

H-RAG at SemEval-2026 Task 8: Hierarchical Parent–Child Retrieval for Multi-Turn RAG Conversations

Passant Elchafei¹, Hossam Emam³, Mohamed Alansary³,
Monorama Swain¹ and Markus Schedl^{1,2}

¹Johannes Kepler University Linz, Institute of Computational Perception, Linz, Austria

²Linz Institute of Technology, Artificial Intelligence Lab, Austria

³Independent Researcher, Egypt

{passant.elchafei, monorama.swain, markus.schedl}@jku.at
{hossamsalahemam, mohamed.alansary.c}@gmail.com

Abstract

We present H-RAG, our submission to SemEval-2026 Task 8 (MTRAGEval), addressing both Task A (Retrieval) and Task C (Generation with Retrieved Passages). Task A evaluates standalone retrieval quality, while Task C assesses end-to-end retrieval-augmented generation (RAG) in multi-turn conversational settings, requiring both accurate answer generation and faithful grounding in retrieved evidence. Our approach implements a hierarchical parent–child RAG pipeline that separates fine-grained child-level retrieval from parent-level context reconstruction during generation. Documents are segmented into overlapping sentence-based child chunks, while full documents are preserved as parent units to provide coherent context. Retrieval combines hybrid dense–sparse search, tunable weighting, and embedding-based similarity rescoring over child chunks. Retrieved evidence is aggregated at the parent level and supplied to an instruction-tuned language model for response generation. H-RAG achieves an nDCG@5 score of 0.4271 on Task A and a harmonic mean score of 0.3241 on Task C (RB_agg: 0.2488, RL_F: 0.2703, RB_llm: 0.6508), underscoring the importance of retrieval configuration and parent-level aggregation in multi-turn RAG performance.

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a standard paradigm for grounding large language model outputs in external knowledge sources (Lewis et al., 2020). By retrieving relevant documents at inference time, RAG aims to improve factual accuracy and reduce hallucinations in knowledge-intensive tasks. While single-turn RAG pipelines have demonstrated strong performance, multi-turn conversational settings introduce additional challenges. User queries may depend on earlier dialogue turns, contain implicit references,

or evolve in intent over time. These characteristics significantly increase the difficulty of both retrieving relevant evidence and maintaining faithful, grounded generation.

SemEval-2026 Task 8, MTRAGEval (Rosenthal et al., 2026b) provides a comprehensive benchmark for evaluating multi-turn RAG systems under realistic conversational conditions (Katsis et al., 2025). The shared task includes Task A (Retrieval), which evaluates standalone retrieval performance, and Task C (Generation with Retrieved Passages), which assesses full end-to-end RAG. Task C jointly measures answer quality, document-level grounding, and faithfulness, reflecting practical deployment requirements for RAG systems.

In this work, we present **H-RAG**, our system submission participating in both Task A and Task C. H-RAG is motivated by a fundamental design trade-off in RAG systems: fine-grained text chunks improve retrieval precision, whereas larger contextual units provide more coherent input for generation. To balance these competing objectives, we adopt a *hierarchical parent–child representation that decouples retrieval granularity from generation context reconstruction*. Unlike prior hierarchical retrieval systems such as HRR (Singh and Mohapatra, 2025) and HiRAG (Huang et al., 2025), which introduce learned reranking or graph-structured knowledge integration, our contribution lies in the integrated design and configuration study of an unsupervised parent–child retrieval pipeline tailored to the multi-turn MTRAGEval setting. We further analyze how retrieval-side design choices influence end-to-end grounding and answer quality.

