from PubMed. To achieve this, the pipeline combines three complementary stages:

- **Reformulation** ensures that both confirming and refuting perspectives are explicitly searched for, rather than relying on a single generic query.
- **Dense retrieval** with FAISS [10] and BioBERT embeddings [8] efficiently retrieves candidate articles from the PubMed corpus.
- **Stance classification and selection** verify whether retrieved documents truly support or contradict the claim, filtering out irrelevant or neutral texts and enforcing task constraints (up to three citations per type).

This modular design grounds each answer sentence in

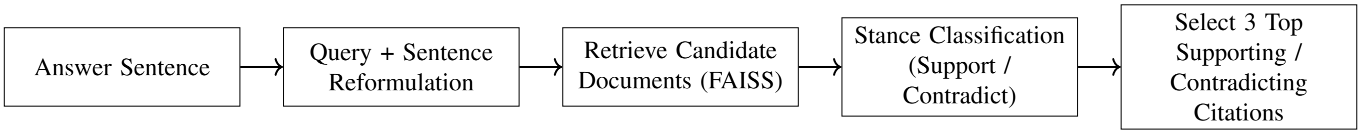
Answer Sentence
Query + Sentence Reformulation
Retrieve Candidate Documents (FAISS)
Stance Classification (Support / Contradict)
Select 3 Top Supporting / Contradicting Citations

Fig. 1. Overview of our approach for Task A (Grounding Answer). Each answer sentence is paired with its corresponding question and reformulated into two distinct queries, a supportive and a contradictory one, using `deepseek-r1` [7]. Both queries are embedded with the BioBERT-based SentenceTransformer model [8], [9] and searched against a FAISS index of PubMed article embeddings [10]. The retrieved candidates are then compared to the original sentence using `Mistral-7B-v0.3` [16], which classifies their stance as *support*, *contradict*, *neutral*, or *irrelevant*. From these pools, up to three supporting and three contradicting citations are selected per sentence, adhering to challenge constraints. This four-stage process - Reformulation, Retrieval, Stance Classification, and Citation Selection - enables identification of both confirming and refuting biomedical evidence. Both submitted runs employ this pipeline, differing only in the prompting strategy applied.

verifiable evidence and ensures that both supportive and contradictory evidence are systematically represented.

3) Prompting Variants: Both runs use the identical pipeline described above; the only difference lies in the prompts used at three points: (1) the Supportive Rewrite, (2) the Contradictory Rewrite, and (3) the Stance Classification prompt. At their core, both the emotional and expert prompts follow the same structure and instructions, but they are reinforced differently depending on the prompting strategy. This choice was not arbitrary: it was guided by recent work showing that different prompting framings can significantly influence LLM output quality.

- In the **emotional prompting** run, the prompts included motivational or affective framing (e.g., phrases such as “This is very important to my career”) to encourage broader, more comprehensive rewrites and greater lexical variety in both retrieval queries and stance checks. This design follows prior findings, which show that affective reinforcement can boost Informativeness and Recall in Large Language Models [17].
- In the **expert prompting** run, the prompts conditioned the model to adopt a biomedical expert persona, emphasizing precise terminology, controlled vocabulary, and domain-specific phrasing. This approach builds on prior work showing that expert-based role prompting can guide models toward more reliable, high-precision outputs [18].

By holding all other components constant and varying only the prompting style, we directly test how prompt design affects the system. A stronger prompt may lead to more effective Query Reformulations, which in turn fetch more relevant candidate documents from FAISS [10]. Likewise, better prompts for Stance Classification may help the model more accurately decide whether an article supports or contradicts a claim. These runs will help assess whether using emotional prompts versus expert prompts, while keeping the rest of the pipeline identical, leads to differences in the quality of retrieval and stance classification. In other words, we aim to understand how alternative prompting approaches might influence performance when applied consistently across the rewrite and stance-checking stages.

B. Task B: Reference Attribution

Task B requires generating a concise biomedical answer (maximum 250 words), where each sentence cites up to three PubMed references from the provided corpus [3]. Unlike Task A, which grounds an existing answer, Task B systems must first generate new answers and then ensure that each claim is supported by verifiable evidence.

