CFDA at TREC 2025 Retrieval-Augmented Generation Track (RAG)

Che-Cheng Wu*, **Po-Jen Ko***, **Chuan-Chun Tseng***, **Shao-Hua Wu***, **Pei-Ju Hsieh***, **Li-Yang Chang***
Ming-Feng Tsai‡ and **Chuan-Ju Wang***
* Research Center for Information Technology Innovation, Academia Sinica
‡ Department of Computer Science, National Chengchi University

Abstract

We present a unified retrieval-augmented generation framework for the TREC 2025 RAG Track, addressing key challenges in both retrieval and answer generation for large language models. Our multi-stage retrieval pipeline combines lexical and neural methods through reciprocal rank fusion, integrates nugget-based query decomposition, and applies dense and late-interaction reranking to maximize evidence recall and coverage. For answer generation, we introduce a hierarchical two-level citation process: concise, evidence-grounded answers are first produced for subqueries and then synthesized into a final response with explicit sentence-level attribution. Evaluation uses nugget-based autograder metrics such as Union Nuggets Coverage and Sentence-Support Rate to jointly assess factual coverage and attribution quality. Experiments on the official TREC RAG benchmark demonstrate that our dual-pipeline approach achieves substantial improvements in retrieval recall, answer faithfulness, and citation reliability compared to strong baselines. These results highlight the effectiveness of structured hybrid retrieval and compositional generation for building reliable and transparent RAG systems.

1 Introduction

Large language models (LLMs) such as GPTs and LLaMA (Touvron et al., 2023) have demonstrated impressive capabilities in open-domain question answering, reasoning, and text generation. However, these models inherently rely on the static knowledge encoded during pretraining, making it challenging to access facts or developments emerging after their training cutoff. While post-hoc methods such as supervised fine-tuning and instruction tuning can inject task-specific knowledge or improve generalization, they are fundamentally constrained by the availability, cost, and timeliness of curated training data.

To overcome these limitations, the retrieval-augmented generation (RAG) paradigm has emerged as a compelling solution. By equipping LLMs with the ability to retrieve and condition on external corpora (Guu et al., 2020; Lewis et al., 2020), RAG systems enable models to produce factually grounded, up-to-date responses that extend beyond the scope of their original training data. However, two persistent challenges remain: retrieval modules often miss fine-grained aspects of complex queries, and generated answers may outpace their supporting evidence, especially under tight output constraints.

To address these gaps, we propose a unified RAG system with targeted improvements in both retrieval (R) and augmented generation (AG). For retrieval, we combine lexical and neural methods via Reciprocal Rank Fusion (Cormack et al., 2009), enhance candidate selection with dense rerankers, and apply query decomposition (Jagerman et al., 2023; Chen et al., 2024) to capture diverse information needs. For augmented generation, we introduce a hierarchical, two-level citation mechanism: first, generating concise answers for each sub-query directly supported by evidence; then, integrating these with explicit sentence-level attribution to ensure faithful and transparent generation.

On the official TREC RAG setup, our approach achieves notable gains in both retrieval recall and generation faithfulness compared to strong baselines show on 4. These results highlight the effectiveness of principled hybrid retrieval and structured attribution in advancing the reliability and transparency of RAG systems.

2 Related Work

2.1 Retrieval: Sparse, Dense, and Hybrid Approaches

Early open-domain retrieval relied on sparse lexical methods such as TF-IDF and BM25, which

long served as strong baselines (Mao et al., 2021; Karpukhin et al., 2020; Thakur et al., 2021). With the rise of dense retrieval, models like ANCE (Xiong et al., 2021), DPR (Karpukhin et al., 2020), and Contriever (Izacard et al., 2021) encode queries and documents into embeddings for improved semantic matching. Late-interaction methods such as ColBERT (Khattab and Zaharia, 2020) further enhance dense retrieval by supporting fine-grained token-level interactions. To combine the strengths of both paradigms, hybrid approaches have been proposed. Techniques like SPLADE (Formal et al., 2021b,a) and fusion strategies such as RRF (Cormack et al., 2009) integrate sparse and dense signals. More recently, LLM-based query expansion (e.g., Q2E, Query2Doc) has been introduced to further improve retrieval effectiveness (Jagerman et al., 2023; Chen et al., 2024).