During document ingestion, each document is segmented into overlapping sentence-based child chunks optimized for semantic and lexical retrieval. In parallel, the original documents are preserved as parent units to provide coherent context once relevant evidence has been identified. At query time, the system optionally applies lightweight

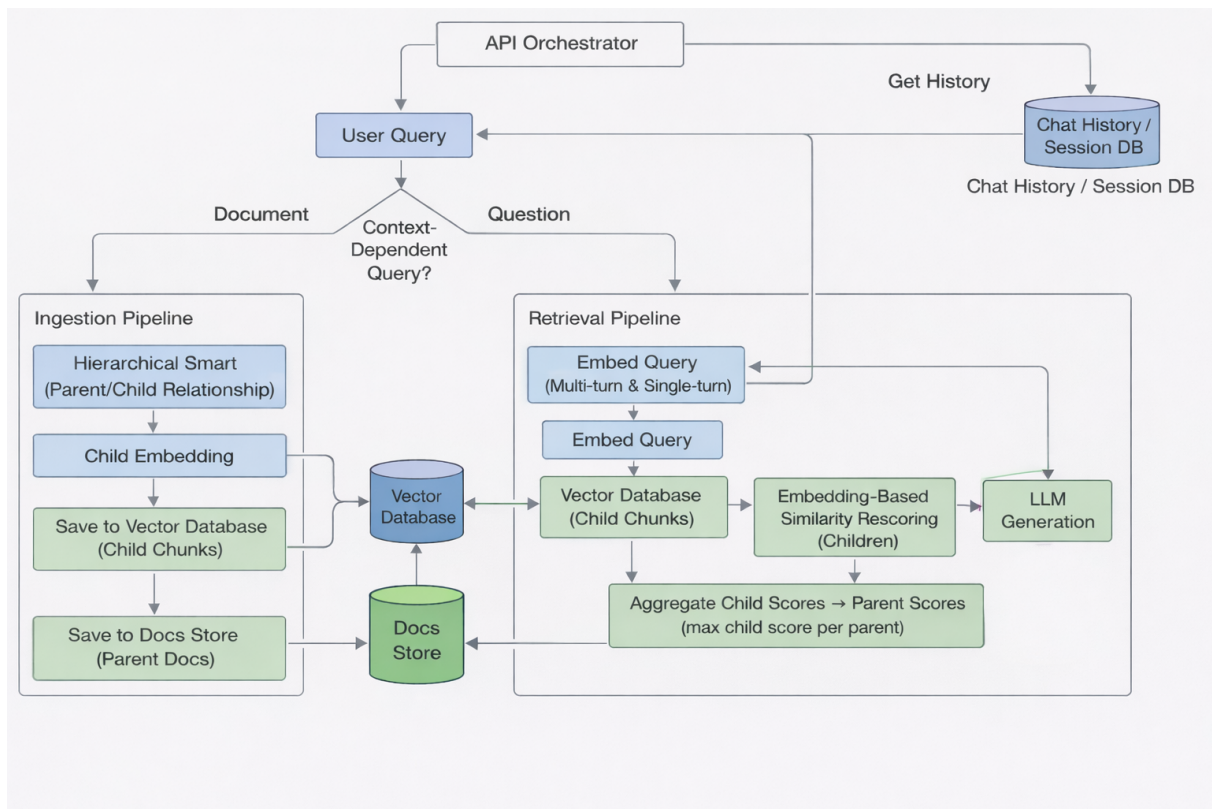


Figure 1: Overview of the proposed H-RAG pipeline for Task C, illustrating hierarchical parent–child document ingestion, hybrid retrieval over child chunks, embedding-based similarity rescoring, parent-level aggregation, and LLM-based answer generation in multi-turn conversational settings.

query rewriting for context-dependent conversational queries. Retrieval is performed using hybrid dense–sparse search with a tunable weighting parameter controlling the balance between embedding similarity and lexical matching (Cuconasu et al., 2024). Retrieved child chunks are refined using embedding-based similarity rescoring, and evidence is aggregated at the parent level by assigning each parent document the maximum score of its associated children (Singh and Mohapatra, 2025; Huang et al., 2025). The top-ranked parent documents are then supplied to an instruction-tuned language model for response generation.

2 Related Work

Multi-turn Conversational RAG. Recent work has emphasized evaluating retrieval-augmented generation (RAG) systems in multi-turn conversational settings, where dialogue context, topic shifts, and underspecified queries increase retrieval complexity. MTRAG (Katsis et al., 2025) introduces a human-authored benchmark for conversational retrieval and grounded generation. CORAL (Cheng et al., 2025) and RAD-Bench (Kuo et al., 2025)

further demonstrate the challenges of retrieval and reasoning across dialogue turns. These benchmarks highlight that conversational RAG performance depends heavily on robust cross-turn evidence selection. Our system is designed for this setting and evaluated under the MTRAGEval protocol.

