AKCIT-UFG at SemEval-2026 Task 8: Structured Chunking and Optimized Query Reformulation for Efficient Multi-Turn Retrieval

David O. C. Ferreira, Priscila R. M. F. Ribeiro, Emanuel B. Passinato, Diogo F. C. Silva & Arlindo R. Galvão Filho
Advanced Knowledge Center for Immersive Technologies (AKCIT)
Federal University of Goiás
{oneil, priscila.maia, diogo_fernandes}@discente.ufg.br
emanuel.passinato@egresso.ufg.br arlindogalvao@ufg.br

Abstract

Multi-turn Retrieval-Augmented Generation (RAG) poses additional challenges due to conversational query underspecification and contextual dependencies. We conduct a systematic empirical study of multi-turn retrieval design choices across multiple domains, analyzing passage granularity, conversational query rewriting, and compact dense retrievers under a unified benchmark setting. Our experiments show that smaller passage segmentation improves early-rank effectiveness, while lightweight query rewriting substantially enhances dense retrieval performance. We further find that supervised fine-tuning with hard negative mining plays a critical role in enabling dense retrievers to outperform lexical baselines. Overall, the results indicate that effective multi-turn retrieval depends not only on model scale but also on principled architectural and preprocessing decisions.

1 Introduction

Retrieval-Augmented Generation (RAG) has become a central paradigm in natural language processing (NLP), integrating document retrieval with neural text generation to improve factual grounding and contextual relevance. By conditioning generation on retrieved evidence, RAG systems mitigate hallucinations and provide greater interpretability compared to purely parametric models.

While most prior work evaluates RAG under single-turn settings, real-world systems increasingly operate in multi-turn conversations. In this setting, retrieval must account for dialog history and contextual dependencies across turns. Multi-turn Retrieval-Augmented Generation (MT-RAG) introduces additional challenges: conversational queries are often underspecified relative to standalone search queries, and retrieval performance becomes sensitive to how context is encoded and segmented.

Designing effective multi-turn retrieval pipelines therefore requires addressing three interdependent factors: (i) conversational query reformulation, (ii) passage granularity, and (iii) retrieval model capacity. Despite growing interest in MT-RAG evaluation frameworks, the interaction between these components, particularly under computationally constrained settings, remains underexplored.

In this work, we conduct a systematic empirical study of multi-turn retrieval design choices using the SemEval MT-RAG benchmark (**?**). We analyze how much retrieval effectiveness can be gained through structural design choices, passage granularity and lightweight conversational query rewriting, when using compact, locally deployable retrievers.

Our experiments show that smaller passage segmentation improves early-rank effectiveness across domains, and that targeted rewriting can substantially amplify dense retrieval performance, with gains depending on the encoder backbone. Overall, the results indicate that robust MT-RAG retrieval can be achieved without large proprietary models, highlighting the role of principled preprocessing and retrieval architecture alongside model capacity.

2 Related Work

Information retrieval has traditionally relied on probabilistic lexical matching models, among which BM25 (Robertson and Zaragoza, 2009) remains a strong and widely adopted baseline. BM25 estimates relevance based on term frequency and inverse document frequency weighting, which makes it particularly effective when there is explicit lexical overlap between queries and relevant passages. Despite advances in neural methods, lexical retrieval continues to demonstrate competitive performance, especially in settings where queries share surface-level terms with relevant documents.