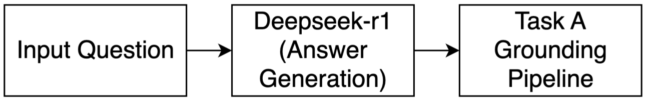

Fig. 2. Overview of our Generate–then–Retrieve pipeline for Task B (Reference Attribution). The input question is first passed directly to `deepseek-r1` [7], which generates a concise biomedical answer without retrieving external documents. Each sentence of this answer is subsequently grounded using the Task A supportive branch, where queries are reformulated and searched against the FAISS index of PubMed embeddings [8]–[10]. Retrieved documents are evaluated for stance using `Mistral-7B-v0.3` [16], and up to three supporting citations are attached per sentence. This approach directly extends the Task A pipeline to answer generation, providing automatic citation attribution while maintaining strict limits of ≤ 250 words and ≤ 3 PMIDs per sentence.

In total, we submitted eight runs for Task B. To keep the description clear, we begin with the two Generate–then–Retrieve systems that extend directly from the prompting strategies introduced in Task A. The remaining six submissions, which involve variations on retrieval–generation pipelines and hybrid designs, are discussed in subsequent subsections.

1) Generate–Then–Retrieve Submissions: The first two runs, *empd* and *expd*, follow a Generate–then–Retrieve design (Fig. 2). In both cases, an answer is first generated by the `deepseek-r1` model [7], and then each sentence is grounded using the pipeline for Task A discussed above.

Unlike Task A, which searches for both supporting and contradicting citations, Task B only requires supporting evidence. Accordingly, the grounding stage here uses only the supportive rewrite and stance-checking components of the Task A system, omitting the contradictory branch.

- *empd*: The question is passed *directly and without modification* to `deepseek-r1`, prompted using the emotional prompting strategy [17]. No Query Reformulation, Pre-Processing, or intermediate retrieval is performed at this stage; the model simply generates an answer. That answer is then grounded using the `emotional_prompt` Task A system, which attaches supporting PubMed citations to each sentence.
- *expd*: The same procedure is repeated with expert prompt-

ing [18]. The question is given as-is to `deepseek-r1` under an expert-role prompt, again with no additional processing. The generated answer is then grounded using the `expert_prompt` Task A system, ensuring consistency between the prompting style used for generation and the grounding stage.

These two submissions directly bridge Task A and Task B: rather than reformulating each sentence into supportive and contradictory queries, the system first produces a complete answer with `deepseek-r1` and then applies the grounding pipeline solely to attach supportive evidence. Because the underlying pipeline is identical, with prompting strategy as the only variable, these runs allow us to directly compare organiser-reported metrics such as Accuracy, Coverage, and Attribution Quality between emotional and expert prompting. We aim to evaluate whether differences in prompting strategy alone lead to measurable variations in the official assessment outcomes.

2) Retrieval–Generation Pipelines: The remaining six runs for Task B use Retrieval–Generation pipelines, where the system first retrieves context from PubMed and then generates an answer conditioned on this evidence. The variants differ along *three* axes: (i) the LLM used for Query Reformulation (`llama-3-70b` [15] vs. `deepseek-r1` [7]); (ii) the control flow: either a fixed pipeline or an *agent-style* loop in which the model judges evidence sufficiency and can trigger further retrieval; and (iii) whether citation validation is applied (with stance-based validation vs. without, retaining only structural checks of $\leq$ 250 words and $\leq$ 3 PMIDs per sentence). We first describe the base system in detail and then highlight how the other variants differ.

The base architecture aligns with common patterns from BioGen 2024 systems [5], but extends them further with strategies not previously documented.

a) System 1 (rrf_monot5-msmarco_llama70b): The base pipeline, illustrated in Fig. 3, consists of a two-branch retrieval stage followed by Re-Ranking, Answer Generation, and Validation. The two retrieval branches are designed to capture complementary evidence: one emphasizes lexical precision via BM25 [11], while the other expands coverage through dense retrieval with query decomposition and embeddings [8], [10]. The full pipeline proceeds as follows:

