IIUoT at TREC 2025 Retrieval-Augmented Generation Track

YATING ZHANG, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan
HAITAO YU, Institute of Library, Information and Media Science, University of Tsukuba, Japan

In this paper, we present the University of Tsukuba's submission to the TREC 2024 Retrieval-Augmented Generation (RAG) Track. Our work addresses the critical challenges of retrieval instability in deep candidate pools and the prevalent issue of hallucinated attributions in Large Language Models (LLMs). We propose a unified framework that tightly couples a progressive retrieval strategy with an evidence-constrained generation mechanism. To handle the limitations of context windows during retrieval, we employ a listwise LLM-based reranker utilizing a sliding window approach, effectively balancing recall and precision across broad document lists. In the generation phase, we introduce a method for claim-citation alignment that enforces a strict structural dependency, ensuring that every generated statement is immediately preceded by and grounded in a specific reference. By constraining the generation process to verified evidence indices, our system aims to produce highly attributable and factually consistent responses for open-domain information needs.

1 Introduction

The TREC 2024 Retrieval-Augmented Generation (RAG) Track represents a significant step towards evaluating the reliability of systems that integrate external knowledge sources with generative models [8]. While the adoption of Retrieval-Augmented Generation has shown promise in mitigating the factual inconsistencies of standalone LLMs [2], current implementations often face a "granularity mismatch." Retrieval systems typically operate at the passage level, providing broad context, while generation models are expected to synthesize specific, sentence-level claims. This disconnect frequently results in hallucinations where the model generates fluent text that is either unsupported by the retrieved context or incorrectly attributed to it [1]. Furthermore, the reliance on exact term overlap in sparse retrieval methods limits the semantic understanding of queries, necessitating more sophisticated neural reranking strategies that can scale to large candidate lists without sacrificing computational efficiency or context integrity [3]. To address these limitations, our team developed an "Evidence-Constrained" RAG framework designed to enforce strict structural dependencies between the retrieval and generation stages. We argue that high-quality generation is contingent upon a retrieval process that can effectively surface relevant information from deep within a candidate pool. To this end, we implement a progressive sliding window strategy that allows powerful listwise LLM rerankers to process long lists of documents, mitigating the context truncation issues often observed in standard pipelines. Moreover, we posit that attribution should be an intrinsic part of the generation process rather than a post-hoc classification task. Therefore, we devised a structured generation protocol where the model is required to identify citation indices before synthesizing claims. This "evidence-first" approach structurally prevents the model from generating content that lacks a corresponding source in the retrieved context. The remainder of this

Authors' Contact Information: YATING ZHANG, s2426078@u.tsukuba.ac.jp, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan; HAITAO YU, yuhaitao@slis.tsukuba.ac.jp, Institute of Library, Information and Media Science, University of Tsukuba, Japan.

paper provides a comprehensive description of our methodological framework, experimental setup, and evaluation plan.

2 Methodology

Our system architecture is designed to optimize the synergy between retrieval precision and generative faithfulness. The pipeline consists of a progressive retrieval and reranking module followed by a citation-guided generation module. We formalize our approach by defining the set of queries as Q and the indexed document collection as C. Figure 1 illustrates the overall framework.

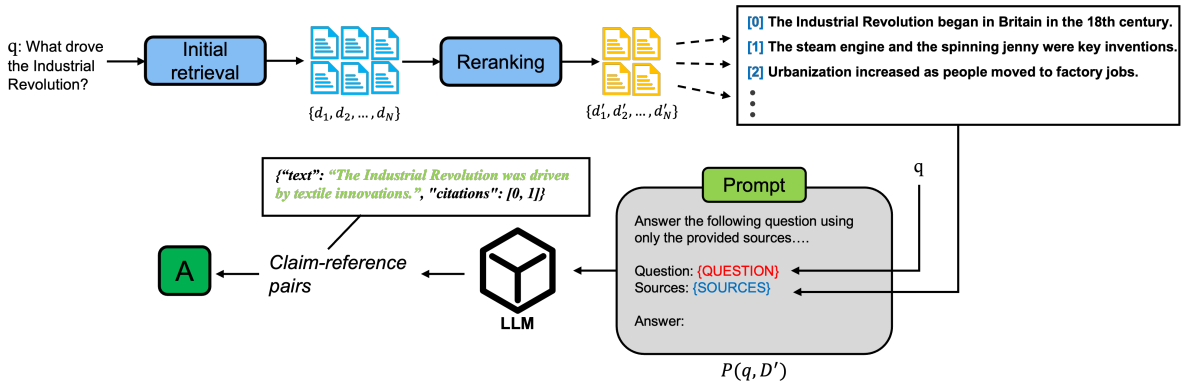

Fig. 1. Illustration of our Evidence-Constrained Retrieval-Augmented Generation pipeline. The process begins with BM25 retrieval and progressive listwise reranking (RankZephyr). The refined documents are then fed into an LLM equipped with dual adapters to generate interleaved citation-claim pairs, ensuring strictly grounded answers.