2.2 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) has become a central paradigm for knowledge-intensive NLP (Guu et al., 2020; Lewis et al., 2020). Lewis et al. (Lewis et al., 2020) introduced the RAG architecture, which integrates a dense retriever with a seq2seq generator for open-domain QA. Advances in LLMs such as GPT, LLaMA (Touvron et al., 2023), and Gemini have further extended AG to multi-hop reasoning and long-context synthesis. Systems including WebGPT (Nakano et al., 2021) and Atlas (Izacard et al., 2023) show that iterative retrieval, query rewriting, and tool augmentation can improve factual grounding. Beyond single architectures, emerging AG pipelines incorporate query decomposition, multi-stage retrieval, and hierarchical answer integration, underscoring AG as a structured framework for aligning LLMs with external knowledge.

2.3 RAG Evaluation

Evaluating RAG remains challenging, as outputs can differ in surface form while remaining factually correct, rendering exact match and passage-level relevance insufficient (Fabbri et al., 2022; Jing et al., 2025). Nugget-based evaluation, first explored in TREC QA (Pradeep et al., 2024, 2025), has recently been revitalized as a fact-sensitive alternative. Central to this direction is the development of autograder frameworks, most notably Dietz's rubric-based workbench (Farzi and Dietz, 2024; Dietz, 2024), which leverages LLMs to assess coverage and groundedness through nugget extraction and Q&A tests. Building on this paradigm, systems such as AutoNuggetizer and ARES (Pradeep et al., 2024; Es et al., 2024) operationalize nugget-level recall and faithfulness, achieving strong correlation with human judgments. Complementary open-source toolkits like TruLens (TruLens) extend these insights into practical pipelines for auditing deployed RAG systems. Together, these advances mark a shift toward nugget-centric, autograder-driven evaluation protocols that enable scalable and reliable assessment of factual coverage in RAG.

3 Method

In this section, we present the methodologies employed for both the Retrieval (R) and Augmented Generation (AG) tasks. For the Retrieval (R) task, our objective is to efficiently identify and rank the most relevant documents for each query. For the Augmented Generation (AG) task, we aim to leverage the retrieved documents to generate responses that are accurate, fluent, and well-supported by the retrieved documents.

3.1 Retrieval (R) Pipeline

Our retrieval pipeline, as illustrated in Figure 1, comprises a sequence of increasingly precise stages. First, we use a hybrid filtering that combines BM25 and a lightweight embedding model to narrow down candidate documents. Second, we apply a more powerful dense embedding model for re-ranking. Third, we finalize the ranking using a transformer-based reranker to ensure the precision of the top-0 results. Detailed descriptions of each stage follow in the subsequent subsections.

Post-processing with Query Expansion To enhance retrieval effectiveness, we adopted a broad query expansion framework, wherein the original query q is automatically decomposed into a set of focused sub-queries (nuggets) $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$. Each nugget captures a distinct aspect or facet of the information need, functioning as a targeted expansion of the initial query. This approach is inspired by recent large language model-based query expansion methods, such as Q2E and Q2E/PRF (Jagerman et al., 2023; Wang et al., 2023; Li et al., 2021). Specifically,

- **Q2E** directly prompts a language model to produce multiple expanded queries (nuggets) in free-text form, thereby broadening the semantic coverage of the original query.