Hybrid Dense–Sparse Retrieval. Hybrid retrieval combining dense embeddings and lexical matching has become a strong baseline in RAG systems. The Power of Noise study (Cuconasu et al., 2024) demonstrates that combining sparse and dense embeddings improves robustness over single-retriever approaches. DAT (Hsu and Tzeng, 2025) proposes query-adaptive weighting of dense and sparse scores. In contrast, H-RAG adopts a lightweight hybrid strategy with a statically tuned interpolation parameter α , optimized via grid search to balance effectiveness and simplicity.

Hierarchical and Parent–Child Retrieval. Chunk granularity in RAG presents a recurring trade-off between retrieval precision and contextual coherence available to the generator. HRR (Singh and Mohapatra, 2025) proposes a learned

hierarchical reranking framework over document trees, while HiRAG (Huang et al., 2025) introduces graph-structured knowledge integration to aggregate multi-granular evidence. Both methods add supervised or structure-aware components on top of the base retriever. H-RAG follows the same intuition of combining fine-grained retrieval with higher-level context reconstruction, but adopts a lightweight unsupervised design. Sentence-level child chunks are retrieved and rescored using embedding similarity, then aggregated to parent documents via maximum-score assignment. Rather than proposing a new retrieval algorithm, our contribution is to study how hierarchical retrieval behaves under the MTRAGEval conversational benchmark and how retrieval-side configuration choices affect grounding quality.

Conversational Query Rewriting and Evaluation. Conversational retrieval often benefits from query rewriting to resolve context dependency. Multi-query rewriting (Kostric and Balog, 2024) and adaptive rewriting methods such as AdaQR (Zhang et al., 2024) improve recall by reformulating dialogue queries. H-RAG incorporates a lightweight rewriting step without introducing additional learned components. Evaluation of RAG systems has evolved beyond traditional IR metrics. eRAG (Salemi and Zamani, 2024) evaluates retrieval quality based on downstream utility, while RAGChecker (Ru et al., 2024) provides fine-grained diagnostic analysis of retriever and generator contributions. The MTRAG benchmark (Katsis et al., 2025) integrates these considerations into a multi-dimensional evaluation framework, under which we analyze both retrieval effectiveness (Task A) and grounded generation quality (Task C).

3 System Overview

Our system, **H-RAG**, implements an end-to-end Retrieval-Augmented Generation (RAG) pipeline for Task C of MTRAGEval, where systems retrieve relevant documents and generate grounded responses in multi-turn conversations. The architecture balances retrieval precision, contextual coherence, and computational efficiency while remaining modular and scalable.

The pipeline consists of two coupled phases: *document ingestion* and *query-time retrieval and generation*, as illustrated in Figure 1.

3.1 Document Ingestion

During ingestion, documents are processed using a hierarchical parent-child chunking strategy. Each document is first segmented into sentences using a rule-based splitter. Overlapping *child chunks* are then constructed with a sliding window over consecutive sentences. In our implementation, each child chunk contains three sentences with a stride of two, preserving local coherence while remaining compact for retrieval. Each child chunk is embedded using a sentence embedding model and indexed in a vector database for efficient semantic search. In parallel, the original full document is preserved as a *parent document* in a document store, with metadata linking each child chunk to its parent. This representation decouples retrieval granularity from generation context size: retrieval uses fine-grained child chunks, while generation benefits from richer parent-level context.

3.2 Query-Time Retrieval

At query time, incoming queries are handled by an API orchestrator that retrieves recent dialogue history from a session store to support multi-turn interactions. When queries depend on prior context or contain implicit references, a lightweight LLM-based rewriting module reformulates them into standalone retrieval queries. The rewritten query is embedded and used for hybrid dense-sparse retrieval over indexed child chunks. Dense semantic similarity is combined with keyword-based lexical matching through a tunable weighting parameter, balancing semantic recall and exact-match precision. An initial set of top- k child chunks is retrieved, where k is configurable. The rewrite prompt instructs the model to return the question unchanged when it is already standalone, which is essential for short factoid turns where LLM-generated rewrites may over-elaborate. Retrieved child chunks are refined using embedding-based similarity rescoring, where cosine similarity between the query and child-chunk embeddings improves ranking precision. Unlike cross-encoder reranking, this step relies only on embedding similarity, enabling efficient large-scale retrieval while maintaining ranking quality.