2.1 Progressive Retrieval and Ranking Synergy

The retrieval stage aims to maximize the quality of the evidence provided to the generator by refining a broad initial search into a concise, high-precision context window. We utilized the official MS MARCO v2.1 segment collection as our evidence base [4]. For each query $q \in Q$, we first extract an initial candidate set of top-N passages using a standard lexical-matching retriever (BM25) via the Anserini toolkit [10]. This process is formally denoted as:

$$D = Retrieve(q, C; N) \tag{1}$$

where $D = \{d_1, d_2, ..., d_N\}$ represents the retrieved candidates and BM25 parameters are set to standard values [5]. While efficient, lexical retrieval often misses semantic nuances. To enhance ranking precision, we subsequently apply a listwise Large Language Model (LLM)-based reranker, specifically RankZephyr, to refine the ordering [6]. However, LLMs have limited context windows, making it impossible to rerank the entire set D in a single pass. To address this, we implemented a **Progressive Sliding Window** strategy. This mechanism efficiently balances recall and precision by reordering only the most promising candidates within localized windows. The reranking function is defined as:

$$D' = Rerank_{Zephyr}(D; w, m, K) \tag{2}$$

Here, w denotes the window size for local reranking (set to 50 in our experiments), m represents the number of top candidates retained and carried over per window (set to 20), and K is the final truncated size ($K \ll N$). By carrying over the top-m documents to the next window, we allow highly relevant documents retrieved at lower ranks (e.g., rank 800) to progressively "bubble up" the list as they are compared against better candidates, effectively propagating relevance signals throughout the entire ranking depth. Following the reranking process, we observed that the top-ranked passages often contained redundant information. To promote evidence diversity while maintaining relevance, we optionally apply a Maximal Marginal Relevance (MMR) adjustment to the top-p ranked passages. The selection logic is governed by:

$$\text{MMR}(d, S) = \arg\max_{d \in D'} \ \lambda \, \text{score}(q, d) - (1 - \lambda) \max_{s \in S} \text{sim}(d, s) \tag{3}$$

where S is the set of already selected passages, sim is a cosine similarity function, and $\lambda \in [0, 1]$ (set to 0.7) controls the trade-off between relevance and novelty. This ensures that the downstream generation stage operates on a compact yet semantically rich document set.

2.2 Citation-Guided Generative Reasoning

The second component of our methodology addresses the generation stage, focusing on integrating retrieved evidence into the reasoning process through explicit citation guidance. This stage enforces a structured alignment between retrieved passages and generated claims, ensuring that all outputs remain verifiable against the supporting documents. Given a query q and its reranked document set D', the generator receives a structured prompt that interleaves the query with the top-K supporting documents. Unlike unconstrained generation, where the model may produce unsupported statements, our approach imposes explicit citation markers during text generation. Formally, we represent the output as a set of claim-reference pairs:

$$A = \{(c_1, R_1), (c_2, R_2), ..., (c_M, R_M)\} \tag{4}$$

where each c_j denotes a generated claim (a natural language statement), and $R_j \subseteq \{0, 1, ..., K-1\}$ is the set of indices of retrieved documents cited as supporting evidence for that specific claim.

To strictly enforce this alignment, we employ a **Dual-Adapter Mechanism** inspired by recent evidence-constrained methods [9]. This mechanism divides the generation process into two distinct logical steps handled by specialized adapter layers within the Llama-3.1 architecture. The first adapter specializes in generating citation spans (the set R_j), while the second adapter produces the associated claims (c_j). This division enforces a sequential dependency: a claim must follow an explicit reference, thereby reducing the risk of hallucinated statements. Concretely, the generator first produces a <reference> block containing a span or index extracted from retrieved evidence, and only then appends a corresponding <claim> block that formulates the grounded statement. This structured sequence enforces explicit attribution for every generated claim, ensuring verifiability against the supporting sources. After generation, a post-hoc filtering module segments the output into sentence-level units. Only those sentences with explicit and valid citations are retained, while redundant sentences are removed via deduplication, and out-of-bound references are discarded to ensure index consistency. This process constrains the generator to operate strictly within the verifiable evidence space.

3 Implementation Details

To ensure reproducibility and robustness, we executed our experiments on a computational cluster equipped with four NVIDIA A100 (40 GB) GPUs running CUDA 12.2. In the retrieval stage, we applied the BM25 algorithm with default parameters ($k_1 = 0.9, b = 0.4$) to the MS MARCO v2.1 segmented corpus. This was followed by listwise reranking using the ZephyrReranker model (`castorini/rank_vicuna_7b_v1_fp16`) via the RankLLM library [6]. The reranking process adopted a progressive sliding-window strategy with a window size of $w = 50$. Within each window, we preserved the top $m = 20$ candidates to carry forward, ensuring that high-quality documents from deep in the initial ranking could propagate to the top. The final selection consisted of the top-100 passages. To further improve result diversity, a Maximal Marginal Relevance (MMR) adjustment with $\lambda = 0.7$ was applied to the top-50 prefix after reranking. For the generation stage, we employed `meta-llama/Meta-Llama-3.1-8B-Instruct` as the base model. This model was equipped with the two aforementioned adapters—`citation_generate` and `claim_generate`—to enable alternating reference-claim generation. We configured the model with deterministic decoding parameters, setting the temperature to 0 and top-p to 1, to minimize generation variance. The prompt context included up to five supporting documents per query to balance information density with context window constraints. The post-processing module utilized regular expressions to parse the XML-like tags (<reference>, <claim>) and validated all indices against the input list D'.