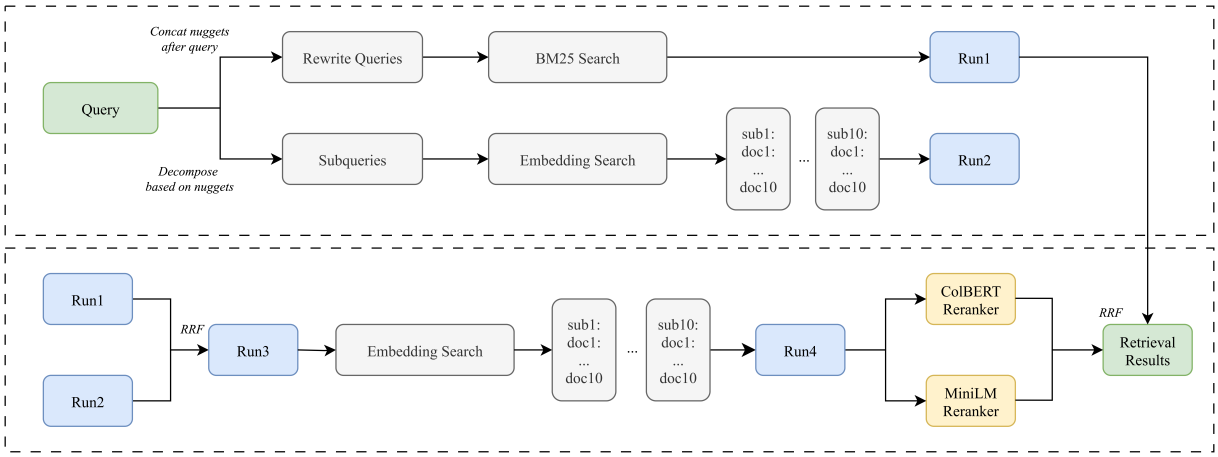

Figure 1: Retrieval effectiveness with different neural reranking strategies on top-50 candidates. Note: in the TREC 2025 Retrieval Only Task, Stage1+2 + MiniLM rerank corresponds to run_id `RRF_minilm_bm25`, and RRF (ColBERT + MiniLM + Stage1+2) corresponds to run_id `RRF_colbert_minlm`.

- **Q2E/PRF** augments this by providing the language model with top-ranked retrieved passages, allowing for context-aware nugget (sub-query) generation that incorporates pseudo relevance feedback.

This nugget-based query expansion framework enables the retrieval pipeline to address different facets of complex queries, improving both recall and the relevance of retrieved evidence. We evaluate the effectiveness of these strategies in Section 4.2.

Stage 1: Hybrid Candidate Selection In this stage, we first issue the expanded query described in Section 3.1 to retrieve its top-5000 documents with BM25, a term-matching model that remains a strong baseline for lexical search (Robertson and Zaragoza, 2009). This list emphasises exact-token overlap, ensuring high-precision matches for rare or domain-specific terms.

In parallel, we embed each decomposed sub-query s_i—designed to cover a single nugget of information—with the compact *all-MiniLM-L6-v2* sentence-transformer [1] and retrieve the top-500 passages per sub-query. With ten sub-queries per query, this yields up to 5000 unique semantic candidates, matching the size of the BM25 pool. Thid semantic retrieval complements BM25 by capturing contextual similarity even when there is little lexical overlap.

Finally, we combine the lexical and semantic rankings with Reciprocal Rank Fusion (RRF), a simple yet effective late-fusion method that rewards documents highly ranked by either source:

$$\mathrm{RRF}(d) = \sum_{r \in \mathcal{R}} \frac{1}{k + \mathrm{rank}_r(d)} \quad (1)$$

where $\mathcal{R}$ is the set of ranked lists (BM25, MiniLM), $\mathrm{rank}_r(d)$ is the rank of d in ranking r, and k is a smoothing constant, yielding a robust and complementary candidate list for downstream stages.

Stage 2: Refined Re-ranking with Qwen3 Embeddings In the second stage of our retrieval pipeline, we further re-rank the candidate set from Stage 1 using a more advanced embedding model. Specifically, we employ *Qwen3-Embedding-0.6B* (Zhang et al., 2025)[2], a lightweight yet high-performing transformer-based encoder from the Qwen3 Embedding series. This model is designed for efficient multilingual representation learning and demonstrates strong performance across various embedding and retrieval benchmarks.

For each sub-query s_i, we embed it with every Stage-1 candidates using Qwen3-Embedding-0.6B, then rank the documents by cosine similarity, and keep the top-10 per s_i. Union the ten result lists yields at most 100 semantically ranked documents per original query. We finally keep 100 top-ranked documents per query to the next-stage cross-encoder re-ranker.

3.2 Reranking Strategies

To further improve retrieval effectiveness, we incorporated an additional reranking stage on top of the

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B

candidate set obtained from Stage 2. Specifically, we applied two neural rerankers:

1. ms-marco-MiniLM-L6-v2[3], a lightweight transformer-based cross-encoder that balances efficiency and semantic precision.

2. ColBERTv1[4], a late-interaction neural retriever that captures fine-grained token-level interactions between queries and documents.

Each reranker independently re-ranked the results from Stage 2. To integrate both lexical and neural signals, we further experimented with Reciprocal Rank Fusion (RRF) among BM25, MiniLM, and ColBERTv1 reranking results.

The detailed comparison of different reranking strategies is presented in Section 4.4.

3.3 Augmented Generation (AG) Pipeline

After constructing a set of candidate passages for each query, we employ an Augmented Generation (AG) pipeline to synthesize accurate and well-supported responses. Our design explicitly addresses the challenge of incorporating a broad range of supporting evidence within a strict output length constraint. To this end, we leverage a two-level citation mechanism, in which sub-queries first gather their own relevant evidence, generate short focused answers, and these answers are then integrated and cited in the final response to the original query.