Conversational question answering introduced

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3149–3155

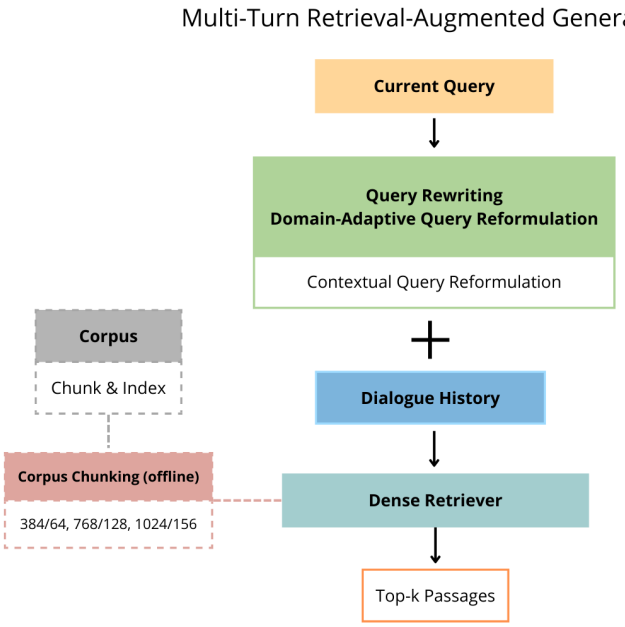

Figure 1: Overview of the proposed Multi-Turn RAG pipeline. The system performs Contextual Query Reformulation on the current query, which is then combined with Dialogue History to inform the Dense Retriever. The retrieval process operates over an offline-indexed corpus with multiple chunking configurations (384, 768, and 1024 tokens).

additional challenges by requiring systems to interpret context-dependent queries. QuAC (Choi et al., 2018) highlighted the need to model dialogue history when handling non-standalone user turns. With the rise of pretrained transformers, cross-encoder reranking became a standard refinement step, as demonstrated by (Nogueira and Cho, 2019). Cross-encoders jointly encode the query and passage, allowing fine-grained token-level interaction between them, which often leads to more accurate relevance estimation compared to independent encoding approaches. This interaction-based modeling is particularly relevant in multi-turn settings where subtle contextual cues influence passage ranking.

Dense retrieval further advanced the field through dual-encoder architectures such as DPR (Karpukhin et al., 2020), which map queries and passages into a shared embedding space. RAG (Lewis et al., 2020) integrated neural retrieval with sequence-to-sequence generation, establishing a paradigm for knowledge-intensive tasks. However, BEIR (Thakur et al., 2021) showed that dense retrievers often require domain adaptation to outperform lexical baselines, particularly under domain shift and zero-shot settings. In conversational retrieval, QReCC (Anantha et al., 2021) enabled systematic study of query reformulation for resolving context-dependent queries.

Our work builds upon these foundations by analyzing lexical and dense retrieval and query rewriting within a unified multi-turn RAG setting.

3 Methodology

Our approach consists of a multi-stage retrieval pipeline designed for multi-turn RAG settings. The design decisions are grounded in the characteristics of the target benchmark and aim to address conversational underspecification and passage granularity effects in retrieval.

3.1 Dataset Analysis and Domain Adaptation

The MTRAG-UN benchmark defines four domains that capture open challenges in multi-turn RAG conversations (Katsis et al., 2025): ClapNQ, Cloud, FiQA, and Govt. These domains vary substantially in terminology, discourse structure, and information density, ranging from financial language (FiQA) to technical documentation (Cloud) and public policy texts (Govt).

Inspection of the benchmark collections and conversational queries shows that user turns are frequently underspecified and rely on implicit entity references or prior dialogue context. Such char-

acteristics create a mismatch between conversational inputs and document-level retrieval units. These domain-specific properties motivate the use of rewriting strategies tailored to conversational resolution rather than a single uniform retrieval configuration.

3.2 Domain-Specific Query Rewriting

To mitigate conversational underspecification, we employ a lightweight rewriting model (`Qwen-0.6B`) that converts context-dependent turns into standalone queries. The rewriting prompts explicitly instruct the model to resolve anaphora, recover implicit entities, and emphasize domain-relevant terminology.

Rather than relying on large proprietary systems, the rewriting stage is guided through carefully designed prompt templates that induce targeted transformations appropriate for each domain. The resulting queries are concise and retrieval-oriented, improving alignment with both lexical and dense retrievers. The full prompt templates are provided in Appendix A.

3.3 Chunking Strategy and Retrieval Models

Passage segmentation plays a central role in retrieval performance, as chunk granularity determines the trade-off between contextual completeness and retrieval noise. We evaluate multiple chunk sizes to analyze how segmentation affects downstream retrieval.

To isolate structural effects, chunk size experiments are conducted using BM25 as a controlled lexical baseline. This setup enables a direct assessment of passage-level design decisions independent of semantic encoding capacity.

After selecting an effective chunk configuration, compact dense retrievers are evaluated on the same segmented collections. This controlled separation makes it possible to distinguish improvements attributable to passage structure from those attributable to semantic modeling, and to examine how dense encoders leverage rewritten queries under optimized chunk boundaries.