1) **Keyword-based Retrieval:** In the first branch, the input query is rewritten into a keyword-rich form using `llama-3-70b`, expanding it with synonyms, biomedical terminology, and related phrases. A Rocchio expansion step [12] is applied, and the reformulated query is cleaned again with `llama-3-70b`. A BM25 index is then queried with this reformulated query, and the top-5,000 documents are returned.
2) **Dense Retrieval with Query Expansion:** In the second branch, the *original query* is decomposed into five sub-queries by `llama-3-70b`, each targeting a different aspect of the topic. For each sub-query, the top-1,000 documents are retrieved from a FAISS index [10] of PubMed embeddings generated with BioBERT [8]. The results are then combined into a union set (up to 5,000 documents before deduplication), ensuring broad semantic coverage across different facets of the question.
3) **Merging and Deduplication:** Results from both branches are combined using Reciprocal Rank Fusion (RRF) [13] and de-duplicated to form a unified ranked list of candidate documents.
4) **Re-ranking and Snippet Extraction:** The top-1,000 merged results are re-ranked with MonoT5 [14], and the top-10 abstracts are selected. From these, 30 relevant snippets (sentences) are extracted using the same MonoT5 model, ensuring the answer is grounded in focused, sentence-level evidence rather than full abstracts.
5) **Answer Generation:** The 30 snippets, along with the query and topic, are passed to `llama-3-70b` [15], which generates a candidate answer in zero-shot mode. Redundant or semantically overlapping sentences are removed by fuzzy matching.
6) **Constraint Validation:** All generated answers are first checked against Task B requirements: maximum 250 words and no more than three PMIDs per sentence. If these limits are exceeded, the answer and reason for failure are re-inserted into the context, and generation is retried. This constraint check is always applied.
7) **Evidence Validation:** In a second step, each citation is verified for logical support of its associated sentence using `Mistral-7b` [16]. Answers failing this check are revised through additional generation attempts (up to 20 attempts). This evidence validation module is later disabled in Systems 4–6 to test its impact on efficiency and citation quality.

This system represents our base Retrieval–Generation pipeline, with layered Retrieval, Re-Ranking, and robust Validation.

b) System 2 (rrf_monot5-msmarco_deepseek-r1): The second system is architecturally identical to Fig. 3; the only change is that all query reformulation and sub-query generation steps use `deepseek-r1` [7] instead of `llama-3-70b`. This variant isolates the impact of using a larger, more powerful model with greater parameter capacity for reformulation, compared to the smaller `llama-3-70b` model.

c) System 3 (afmmd): The third system, shown in Fig. 4, extends the previous design with an agent-style loop. After retrieving the top 30 snippets, the snippets are passed back to `deepseek-r1`, which is explicitly asked whether the information is sufficient to answer the question. If the model judges the context incomplete, it reformulates the query to target the missing information, triggers another retrieval cycle, and appends the newly retrieved context to all evidence collected so far. The combined context is then checked again against the original question by prompting the model. This loop can repeat multiple times; if more than five iterations occur without satisfying the sufficiency check, the model is forced to generate an answer using the full set of retrieved context. This adaptive process allows iterative query refinement and progressive context accumulation, making retrieval