An overview of the entire AG pipeline is illustrated in Figure 2. This schematic highlights the interplay between query decomposition, multi-stage evidence retrieval, hierarchical answer integration, and sentence support verification, delineating the flow from input query to the final fact-checked output.

Sub-query Decomposition and Evidence Pooling Given an input query q, we first decompose it into a set of focused sub-queries $\mathcal{S} = s_1, s_2, \ldots, s_m$. Each s_i captures a different aspect or nugget of information sought by q. For each sub-query s_i, we select its own evidence pool by re-ranking the initial set of candidate documents retrieved for q. Specifically, we employ the ms-marco-MiniLM-L-6-v2 cross-encoder as the relevance function $f(s_i, d)$ to score each candidate d, and select the top K passages as:

$$T_i = \text{TopK}_{d \in \mathcal{D}_q} f(s_i, d) \tag{2}$$

where $\mathcal{D}_q$ denotes the set of passages retrieved in the initial retrieval stage for query q. This stage ensures that the support for each sub-query is both precise and contextually aligned.

First-Level Citation: sub-query Answer Generation Given the evidence pool T_i, a language model generates a concise, fact-based answer a_i to each sub-query s_i, conditioned only on the texts $\{\text{seg}(d) : d \in T_i\}$. The generation of a_i is subject to several constraints: it must be a single, well-formed sentence, strictly limited in length (e.g., $<$35 words), and every factual claim within must be directly supported by the retrieved passages in T_i. By enforcing these constraints, we ensure that each sub-query answer is both focused and verifiable. Specifically, the answer a_i is produced according to:

$$a_i = G(s_i, \{\text{seg}(d) \mid d \in T_i\}), \tag{3}$$

where G denotes the answer generation model and $\text{seg}(d)$ is the textual content of segment d. The primary objective of this stage is to maximize the density of factual content while minimizing the risk of hallucination or irrelevant elaboration. Each a_i thus forms a *first level* of citation, with its content precisely traceable to a small, query-specific subset of supporting segments. This mechanism enhances the factual reliability of each answer and prepares the outputs for effective downstream integration and traceable citation.

Second-Level Citation: Answer Integration The set of sub-query answers $\{a_1, ..., a_m\}$ is subsequently passed to a separate language model for answer integration. The purpose of this stage is twofold: (1) to synthesize a coherent and comprehensive paragraph y that holistically addresses the original query q, and (2) to maintain explicit traceability between each sentence in the generated response and the specific sub-query answers upon which it is based. During integration, the model reorganizes, rephrases, and combines the sub-query answers as necessary to ensure logical flow, topical cohesion, and avoidance of redundancy. The final response y is thus constructed as:

$$y = F(q, \{[i] : a_i\}_{i=1}^{m}), \tag{4}$$

[3] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

[4] https://huggingface.co/jinaai/jina-colbert-v2

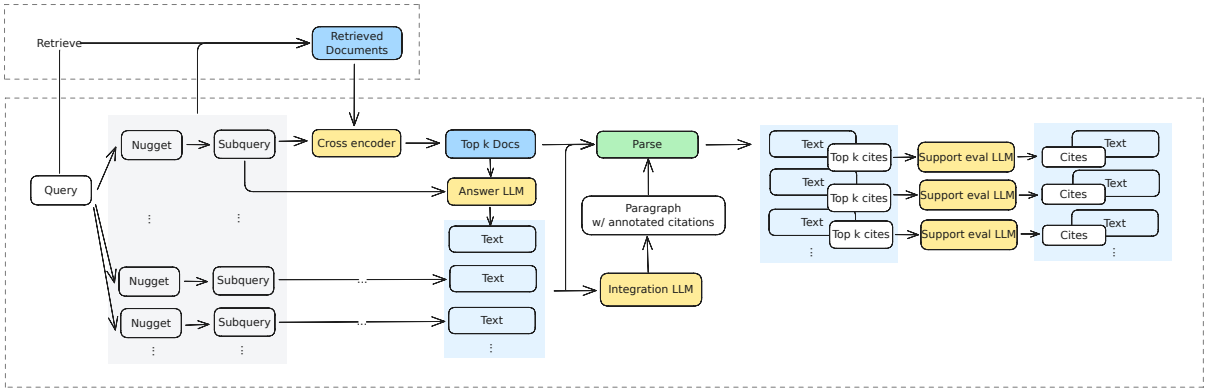

Figure 2: Our AG Pipeline with Query Decomposition, Multi-Stage Retrieval, and Answer Integration.