4 Experiments

This section evaluates how far retrieval quality can be improved through principled architectural and preprocessing decisions within strict resource limits. Our experimental design is guided by two complementary objectives: (i) improving multi-turn retrieval effectiveness, and (ii) maintaining computational efficiency under single-GPU, locally deployable constraints. Rather than relying on large proprietary models or heavy reranking pipelines, we prioritize compact backbones and structural optimizations—specifically passage granularity and conversational query rewriting.

4.1 Setup

All experiments are conducted on the four Task A domains: ClapNQ, Cloud, FiQA, and Govt. Each conversational turn is treated as a retrieval instance: given a query turn, the system retrieves the top-k passages from the corresponding domain collection. We report standard ranking metrics, including Recall@k and nDCG@k (official) and MRR (for early-rank sensitivity). Unless stated otherwise, we evaluate `lastturn` queries, i.e., the final user turn of each conversation.

Lexical retrieval and chunking. BM25 serves as the lexical baseline for the chunking study, allowing us to isolate passage segmentation effects independent of dense semantic encoding. We evaluate three passage configurations: 384 tokens with 64 overlap, 768 tokens with 128 overlap, and 1024 tokens with 156 overlap. After selecting the best-performing configuration, we fix chunking to that default for subsequent experiments unless explicitly stated.

Dense retrievers. We benchmark compact dense encoders including `BGE-Base`, `BGE-M3`, `multilingual-e5-large`, and a `Qwen`-based dense encoder, following the same collections and retrieval protocol used for BM25. All dense models are evaluated on the same indexed corpora constructed under each chunking configuration.

Query rewriting. Conversational query rewriting is performed with `Qwen-0.6B-Instruct` to transform context-dependent turns into standalone queries. We compare *Expand*, *Clarify*, *Decompose*, *GQR*, *KWR*, *PAR*, and *CCE*. Rewritten queries are used as input to both BM25 and dense retrieval to quantify the impact of reformulation on different retrievers.

Computational environment. All experiments were executed on a single NVIDIA RTX 4090 (24GB).

	MRR			nDCG@5			Recall@5		
Dataset	**384**	**768**	**1024**	**384**	**768**	**1024**	**384**	**768**	**1024**
ClapNQ	**0.69**	0.67	0.65	**0.70**	0.69	0.69	**0.66**	0.65	**0.66**
Cloud	0.23	**0.24**	0.23	0.20	**0.23**	0.22	0.19	**0.22**	0.21
FiQA	0.27	**0.28**	**0.28**	**0.21**	0.20	**0.21**	**0.20**	0.19	0.19
Govt	**0.29**	0.27	0.25	**0.32**	0.31	0.30	**0.29**	0.28	0.27

Table 1: Impact of document chunk size on ranking performance. **Bold** indicates the best value per dataset and metric (ties included).

	nDCG@1			Recall@1		
Dataset	**384**	**768**	**1024**	**384**	**768**	**1024**
ClapNQ	**0.67**	0.63	0.60	**0.67**	0.63	0.60
Cloud	0.15	0.16	**0.17**	0.15	0.16	**0.17**
FiQA	0.20	**0.22**	**0.22**	0.14	**0.15**	**0.15**
Govt	**0.23**	0.21	0.19	**0.23**	0.21	0.19

Table 2: Performance sensitivity to chunk size across various datasets. **Bold** indicates the best value per dataset (ties included).

4.2 Chunks

We evaluate chunk granularity with BM25 under three passage configurations: 384/64, 768/128, and 1024/156 (chunk size / overlap). Results are reported with the official competition metrics (nDCG@k and Recall@k) and complemented with MRR to capture early-rank sensitivity. All values in Tables 1 and 2 are computed on `lastturn` queries. Best values per dataset row are shown in bold.

Across the three configurations, 384/64 achieves the strongest overall early-rank performance, yielding the best MRR and nDCG@5 on *ClapNQ* and *Govt*, while remaining competitive on *Cloud* and *FiQA* (Tables 1 and 2). Based on these results, we use 384/64 as the default chunking configuration in subsequent experiments.

Smaller passages appear to reduce topical drift in conversational retrieval, where turns are often short and context-dependent. In this setting, finer segmentation increases the likelihood that highly ranked passages directly match the localized information need.