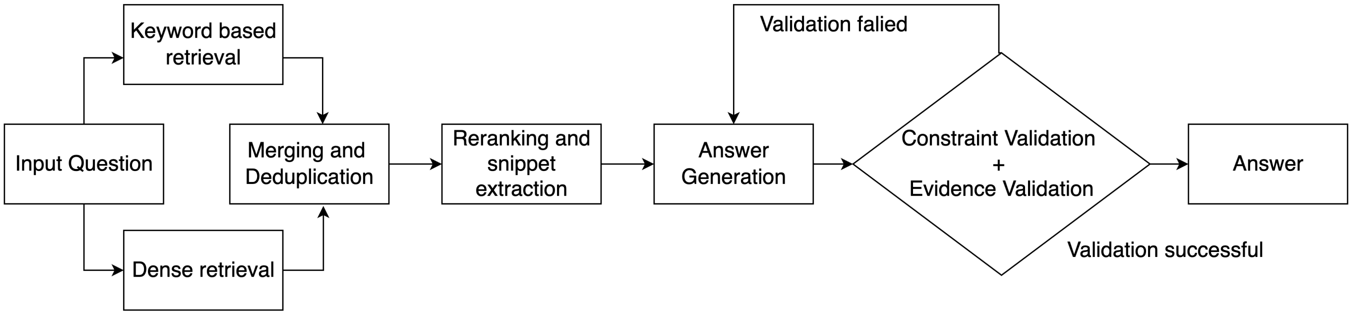

Fig. 3. Overview of the base retrieval–generation architecture for Task B (System 1). The system combines two complementary retrieval branches: a lexical branch using BM25 [11] with Rocchio expansion [12] guided by Llama-3-70B [15], and a dense branch using BioBERT embeddings [8] stored in a FAISS index [10]. Results from both branches are merged via Reciprocal Rank Fusion (RRF) [13] and re-ranked using MonoT5 [14]. The top-ranked abstracts are reduced to 30 evidence snippets and passed, along with the query, to Llama-3-70B for zero-shot answer generation. Validation ensures compliance with BioGen constraints (word and citation limits) and verifies citation–sentence alignment using Mistral-7B-v0.3 [16]. This layered pipeline integrates Hybrid Retrieval, Re-Ranking, and Evidence Validation for robust Citation-Grounded Generation.

TABLE I
TASK A AUTOMATIC EVALUATION (%).

Run	Sup-P	Sup-R	Sup-F1	Con-P	Con-R	Con-F1
emotional_prompt	50.60	67.23	55.53	1.29	1.29	1.20
expert_prompt	45.62	55.67	48.80	0.52	0.26	0.34

more interactive compared to the fixed pipelines in Systems 1 and 2.

d) Systems 4–6 (rmmln, rmmdn, afmmdn): These runs replicate the architectures in Figs. 3 and 4 but with the evidence validation block disabled. In these runs, answers are still constrained to ≤ 250 words and ≤ 3 PMIDs per sentence, but citations are not verified by Mistral-7b for logical support. These lighter pipelines test the significance of the evidence validation step.

IV. Results and Discussion

A. Evaluation Setup

We report the official organizer evaluations for both tasks. As part of the BioGen 2025 protocol [3], results include human judgments for selected top-priority submissions and automatic evaluations using the BioACE framework [4]. For Task A, automatic scoring reports precision, recall, and F1 for supported and contradicted citations. For Task B, automatic scoring reports answer quality metrics (precision, recall, completeness, correctness) and citation attribution metrics (citation coverage, citation support rate, citation contradict rate) [4].

B. Task A: Grounding Answer

Table I shows automatic Task A performance for our two submitted runs. The emotional prompting variant consistently outperformed the expert prompting variant across all six automatic metrics, with the largest gains on supported recall (+11.56 points) and supported F1 (+6.73 points). This suggests that the broader reformulations produced by emotional prompting improved evidence retrieval coverage for supporting citations.

Human assessment for Task A was released for a subset of top-priority runs. For our team, expert_prompt was included in the assessed subset. As shown in Table II, strict support metrics are modest, while relaxed support metrics improve substantially (30.00%), indicating that near-miss but semantically related evidence was often retrieved even when strict matching criteria were not met.

C. Task B: Reference Attribution

Human and automatic results for Task B are summarized in Tables III and IV. The human evaluation covered six of our runs. Two runs were assessed on the full 30-question set, while four runs (expd, empd, afmmd, afmmdn) were assessed on partial sets of 26, 20, 14, and 22 questions, respectively. This partial coverage should be considered when comparing absolute values across all runs.

Across Task B runs, retrieval–generation variants (e.g., rrf_⋆, rmm⋆, afmm⋆) generally achieved stronger automatic citation attribution than the direct generate–then–retrieve variants (expd, empd), particularly in citation support rate and coverage. We also observe a trade-off between maximizing citation support and minimizing contradiction rate: for example, afmmd attains very high citation support (98.20%) with low contradiction (0.45%), whereas runs with near-maximal coverage (e.g., rrf_monot5-msmarco_llama70b, 99.36%) exhibit higher contradiction rates (1.11%). Overall, these results indicate that our hybrid retrieval and re-ranking design is effective for high-coverage attribution, while prompting and validation choices substantially influence citation consistency and answer-level correctness.

V. Conclusion

This paper presented the design and implementation of Retrieval-Augmented Generation (RAG) pipelines developed for the TREC BioGen 2025 Challenge. Our work focused on constructing modular systems that integrate dense and lexical Retrieval, Reranking, and Large Language Model-based reasoning to support Biomedical Question Answering.