Method	nDCG@10	Recall@1000	Recall@2000	Recall@5000
BM25 Only	0.3290	0.3466	0.4213	0.5200
BM25 + Q2E	0.4199	0.4458	0.5362	0.6393
BM25 + Q2E + PRF	**0.4337**	**0.4466**	**0.5384**	**0.6423**

Table 1: Retrieval performance of BM25, BM25+Q2E, and BM25+Q2E+PRF on the TREC 2024 RAG dataset.

where F denotes the integration model that generates the output paragraph from the original query and the set of indexed sub-query answers. Each output sentence is annotated with markers $[i]$ indicating its provenance—i.e., which sub-query answer(s) it draws upon. This structure constitutes a *second level* of citation, enabling each sentence in the final response to be mapped not only to the sub-query answer(s) that support it, but, via the first level, transitively to the original evidence passages retrieved for those sub-queries. Beyond maximizing factual coverage and transparency, this hierarchical integration supports fluent transitions and coherence across sentences, resulting in an answer that is both information-rich and naturally readable within the target length constraint.

Sentence Support Verification For each sentence y_k in the final output, we identify its set of sub-query markers I_k. We then aggregate all candidate supporting passages by taking the union of the evidence pools for those sub-queries, i.e.,

$$U_k = \bigcup_{i \in I_k} T_i \tag{5}$$

To ensure that every claim in y_k is well-supported, we apply a binary support verification function $h(y_k, \text{seg}(d)) \in \text{yes}, \text{no}$ to each candidate $d \in U_k$, retaining only those passages that directly support the sentence content.

Key Insights This multi-stage design allows our AG pipeline to maximize evidence coverage and factual correctness, even under strict length constraints. By structuring generation around both sub-query-level and integrated-level citation, we provide clear chains from each sentence in the final response back to concrete evidence passages. Experimental results in Section 4.5 demonstrate the effectiveness of this approach.

3.4 Augmented Generation Metrics

In this section, we introduce metrics that quantify factual coverage and evidential grounding for our augmented-generation pipeline.

Factual Coverage We report *Union Nuggets Coverage* (UNC), a nugget-based metric designed to quantify the degree to which a system covers the assessor-defined atomic facts ("nuggets") for each topic. Our evaluation employs the **autograder** toolkit (Dietz, 2024)[5], a third-party system that automatically constructs topic-specific nugget rubrics and leverages a large language model (LLM) to assign support scores to system outputs. Given a topic q, let $\mathcal{N}_q^\star$ denote the set of gold nuggets produced by autograder, and $\mathcal{Y}_q$ the set of system-generated sentences.

For each nugget $n \in \mathcal{N}_q^\star$, let $s(n)$ denote the highest LLM-assigned support score (on a 1–5 scale) across all sentences in $\mathcal{Y}_q$ that mention n.

[5] https://github.com/TREMA-UNH/rubric-grading-workbench

UNC at a support threshold τ is defined as

$$\mathrm{UNC@}\tau(q) = \frac{\left|\{n \in \mathcal{N}_q^\star \mid s(n) \geq \tau\}\right|}{|\mathcal{N}_q^\star|} \quad (6)$$

that is, the proportion of unique gold nuggets that receive a support score of at least τ from the LLM grader.

We report UNC@τ macro-averaged across all topics; higher values indicate improved nugget recall under the specified support criterion. It should be noted that both autograder and LLM-based scoring are third-party tools and are not part of the official TREC RAG protocol.

Evidential Grounding To rigorously evaluate factual attribution in generated answers, we introduce the *Sentence-Support Rate (SSR)*, which quantifies the extent to which each generated sentence is verifiably grounded in its cited evidence. For a generated sentence, denoted as y_k, and its associated set of citations C_k, we resolve these citations to obtain the corresponding evidence segments $\Gamma(C_k)$. Given the topic narrative q, the sentence y_k, and the concatenated evidence segments, a large language model grader assigns a discrete support score $g_k \in \{1, 2, 3, 4, 5\}$ according to a rubric ranging from "Not supported" to "Strongly supported."

We report two complementary indices to summarize sentence-level support. The first, SSR@τ, is defined as

$$\mathrm{SSR@}\tau = \frac{1}{|\mathcal{Y}|} \sum_{y_k \in \mathcal{Y}} \mathbf{I}[\, g_k \geq \tau \,] \quad (7)$$

where $\mathcal{Y}$ denotes the set of all generated sentences, g_k is the support score assigned to sentence y_k, and $\mathbf{I}[\cdot]$ is the indicator function that equals 1 if the argument is true and 0 otherwise. This metric measures the proportion of sentences that achieve at least a specified support threshold τ, thereby reflecting the system's ability to produce claims that are well supported by explicit evidence.