3.3 Parent Aggregation and Context Reconstruction

Following rescoring, child chunks are mapped back to their corresponding parent documents. Parent-

level relevance scores are computed by aggregating the scores of associated child chunks, using the maximum child score as the parent score. The system supports two ranking strategies: (1) ranking child chunks first and then selecting parent documents based on aggregated scores, or (2) identifying candidate parent documents and performing an additional embedding-based rescoring step directly on parent document content. These strategies are treated as configurable options during experimentation. The top- n ranked parent documents are retrieved from the document store and used as context for generation. Supplying parent-level documents ensures that the generation module has access to sufficient surrounding information while avoiding excessive input length and redundancy.

3.4 Answer Generation

Finally, an instruction-tuned large language model generates a response conditioned on the reconstructed parent-level context and the rewritten query. The generation module is designed to either produce a grounded answer supported by the retrieved documents or abstain when the available evidence is insufficient, in accordance with Task C requirements. Overall, H-RAG integrates hierarchical document representation, hybrid retrieval, embedding-based rescoring, and parent-level context reconstruction into a unified pipeline for multi-turn RAG. This design aligns closely with the MTRAGEval evaluation framework, which assesses both answer quality and document-level evidence grounding.

3.5 Implementation Details

Documents are segmented into overlapping three-sentence child chunks with a stride of two sentences using a rule-based sentence splitter. Child chunks and parent documents are indexed in a Weaviate hybrid vector store with $\alpha = 0.7$ (70% dense, 30% sparse). Child embeddings are produced using BAAI/bge-large-en-v1.5, while rescoring uses BAAI/bge-reranker-v2-m3 through cosine similarity over embeddings.

The initial candidate set size is $k = 50$ child chunks. Parent scores are computed using maximum-score aggregation over retrieved child chunks, and the top- $n = 5$ parent documents are passed to the generator. Both query rewriting and answer generation use OpenAI gpt-5. Query rewriting uses temperature 0.2 to minimize unnecessary reformulation, while answer generation uses

temperature 0.7 with `max_completion_tokens = 4096`. No component is fine-tuned; all models are used zero-shot. The generation prompt instructs the model to respond conversationally while avoiding explicit references to “the context” or numeric citations. The full prompt templates are provided in Appendix A. Additional configuration ablation results are provided in Appendix B. Table 1 summarizes the implementation settings used for the submitted H-RAG system, including embedding model, hybrid retrieval configuration, chunking strategy, ranking method, and LLM parameters.

4 Experimental Setup

4.1 Dataset

We evaluate our system on the MTRAG dataset released as part of SemEval-2026 Task 8: MTRAGEval. The dataset consists of human-curated multi-turn conversational interactions grounded in a large document collection. Each instance contains a sequence of dialogue turns followed by a final user query, which may depend on earlier conversational context through coreference, ellipsis, or implicit constraints.

In **Task C: Generation with Retrieved Passages**, systems are required to retrieve a fixed number of relevant documents from the corpus and generate a grounded response conditioned on the retrieved evidence. The dataset includes both answerable and unanswerable queries, requiring systems to either produce a faithful answer or abstain when insufficient evidence is available. We use the official dataset splits and evaluation protocol provided by the task organizers.

4.2 Retrieval Configuration and Hyperparameters

Our experiments focus on analyzing the impact of key retrieval and ranking components within the H-RAG pipeline. Rather than introducing additional architectural components, we perform a systematic exploration of retrieval-related hyperparameters that directly influence evidence selection and grounding quality. We explore the following configuration dimensions:

Embedding Model. We evaluate two sentence embedding models for indexing and retrieval: `intfloat/multilingual-e5-large` and `bge-large-en-v1.5`. These models are used consistently for child chunk embedding and similarity computation during retrieval and rescoring.