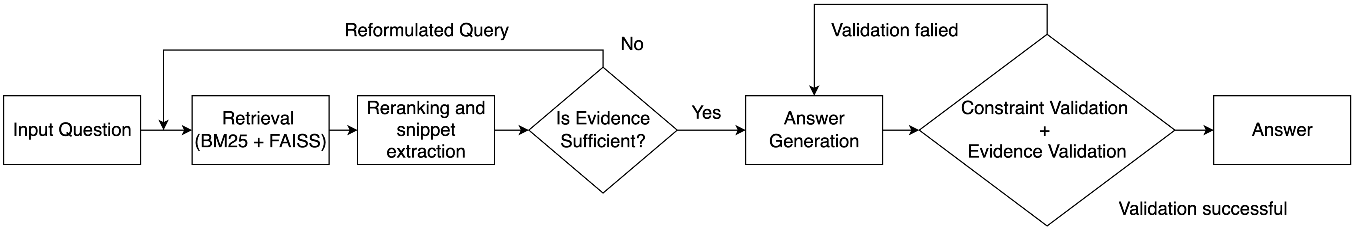

Fig. 4. Overview of the Agent-Style Retrieval–Generation pipeline for Task B (System 3). Following hybrid retrieval with BM25 [11] and FAISS-based dense search [10], the model assesses *evidence sufficiency* using `deepseek-r1` [7]. If insufficient context is detected, the model reformulates the query and triggers an additional retrieval cycle, progressively accumulating relevant evidence. This loop continues until adequacy is reached or a maximum of five iterations is performed. The final answer is then generated and validated to ensure adherence to the limits of the task(≤ 250 words, ≤ 3 PMIDs per sentence) and logical citation support via `Mistral-7B-v0.3` [16]. The adaptive, loop-based design enables iterative reasoning and targeted retrieval refinement for more complete and trustworthy biomedical answers.

TABLE II
Task A human evaluation (%). SSP/SSR: strict support precision/soft recall; SCP/SCR: strict contradiction precision/soft recall; RSP/RSR: relaxed support precision/soft recall; RCP/RCR: relaxed contradiction precision/soft recall.

Run	SSP	SSR	SCP	SCR	RSP	RSR	RCP	RCR
expert_prompt	13.33	13.33	3.33	3.33	30.00	30.00	3.33	3.33

TABLE III
Task B human evaluation (%).

Run	Assessed Qs	Accuracy	Precision	Recall (ST S+R)	Citation Coverage	Citation Support Rate	Citation Contradict Rate	Scope
rrf_monot5-msmarco_llama70b	30	90.00	65.16	23.00	78.74	73.79	3.88	Full
rrf_monot5-msmarco_deepseek-r1	30	100.00	68.05	34.67	85.24	85.53	4.40	Full
expd	26	100.00	71.54	40.00	65.72	53.53	2.65	Partial
empd	20	100.00	68.43	40.50	77.31	60.65	3.56	Partial
afmmd	14	100.00	81.26	35.00	79.69	90.25	1.19	Partial
afmmdn	22	100.00	54.22	29.09	75.99	77.07	6.19	Partial

TABLE IV
Task B automatic evaluation (%).

Run	Precision	Recall	Completeness	Correctness	Citation Coverage	Citation Support Rate	Citation Contradict Rate
rrf_monot5-msmarco_llama70b	94.72	37.63	88.70	66.70	99.36	97.22	1.11
rrf_monot5-msmarco_deepseek-r1	93.82	36.99	77.70	67.19	95.58	96.72	0.41
rmmln	93.94	37.68	88.45	67.26	99.30	97.75	1.12
rmmdn	93.23	36.10	82.33	63.07	97.73	96.36	0.40
afmmd	93.99	37.66	83.29	64.32	85.90	98.20	0.45
afmmdn	94.40	38.08	81.09	64.96	88.24	93.49	0.77
expd	92.20	34.28	80.82	67.92	85.71	81.61	1.24
empd	88.01	37.30	79.75	68.42	94.58	83.52	0.50

For Task-A, we introduced a Stance-Aware grounding pipeline capable of identifying both supporting and contradicting PubMed citations through Reformulation, Retrieval, and Stance Classification. For Task-B, we extended these ideas to Generation, designing multiple Retrieval–Generation and Generate–then–Retrieve variants with consistent validation and attribution constraints.

By isolating prompting strategies and modularizing components, the proposed architectures provide a foundation for analyzing how Prompt Framing, Retrieval Fusion, and Evidence Validation jointly affect the reliability of Biomedical QA systems. While our framework adopts structural elements proven effective in BioGen 2024—such as UR-IW's LLM-based query expansion [19], H2oloo's lexical retrieval with Rocchio [5], iiresearch's MonoT5 reranking [5], we introduce new mechanisms such as hybrid RRF fusion, contradiction-aware retrieval (Task A), and prompt-style interventions. The methodology and design insights described here aim to support reproducible experimentation and further advances in Retrieval-Grounded Generation for the biomedical domain.