The second index, the *Mean Support Score (MSS)*, is given by

$$\mathrm{MSS} = \frac{1}{|\mathcal{Y}|} \sum_{y_k \in \mathcal{Y}} \frac{g_k}{S_{\max}} \quad (8)$$

where g_k is the support score for sentence y_k, and $S_{\max}$ denotes the maximum possible support score (e.g., $S_{\max} = 5$ in our assessment), thus normalizing MSS to the $[0, 1]$ range. MSS provides a normalized average of support scores across all sentences, enabling fine-grained comparison of grounding quality, including partially supported content.

4 Experiments and Results

This section presents our experimental evaluation across three dimensions: query expansion, retrieval pipeline performance, and the effectiveness of different Augmented Generation (AG) strategies. We begin with a high-level overview of our evaluation framework, followed by detailed breakdowns of each component.

4.1 Evaluation Methods

To ensure methodological consistency and comparability, we adopt a unified evaluation framework anchored in the MS MARCO V2.1 Segmented Corpus[6]. Our evaluation spans two complementary settings: retrieval-oriented evaluation (**R**) and augmented generation evaluation (**AG**), both following the official TREC RAG track resources and protocols.

Retrieval (R) Retrieval performance is assessed using the official RAG 2024 topics (301 queries) alongside the corresponding UMBRELA-based relevance judgments (Upadhyay et al., 2024b,a), which leverage LLM-based assessors as adopted in the TREC 2024 RAG track.

We employ the official `trec_eval` toolkit[7] to report standard information retrieval metrics, including Recall and nDCG. This evaluation setup enables us to comprehensively measure both the breadth (coverage) and depth (ranking precision) of our retrieval pipeline and query expansion strategies.

Augmented Generation (AG) For augmented generation, we evaluate on the official RAG 2025 topics using the nugget-based automatic evaluation framework. Specifically, we employ an autograder (Dietz, 2024)[8] to compute **nugget-level coverage**, which quantifies the inclusion of relevant information units within system outputs.

To further assess factual grounding, we implement an evaluation module that leverages LLMs to verify the **support relationship** between each generated sentence and its cited passage(s). This

[6] https://trec-rag.github.io/annoucements/2025-rag25-corpus/

[7] https://github.com/usnistgov/trec_eval

[8] https://github.com/TREMA-UNH/rubric-grading-workbench

dual evaluation framework captures both (1) nugget coverage as a proxy for informativeness and (2) citation-supported correctness as a measure of factual accuracy. Together, these metrics offer a comprehensive assessment of AG quality that goes beyond surface-level lexical matching.

4.2 Query Expansion Effectiveness

We evaluate the impact of query expansion strategies introduced in Section 3.1, specifically comparing vanilla query expansion (Q2E) and Q2E with pseudo relevance feedback (Q2E/PRF). Our experimental setup contrasts two conditions: (1) expansion based solely on the original query (Q2E, "without context"), and (2) expansion with the top-3 retrieved passages provided to the LLM for contextual grounding (Q2E/PRF, "with context").

Table 1 shows the retrieval effectiveness of these methods on the development set. We observe that Q2E/PRF consistently outperforms both vanilla Q2E and the BM25 baseline across all metrics (nDCG@10, Recall@1000/2000/5000). This demonstrates that contextualized query expansion leveraging pseudo relevance feedback allows the LLM to generate more targeted and effective subqueries, thereby significantly improving downstream retrieval recall and ranking quality.

4.3 Retrieval Pipeline Performance

As we can see, Table 2 presents the retrieval effectiveness across our multi-stage pipeline, and reveals several important insights.

In the first stage, our chief goal is efficient candidate filtering. Despite its simplicity, BM25 performs strongly here, achieving an nDCG@10 of 0.4576 and Recall@5000 of 0.6554. This high recall makes it particularly well-suited for narrowing the corpus before further processing. By contrast, a lightweight embedding-only retriever yields lower recall (0.6145), suggesting that while embeddings can capture semantic relevance, they may lack the lexical precision needed for broad coverage.

We observe that combining BM25 and embedding retrieval via Reciprocal Rank Fusion (RRF) substantially boosts coverage: Recall@5000 rises to 0.7555. This underscores the well-known complementarity of lexical and semantic signals in hybrid retrieval systems.