4.3 Query Rewriting

This section evaluates conversational query rewriting to mitigate context dependency in multi-turn turns (e.g., ellipsis and coreference) by transforming the last user turn into a standalone query prior to retrieval. The rewriting stage is inspired by recent work on query rewriting for multi-turn RAG, particularly DMQR (Li et al., 2024). Rewrites are generated with `Qwen-0.6B-Instruct` and evaluated on the four domains (ClapNQ, Cloud, FiQA, Govt) under the retrieval protocol described in Section 4.1. Unless stated otherwise, these experiments use the default chunking configuration 384/64 selected above.

Dataset	**Expand**	**Clarify**	**Decompose**	**GQR**	**KWR**	**PAR**	**CCE**
ClapNQ	0.4041	0.3977	0.3739	0.3957	**0.4610**	0.3658	0.4045
Cloud	0.5223	0.5580	0.5270	0.5399	**0.5729**	0.4912	0.5444
FiQA	0.3690	0.3752	0.3653	0.3709	**0.4003**	0.3572	0.3724
Govt	0.4711	0.4718	0.4596	0.4679	**0.5104**	0.4482	0.4730

Table 3: Average retrieval score by query rewrite technique using the Qwen-based retriever across domains. Best value per dataset is shown in **bold**.

Rewrite techniques. The following strategies are considered:

- **Expand**: enriches the last-turn query by explicitly adding missing context from the dialogue history.
- **Clarify**: rewrites the query to resolve ambiguous references (e.g., pronouns, underspecified entities) while preserving intent.
- **Decompose**: splits complex information needs into simpler sub-questions and rewrites the query accordingly.
- **GQR (General Query Rewrite)**: produces a cleaned, well-formed standalone query that preserves the original meaning while removing noise.
- **KWR (Keyword Rewrite)**: expresses the information need as a compact keyword-style query (entities + key terms), aiming to maximize retrievability.
- **PAR (Pseudo-Answer Rewriting)**: generates a short pseudo-answer (or answer-like hypothesis) and uses it to form a retrieval-oriented query.
- **CCE (Core Context Extraction)**: extracts only the minimal core constraints/slots required to answer the turn, discarding peripheral details.

Technique selection based on retrieval metrics. Technique selection is based on retrieval effectiveness on development data. Table 3 reports nDCG@10 on dev for each rewrite strategy in each

Metric	Chunk 384	Chunk 768	Chunk 1024	Overall Avg.
Recall@1	**+14.96**	+13.13	+11.06	+13.05
Recall@10	+10.81	+10.57	**+12.91**	+11.43
nDCG@1	+14.49	**+15.07**	+13.13	+14.23
nDCG@10	+18.61	+16.69	**+18.84**	+18.05

Table 4: Relative metric change (%) after query rewriting for `Qwen 0.6B`, macro-averaged across datasets.

Metric	Chunk 384	Chunk 768	Chunk 1024	Overall Avg.
Recall@1	-9.30	-8.66	-9.47	-9.14
Recall@10	-5.81	-6.27	-6.29	-6.12
nDCG@1	-8.14	-8.44	-7.87	-8.15
nDCG@10	-6.61	-6.29	-6.81	-6.57

Table 5: Relative metric change (%) after query rewriting for `BGE-M3`, macro-averaged across datasets.

domain (higher is better), computed using the same dense retrieval backbone across techniques. Across the four datasets, KWR yields the strongest and most consistent gains, and is therefore adopted as the default rewriting strategy in subsequent experiments.

Qualitative prompt examples. For qualitative inspection and reproducibility, prompt templates (one example per technique) are provided in Appendix A.

4.4 Dense Models

This section analyzes how compact dense retrievers respond to query refinement under a small, local rewriting model. The central hypothesis is that encoder backbones differ in how strongly they benefit from improved query specificity. In particular, a `Qwen`-based dense encoder exhibits higher *semantic elasticity*, improving substantially under rewriting, whereas `BGE-M3` behaves more *rigidly* and may degrade under the same refinement.

Rewriter vs. retriever roles. We use `Qwen-0.6B-Instruct` as a lightweight *rewriter*. Dense retrieval is performed by independent dual-encoder backbones (including `BGE-M3` and a `Qwen`-based dense encoder). In all dense experiments below, we apply the default rewrite strategy KWR selected in Section 4.3. Additional encoders were also evaluated (e.g., `E5` and `bge-base`); however, these backbones achieved consistently lower performance than the two contrasted models and are omitted for brevity.