Acknowledgments

The authors would like to thank Dalhousie University and the MALNIS lab for their guidance and technical support throughout this work. We also acknowledge funding support from the Natural Sciences and Engineering Research Council of Canada (NSERC). We also acknowledge the methodological groundwork established by the participants of the TREC BioGen 2024 Challenge [5], [6], whose systems laid the foundation for the hybrid retrieval and attribution strategies we expanded upon in this work. Finally, we thank the organizers of the TREC BioGen Challenge for providing the datasets, evaluation framework, and collaborative research environment that enabled this study.

REFERENCES

[1] Z. Ji, N. Yu, J. Xu, *et al.*, “Survey on Hallucination in Natural Language Processing,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[2] P. Lewis, E. Perez, A. Piktus, *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] NIST, “TREC BioGen 2025 Challenge,” 2025. [Online]. Available: https://trec-biogen.github.io/

[4] D. Gupta and D. Demner-Fushman, “BioACE: Automated Citation and Answer Evaluation for Biomedical Generative QA,” *arXiv preprint* arXiv:2602.04982, 2026. [Online]. Available: https://arxiv.org/abs/2602.04982

[5] D. Gupta, D. Demner-Fushman, W. Hersh, S. Bedrick, and K. Roberts, “Overview of the TREC 2024 Biomedical Generative Retrieval (BioGen) Track,” in The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings, NIST Special Publication 1329, 2024. [Online]. Available: https://trec.nist.gov/pubs/trec33/papers/Overview_biogen.pdf

[6] M. Fröbe, L. Gienapp, H. Scells, E. O. Schmidt, M. Wiegmann, M. Potthast, and M. Hagen, “Webis at TREC 2024: Biomedical Generative Retrieval, Retrieval-Augmented Generation, and Tip-of-the-Tongue Tracks,” in The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings, 2024. [Online]. Available: https://trec.nist.gov/pubs/trec33/papers/webis.biogen.rag.tot.pdf

[7] DeepSeek Team, “DeepSeek-R1: Advancing Reasoning in Open Large Language Models,” 2025. [Online]. Available: https://huggingface.co/deepseek-ai/deepseek-r1

[8] J. Lee, W. Yoon, S. Kim, *et al.*, “BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[9] P. Deka, “BioBERT-mnli-snli-scinli-scitail-mednli-stsb (SentenceTransformer),” 2021. [Online]. Available: https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb

[10] J. Johnson, M. Douze, and H. Jégou, “Billion-Scale Similarity Search with GPUs,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[11] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[12] J. Rocchio, “Relevance Feedback in Information Retrieval,” in *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323, Prentice-Hall, 1971.

[13] G. Cormack, C. Clarke, and S. Buettcher, “Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,” in *Proceedings of the 32nd International ACM SIGIR Conference*, pp. 758–759, 2009.

[14] R. Nogueira, Z. Jiang, and J. Lin, “Document Ranking with a Pretrained Sequence-to-Sequence Model,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1412–1418, 2020.

[15] Meta AI, “The Llama 3 Herd of Models,” 2024. [Online]. Available: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

[16] A. Jiang, G. Mitchell, T. Raúl, *et al.*, “Mistral: A Family of Open-Weight Large Language Models,” *arXiv preprint* arXiv:2310.06825, 2023.

[17] Y. Xie, H. Li, B. Y. Lin, *et al.*, “EmotionPrompt: Leveraging Psychology for Large Language Models,” *arXiv preprint* arXiv:2307.11760, 2023.

[18] Z. Wang, C. Xu, X. Tang, *et al.*, “Role prompting: Large Language Models Are Better When Playing Roles,” *arXiv preprint* arXiv:2305.10229, 2023.

[19] S. Ateia and U. Kruschwitz, “Exploring the Few-Shot Performance of Low-Cost Proprietary Models in the 2024 TREC BioGen Track,” in *The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings*, NIST Special Publication 1329, 2024. [Online]. Available: https://trec.nist.gov/pubs/trec33/papers/ur-iw.biogen.pdf