4 Evaluation Metrics

Our system's performance will be officially assessed by NIST using the ground truth relevance judgments (qrels) for the TREC 2024 RAG Track. As these official results are currently pending, we verify our submission against track baselines using a framework that scrutinizes two key dimensions corresponding to our pipeline's architecture. First, regarding **Retrieval Quality**, the effectiveness of our progressive windowing strategy will be quantified using the UMBRELA score [7]. This holistic metric is specifically chosen to assess overall evidence recall and ranking quality beyond simple binary relevance, thereby capturing the granular performance of our listwise reranking approach. Complementing the retrieval analysis, the evaluation of **Generative Faithfulness** will focus on measuring the success of our claim-citation alignment strategy. We anticipate the utilization of citation-aware precision and recall metrics to rigorously verify the accuracy with which generated references support their associated claims [9]. Furthermore, nugget-based evaluation frameworks will be employed to assess factual consistency and information density. We hypothesize that this evidence-constrained approach will demonstrate a significant reduction in hallucinated content and a higher rate of verifiable attribution compared to standard end-to-end RAG baselines in the final official standings.

5 Results and Analysis

The evaluation results for our submitted run (`bm25-rz7b-2025a`) are summarized in Table 1, which compares our system's performance against the median and maximum scores across all TREC 2024 RAG track participants. Our framework demonstrates a distinct trade-off between information recall and attribution reliability.

In terms of citation accuracy (Support Evaluation), our system significantly outperforms the track median, achieving a *weighted_precision* of 0.77 and a *weighted_recall* of 0.76 (compared to the median

of approximately 0.40). These results validate the effectiveness of our proposed evidence-constrained generation mechanism, which ensures that generated claims are strictly grounded in retrieved document segments, thereby effectively mitigating hallucinated attributions.

However, in the nugget-based evaluation (Information Coverage), our system recorded a *strict_vital_score* of 0.20 and a *sub_coverage* of 0.24, which are below the track medians. This suggests that while our model produces highly reliable and attributable responses, it currently prioritizes precision over exhaustive coverage of all sub-narratives. Future work will focus on expanding the retrieval breadth and enhancing the LLM's capability to synthesize more comprehensive summaries without sacrificing factual consistency.

Metric	**Our Run**	**Median**	**Best**	**Worst**
strict_vital_score	0.20	0.41	0.87	0.00
sub_coverage	0.24	0.67	1.00	0.00
weighted_precision	0.77	0.40	1.00	0.03
weighted_recall	0.76	0.39	1.00	0.02

Table 1. Overall evaluation results of our system compared with TREC 2024 RAG track statistics.

6 Conclusion

In this paper, we described the University of Tsukuba's approach to the TREC 2024 Retrieval-Augmented Generation Track. We introduced an evidence-constrained framework that jointly enhances retrieval precision and generative grounding through dual-stage reranking and structured citation alignment. By integrating BM25-based retrieval with a listwise LLM reranker using a sliding window strategy, we addressed the issue of retrieval instability in deep candidate pools. Furthermore, by enforcing explicit claim-citation structure during generation via a dual-adapter mechanism, our system ensures that every generated statement is strictly attributable to a retrieved source. We believe that this explicit modeling of evidence attribution fosters transparency and interpretability—key factors for deploying RAG systems in knowledge-intensive domains. Future work will explore adaptive reranking strategies and reinforcement-based generation optimization to further improve evidence utilization.

References

[1] Tianyu Gao, Howard Yen, Jiahua Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. arXiv preprint arXiv:2305.14627.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[3] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. arXiv preprint arXiv:2309.16769.

[4] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. arXiv preprint arXiv:1611.09268.

[5] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* (2009).

[6] Sahel Sharifymoghaddam, Ronak Pradeep, Andre Slavescu, Ryan Nguyen, Andrew Xu, Zijian Chen, Yilin Zhang, Yidi Chen, Jasper Xian, and Jimmy Lin. 2024. RankLLM: A Python Package for Reranking with LLMs. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[7] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. arXiv preprint arXiv:2406.06519.

[8] Ellen Voorhees, Kirk Roberts, and Abdulaziz Alamri. 2024. TREC 2024 Retrieval-Augmented Generation Track Overview. In *Proceedings of the Text Retrieval Conference (TREC)*.

[9] Shuofei Xia, Xingyao Wang, Jiaqi Liang, Yuxuan Zhang, Wenxuan Zhou, Jie Deng, Feng Yu, and Yang Xiao. 2024. Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation. arXiv preprint arXiv:2407.01796.

[10] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.