Experimental view. To isolate the effect of query refinement, we report relative percentage change (in %) in retrieval metrics after applying query rewriting. Improvements are computed for three chunk sizes (384, 768, 1024) and then macro-averaged across datasets. All improvements are measured with respect to the baseline using the original (non-rewritten) `lastturn` queries. The reported metrics include official competition metrics (Recall@k and nDCG@k).

Results. Tables 4 and 5 show a clear contrast: the `Qwen`-based dense encoder consistently benefits from rewriting across chunk sizes, with strong gains in early-rank quality (Recall@1 and nDCG@1) and substantial improvements at larger cutoffs (nDCG@10). In contrast, `BGE-M3` degrades across all reported metrics after rewriting, suggesting that the same refinement introduces distribution shifts that the model does not translate into improved retrieval.

4.5 Local Evaluation vs. Official Leaderboard

The absence of results and the official leaderboard for Task 8 of SemEval-2026 is due to the fact that the final optimized pipeline (segmentation into 384/64 blocks) could not be officially evaluated. Central hypothesis of maximizing performance under strict computational constraints, a last-minute switch to this more refined segmentation required a complete reconstruction of the index, preventing the final submission.

5 Conclusion

We presented the AKCIT-UFG submission to SemEval-2026 Task 8 (MTRAGEval), Track A, focusing on efficient multi-turn retrieval under constrained computational settings. Our results show that retrieval quality in conversational settings depends not only on encoder scale, but on principled structural design choices.

Across domains, finer passage segmentation and lightweight keyword-oriented query rewriting consistently improved early-rank effectiveness. Moreover, dense retrievers responded differently to query refinement, highlighting that rewriting interacts with encoder inductive biases rather than universally improving performance.

Overall, our findings indicate that competitive multi-turn retrieval can be achieved with compact, locally deployable components when chunking, rewriting, and model selection are treated as first-class design decisions. Future work will investigate adaptive rewrite policies and hybrid lexical–dense pipelines under strict efficiency constraints.

Acknowledgements

This work has been fully funded by the project Research and Development of Gênese Digital: Scaling Interactive and Culturally Adapted Digital Humans supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT of the MCTI grant number 057/2023, signed with EMBRAPII. The authors are also grateful to the Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) for the financial support provided for this research (Grant 64448878/2024).

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. 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.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Zhicong Li, Jiahao Wang, Zhishu Jiang, Hangyu Mao, Zhongxia Chen, Jiazhen Du, Yuanxing Zhang, Fuzheng Zhang, Di Zhang, and Yong Liu. 2024. Dmqr-rag: Diverse multi-query rewriting for rag. *arXiv preprint arXiv:2411.13154*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

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

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

A Query Rewriting Prompt Templates

This appendix provides the prompts for the rewriting technique used in Section 4.3. The examples are intended for qualitative inspection and reproducibility.

A.1 Expand

```
[Rewrite this search query to be more comprehensive. Add relevant synonyms, related terms, and variations while keeping it natural. Query: query Rewritten query (one sentence only):"""]
```

A.2 Clarify

```
[Rewrite this search query to be clearer and more specific. Add necessary context and clarify ambiguous terms. Query: query Rewritten query (one sentence only)]
```

A.3 Decompose

```
[Rewrite this complex query as a single comprehensive question. Combine all aspects into one detailed query. Query: query Rewritten query (one sentence only)]
```

A.4 GQR (General Query Rewrite)

```
[Clean and standardize this search query. Remove filler words, fix grammar, use proper terminology. Keep it concise. Query: query Clean query (short):]
```

A.5 KWR (Keyword Rewrite)

```
[Extract only the essential keywords from this query. Return only important terms, separated by spaces. No explanations. Query: query Keywords:"""]
```

A.6 PAR (Pseudo-Answer Rewriting)

```
[Expand this query by imagining what a relevant answer would contain. Include key terms and concepts that would appear in good results. Query: query Expanded query (1-2 sentences)]
```

A.7 CCE (Core Context Extraction)

```
[Extract ONLY the core essential content from this query. Remove all redundancy. Return the absolute minimum needed. Be extremely brief.  Query:  query Core content:"""]
```