Hybrid Retrieval Weight. Hybrid dense–sparse retrieval combines semantic similarity search with keyword-based lexical matching. We tune the hybrid weighting parameter α , which controls the relative contribution of dense embeddings versus sparse keyword matching. We consider values $\alpha \in \{0.9, 0.7, 0.5\}$.

Ranking Strategy. We evaluate two parent document ranking strategies. In the first strategy, child chunks are ranked and rescored first, after which unique parent documents are selected based on aggregated child scores. In the second strategy, unique parent documents are identified earlier in the pipeline and subjected to an additional embedding-based rescoring step using parent-level content. This setting allows us to assess the effect of ranking order on evidence selection.

Initial Candidate Set Size. We vary the number of retrieved child chunks prior to parent aggregation. Specifically, we consider candidate set sizes of $k \in \{20, 30, 50\}$, enabling analysis of the trade-off between retrieval recall and ranking noise. Unless otherwise specified, a single configuration is designated as the base setup for controlled comparisons, and remaining parameters are held constant while individual dimensions are varied.

4.3 Evaluation Protocol

We evaluate all configurations using the official MTRAGEval evaluation framework. The framework combines reference-based and reference-less metrics designed specifically for RAG systems. Evaluation assesses answer quality, document-level evidence grounding, and appropriate abstention behavior for unanswerable queries. All system outputs follow the official Task C submission schema, which requires generated responses to be accompanied by a ranked list of retrieved documents. Scores are computed automatically using the task-provided scripts, and all evaluations are conducted under identical conditions.

5 Results and Analysis

We report results using the official MTRAGEval evaluation framework. Task A evaluates standalone retrieval quality using nDCG@5, while Task C evaluates end-to-end RAG using the harmonic mean of document relevance (RB_agg), faithfulness (RL_F), and LLM-based answer quality (RB_llm). The reported scores correspond to the configuration using BAAI/bge-large-en-v1.5, $\alpha = 0.7$,

Component	Name
Embedding model	BAAI/bge-large-en-v1.5
Reranker	BAAI/bge-reranker-v2-m3
Generation LLM	GPT-5 (T=0.7)
Query rewrite LLM	GPT-5 (T=0.2)
Hybrid weight α	0.7
Initial candidates k	50
Top- n parents	5
Chunking	3 sentences, stride 2
Aggregation	Max child score
Ranking strategy	Child-first
Vector store	Weaviate hybrid

Table 1: Implementation settings used for the submitted H-RAG system.

ranking by child-first aggregation, and an initial candidate size of $k = 50$.

Table 2 summarizes the configuration sweep across $\alpha \in \{0.5, 0.7, 0.9\}$, $k \in \{20, 30, 50\}$, and both ranking strategies on the development set. Ranking strategy emerges as the dominant factor: enabling parent-level rescoring (rank_parents = true) consistently improves retrieval metrics across all configurations, with mean gains of +0.0197 nDCG@5 and +0.0108 Recall@5. In contrast, hybrid weighting has limited sensitivity, with less than 0.003 variation in mean nDCG@5 across α values. Candidate size peaks at $k = 30$, suggesting that larger retrieval pools may introduce additional ranking noise. The strongest configuration is achieved with $\alpha = 0.7$, parent-level ranking enabled, and $k = 30$, yielding nDCG@5 = 0.4872 and Recall@5 = 0.5184. Additional configuration results across hybrid weighting, ranking strategy, and candidate set size are reported in Appendix B.

5.1 Task A: Retrieval Performance

On Task A, H-RAG achieves an nDCG@5 score of 0.4271 on the final submission dataset. It further achieves an nDCG@5 score of 0.4728 and Recall@5 of 0.502 on the baseline dataset (Rosenthal et al., 2026b), while the baseline model reports an nDCG@5 score of 0.45 and Recall@5 of 0.49. These results reflect the effectiveness of hybrid dense–sparse retrieval combined with hierarchical child-level indexing. Sentence-level child chunks improve fine-grained matching, while hybrid fusion balances lexical precision and semantic similarity. Overall, retrieval design choices strongly influence performance in multi-turn conversational settings. After comparing document-level and passage-level datasets (Rosenthal et al., 2026a), we observed an approximately 10% improvement

when evaluating on document-level data. This gain suggests that the proposed system is particularly well suited for long-reference-document scenarios, aligning with its design objective of supporting short-query, long-context retrieval settings. Compared with baseline results reported by the task organizers (Rosenthal et al., 2026a,b), H-RAG achieves slight improvements, demonstrating its effectiveness relative to established approaches.