The refined re-ranking with the Qwen model in Stage 2 delivers clear improvements—both in ranking quality (nDCG of 0.5027) and recall (0.7667)—affirming the value of additional semantic precision after initial filtering.Notably, when we apply RRF between Stage 1 (BM25) and Stage 2, the best performance emerges: nDCG improves to 0.5432 and Recall@5000 reaches 0.7734. This demonstrates that even after semantic re-ranking, the distinct contributions of BM25 continue to enhance the final retrieval outcomes.

Together, these findings support the multi-stage pipeline design: Stage 1 (especially BM25-based retrieval) efficiently achieves high recall, providing a robust foundation; Stage 2 fine-tunes relevance with deeper semantic modeling; and fusing both stages via RRF yields the most robust overall performance.

4.4 Impact of Neural Rerankers

To further investigate the contribution of neural reranking strategies, we applied both ColBERTv1 and MiniLM cross-encoder on the top-50 candidates obtained from Stage 2. We additionally experimented with Reciprocal Rank Fusion (RRF) to combine the outputs of both rerankers with the Stage 1 and Stage 2 results. Table 3 summarizes the comparative performance across different reranking configurations.

From Table 3, we observe that ColBERTv1 provides the strongest individual reranking performance, achieving the best nDCG@10 of 0.5928, while MiniLM shows slightly weaker performance across all cutoffs. However, when combining both rerankers with the Stage 1 and Stage 2 results using RRF, the fused run surpasses individual rerankers at shorter cutoffs, achieving the best nDCG@3 (0.6307) and nDCG@5 (0.6167). These findings highlight the complementary strengths of cross-encoder and late-interaction rerankers, suggesting that multi-reranker fusion can further improve retrieval effectiveness beyond relying on a single reranker.

4.5 Augmented Generation Pipeline Performance

To establish a fair and informative comparison, we benchmark our augmented-generation (AG) pipeline against a single-pass baseline representative of standard RAG practices. For each topic, the baseline system retrieves relevant passages, concatenates them, and invokes the LLM once to produce an answer, requesting explicit citations but omitting query decomposition, passage reranking, and iterative evidence integration. This reflects

Method	nDCG@10	Recall@1000	Recall@2000	Recall@5000
Stage 1 – BM25 only	0.4576	0.4654	0.5553	0.6554
Stage 1 – Embedding only	0.3100	0.4090	0.5013	0.6145
Stage 1 – RRF Fusion	**0.4636**	**0.5451**	**0.6479**	**0.7555**
Stage 2 – Qwen re-ranking	0.5027	0.6180	0.7024	0.7667
RRF (Stage 1 + Stage 2)	0.5209	**0.6300**	**0.7151**	0.7679
RRF (BM25 + Stage 2)	**0.5432**	0.6293	0.7130	**0.7734**

Table 2: Comparative Retrieval Performance of Stage 1, Stage 2, and RRF-based Fusion.

Method	nDCG@3	nDCG@5	nDCG@10
Stage1+2 + ColBERTv1 rerank	0.6063	0.6010	0.5928
Stage1+2 + MiniLM rerank	0.5894	0.5768	0.5568
RRF (ColBERT + MiniLM + Stage1+2)	**0.6307**	**0.6167**	**0.5976**

Table 3: Retrieval effectiveness with different neural reranking strategies on top-50 candidates. Note: in the TREC 2025 Retrieval Only Task, Stage1+2 + MiniLM rerank corresponds to run_id `RRF_minilm_bm25`, and RRF (ColBERT + MiniLM + Stage1+2) corresponds to run_id `RRF_colbert_minlm`.

a standard, non-compositional approach in recent RAG systems.

We evaluate system performance along two key dimensions: **(1) Factual Coverage**, which measures how comprehensively the generated answers address all relevant information needs; and **(2) Evidential Grounding**, which assesses the degree to which individual claims in the output are explicitly supported by retrieved evidence.

These complementary perspectives provide a holistic assessment of both the breadth of information provided and the factual reliability of the generated content.

LLM-Assessed Union Nuggets Coverage As shown in Table 4, our augmented-generation (AG) pipeline achieves markedly higher UNC@4 scores compared to the baseline for both LLMs. In particular, for Llama-3.1-8B, our method raises the UNC@4 from 0.4819 to 0.9352, representing a 94% relative improvement. For GPT-4.1-Mini, the UNC@4 increases from 0.6981 to 0.9457, a 35% relative gain.