5.2 Task C: End-to-End RAG Performance

For Task C, H-RAG achieves a harmonic mean score of 0.3241, with component scores of $RB_agg = 0.2488$, $RL_F = 0.2703$, and $RB_llm = 0.6508$. The relatively high RB_llm score suggests that the generation module produces fluent and semantically appropriate responses. However, lower RB_agg and RL_F scores indicate that evidence grounding remains the primary bottleneck. The gap between RB_llm and grounding-related metrics suggests that retrieval quality alone does not guarantee faithful evidence usage during generation. While retrieved parent documents often contained relevant supporting information, the generation model occasionally produced fluent responses that were only partially aligned with retrieved evidence. This indicates that evidence selection and grounding remain more challenging than answer fluency in conversational RAG.

6 Conclusion

We presented H-RAG, a hierarchical parent-child retrieval system developed for SemEval-2026 Task 8 (MTRAGEval), participating in both Task A and Task C. The proposed architecture decouples fine-grained child-level retrieval from parent-level context reconstruction, integrating hybrid dense-sparse search, embedding-based rescoring, and lightweight query rewriting within a unified pipeline. Our results demonstrate that, in multi-turn conversational RAG, fluent answer generation alone is insufficient for strong end-to-end performance; robust evidence selection and aggregation remain dominant limiting factors. The observed gap between answer quality and grounding metrics highlights the role of retrieval design in conversational settings. Future work may explore adaptive hybrid weighting, refined parent-level aggregation, uncertainty-aware retrieval mechanisms, context-aware segmentation, and conversational-history summarization to strengthen grounded generation

in long-context dialogue systems.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): [10.55776/COE12](#) and [10.55776/DFH23](#). We would like to thank the task organizers for the opportunity to participate in this challenge. We used generative AI tools to assist with language polishing and clarity of presentation. All scientific content, experimental design, analysis, and conclusions were developed and verified by the authors.

References

- Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. 2025. [CORAL: Benchmarking multi-turn conversational retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1308–1330, Albuquerque, New Mexico. Association for Computational Linguistics.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, and Fabrizio Silvestri. 2024. [The power of noise: Redefining retrieval for rag systems](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 719–729, New York, NY, USA. Association for Computing Machinery.
- Hsin-Ling Hsu and Jengnan Tzeng. 2025. [Dat: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation](#). *arXiv preprint arXiv:2503.23013*.
- Haoyu Huang, Yongfeng Huang, Yang Junjie, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. [Retrieval-augmented generation with hierarchical knowledge](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6044–6060, Suzhou, China. Association for Computational Linguistics.
- Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. [Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems](#). *Transactions of the Association for Computational Linguistics*, 13:784–808.
- Ivica Kostic and Krisztian Balog. 2024. [A surprisingly simple yet effective multi-query rewriting method for conversational passage retrieval](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, pages 2271–2275. ACM.

Tzu-Lin Kuo, FengTing Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-shan Shiu. 2025. [RAD-bench: Evaluating large language models’ capabilities in retrieval augmented dialogues](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 868–902, Albuquerque, New Mexico. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.

Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026a. Mtragun: A benchmark for open challenges in multi-turn rag conversations. *arXiv preprint arXiv:2602.23184*.

Sara Rosenthal, Vraj Shah, Yannis Katsis, and Marina Danilevsky. 2026b. Semeval-2026 task 8: Mtrageval: Evaluating multi-turn rag conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California. Association for Computational Linguistics.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: a fine-grained framework for diagnosing retrieval-augmented generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.

Alireza Salemi and Hamed Zamani. 2024. [Evaluating retrieval quality in retrieval-augmented generation](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24*, pages 2395–2400. ACM.

Ashish Singh and Priti Mohapatra. 2025. Hierarchical re-ranker retriever (hrr). *arXiv preprint arXiv:2503.02401*.

Tianhua Zhang, Kun Li, Hongyin Luo, Xixin Wu, James R. Glass, and Helen M. Meng. 2024. [Adaptive query rewriting: Aligning rewriters through marginal probability of conversational answers](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13444–13461, Miami, Florida, USA. Association for Computational Linguistics.

A Prompt Templates

Query Rewrite Prompt

Task: Given the conversation history, rewrite the new user question into a standalone and specific query suitable for retrieval.

Important Rules:

- If the question is already clear and standalone, return it **EXACTLY** as is
- If the question contains pronouns or references to earlier dialogue, rewrite it using the necessary context
- Do **NOT** invent information or change the original meaning
- Return only the rewritten question with **NO explanation**

Conversation History:

{history_text}
(last three Q/A turns, formatted as “Q: ...” and “A: ...”)

New Question: {user_question}

Expected Output:

Rewritten standalone question (or unchanged original question if already standalone).

Generation Prompt

Task: Generate a natural conversational response grounded in retrieved parent-level documents while maintaining dialogue continuity. **System Instruction:**

You are a helpful AI assistant engaged in a conversation with a user. Answer the user’s question naturally and directly, as if you already know the information.

Important Rules:

- Do **NOT** mention “the context”, “the provided information”, “according to the documents”, or similar phrases
- Do **NOT** reference source numbers such as [1], [2], etc.
- Respond conversationally as if the knowledge is your own
- If insufficient information is available, state this naturally without mentioning missing context
- Maintain continuity with the conversation history

Conversation History:

[Conversation history block, when present]

Retrieved Background Knowledge:

[Top-*n* parent documents with numbered references]

(used internally for grounding; never referenced explicitly)

Current User Question: {user_question}

Generation Trigger:

Respond naturally and conversationally:

α	rank_parents	k	nDCG@1	nDCG@3	nDCG@5	Recall@1	Recall@3	Recall@5
0.5	true	20	0.4938	0.4530	0.4832	0.2111	0.4243	0.5120
0.5	true	30	0.4952	0.4511	0.4826	0.2120	0.4204	0.5108
0.5	true	50	0.4931	0.4449	0.4765	0.2093	0.4105	0.5046
0.7	true	20	0.4913	0.4513	0.4822	0.2092	0.4236	0.5125
0.7	true	30	0.4952	0.4559	0.4872	0.2111	0.4288	0.5184
0.7	true	50	0.4938	0.4506	0.4806	0.2110	0.4204	0.5084
0.9	true	20	0.4913	0.4515	0.4828	0.2095	0.4227	0.5125
0.9	true	30	0.4938	0.4562	0.4856	0.2104	0.4296	0.5145
0.9	true	50	0.4926	0.4537	0.4852	0.2113	0.4258	0.5164
0.5	false	20	0.4425	0.4248	0.4618	0.1910	0.4068	0.5022
0.5	false	30	0.4373	0.4245	0.4622	0.1884	0.4076	0.5048
0.5	false	50	0.4464	0.4285	0.4648	0.1931	0.4105	0.5046
0.7	false	20	0.4514	0.4310	0.4653	0.1944	0.4115	0.5029
0.7	false	30	0.4438	0.4291	0.4657	0.1916	0.4115	0.5064
0.7	false	50	0.4476	0.4316	0.4662	0.1934	0.4139	0.5050
0.9	false	20	0.4360	0.4153	0.4433	0.1878	0.3967	0.4746
0.9	false	30	0.4502	0.4305	0.4673	0.1942	0.4117	0.5069
0.9	false	50	0.4361	0.4250	0.4624	0.1880	0.4095	0.5058

Table 2: Configuration ablation across hybrid weighting α , ranking strategy, and initial candidate size k . Best-performing configuration is highlighted in bold.

B Configuration Ablation

Table 2 reports the full configuration sweep across hybrid weighting α , ranking strategy, and candidate set size k . These experiments were conducted to analyze retrieval sensitivity under the MTRAGEval framework.