These results demonstrate that hierarchical query decomposition and iterative evidence integration substantially enhance factual coverage. The performance gap is especially pronounced for the smaller Llama model, suggesting that a structured AG pipeline can effectively compensate for the limited capacity of smaller LLMs.

Sentence-Level Support Assessment As shown in Table 5, our AG pipeline attains sentence-level support scores broadly comparable to the baseline. While the baseline achieves slightly higher values under stricter thresholds (SSR@4 and MSS), our method performs on par at SSR@3. This reflects a natural trade-off: by integrating more diverse evidence, the system generates denser outputs where maintaining uniformly high support for every sentence becomes more challenging.

When viewed alongside the UNC results in Table 4, this trade-off proves favorable. The pipeline yields substantial gains in factual coverage—up to a 94% relative improvement—while only marginally lagging in per-sentence support. This indicates that our approach successfully expands coverage without materially compromising evidential grounding, producing answers that are both more comprehensive and comparably reliable.

5 Conclusion

This paper introduces a unified retrieval-augmented generation (RAG) framework that systematically advances both retrieval and augmented generation through pipeline designs and principled evaluation. On the retrieval side, the proposed multi-stage architecture incorporates hybrid lexical and neural retrieval, nugget-based query decomposition, and successive dense and late-interaction reranking, collectively facilitating robust evidence assembly for complex information needs.

Model	Method	UNC@4
GPT-4.1-Mini	Ours	**0.9457**
	Baseline	0.6981
Llama-3.1-8B	Ours	**0.9352**
	Baseline	0.4819

Table 4: Union Nuggets Coverage (UNC) for different models and methods. Note: for Ours, GPT-4.1-Mini corresponds to run_id `ag-v2-gpt`, and Llama-3.1-8B corresponds to run_id `ag-v2-llama`.

Model	Method	SSR@3	SSR@4	MSS
GPT-4.1-Mini	Ours	**0.9858**	0.8321	0.7516
	Baseline	0.9794	**0.8336**	**0.7624**
Llama-3.1-8B	Ours	**0.9813**	0.8251	0.7473
	Baseline	0.9750	**0.8419**	**0.7635**

Table 5: Sentence-level attribution results: SSR@3, SSR@4, and MSS for each system. Note: for Ours, GPT-4.1-Mini corresponds to run_id `ag-v2-gpt`, and Llama-3.1-8B corresponds to run_id `ag-v2-llama`.

For augmented generation, we present a hierarchical pipeline that integrates sub-query decomposition, two-level citation, and sentence-level support verification. This design not only improves the factual comprehensiveness of system outputs but also strengthens the transparency and traceability of evidence attribution, even under stringent output constraints. To enable rigorous and nuanced evaluation of our approach, we develop dedicated metrics, including Union Nuggets Coverage and Sentence-Support Rate, which provide a fine-grained and faithful assessment of both coverage and attribution.

Empirical results on the TREC RAG benchmark, in conjunction with our custom evaluation framework, demonstrate substantial improvements in retrieval effectiveness and answer quality relative to strong baselines. These findings affirm the efficacy of a principled, dual-pipeline approach that aligns structured retrieval, hierarchical generation, and fine-grained evaluation. Future research will focus on further unifying retrieval and generation through end-to-end modeling, extending attribution strategies to more abstractive and multi-hop scenarios, and advancing robust, scalable evaluation for real-world RAG deployments. Collectively, our contributions offer a transparent and extensible foundation for the next generation of evidence-grounded language systems.

References

Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, generate and refine: Query expansion with LLMs for zero-shot open-domain QA. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11908–11922, Bangkok, Thailand. Association for Computational Linguistics.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Laura Dietz. 2024. A workbench for autograding retrieve/generate systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1963–1972, New York, NY, USA. Association for Computing Machinery.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Naghmeh Farzi and Laura Dietz. 2024. Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '24)*, Washington, DC, USA. ACM.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models.

Xiaonan Jing, Srinivas Billa, and Danny Godbout. 2025. On a scale from 1 to 5: Quantifying hallucination in faithfulness evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7765–7780, Albuquerque, New Mexico. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

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. NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2021. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 180–190, New York, NY, USA. Association for Computing Machinery.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

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 (Round 2)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

TruLens. Trulens: Move from vibes to metrics.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2024a. A large-scale study of relevance assessments with large language models: An initial look.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024b. Umbrela: Umbrela is the (open-source reproduction of the) bing relevance assessor.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models.