Citation-Aware RAG Framework Leveraging Hierarchical Fusion Representation for 3GPP Specifications

Tien-Quang Nguyen, Quang-Hai Vu, Minh-Hanh Tran,
Phi-Hung Nguyen, Quan-Huan Vu
Viettel High Technology – Viettel Group
Hanoi, Vietnam
{quangnt22, haivq1, hanhtm13, hungnp22, huanvq1}@viettel.com.vn

Thai Trong Nguyen
Viettel Networks – Viettel Group
Hanoi, Vietnam
thaint@viettel.com.vn

***Abstract*—Third Generation Partnership Project (3GPP) specifications are fundamental to the development and standardization of modern telecommunication systems. However, answering questions from 3GPP specifications is challenging, as information is distributed across a large collection of documents and interconnected through extensive citation relationships. While recent retrieval-augmented generation (RAG) methods have improved access to 3GPP information, they primarily rely on semantic similarity at the chunk level, overlooking the hierarchical organization of information across series and documents, as well as the citation relationships among specifications. These limitations become particularly pronounced for complex queries that require evidence to be retrieved and integrated from multiple documents. To address these challenges, we introduce MUCIRAG, a RAG framework that jointly leverages hierarchical context and citation relationships to guide evidence retrieval. MUCIRAG enriches chunk-level semantic representations with higher-level specification information and uses citation relationships to progressively connect and retrieve complementary evidence across specifications. Experimental results on a 3GPP-related question-answering dataset demonstrate that MUCIRAG consistently outperforms state-of-the-art RAG approaches, with relative improvements of up to 28.83% in accuracy. Furthermore, MUCIRAG covers 91.04%–92.28% of the questions successfully handled by the other approaches.**

***Index Terms*—Third Generation Partnership Project, RAG for Telecom**

I. INTRODUCTION

In global telecommunications, the Third Generation Partnership Project (3GPP) specifications provide the foundational standards that guide vendors in designing and implementing interoperable telecommunication systems [1]. The 3GPP specification corpus contains a vast amount of technical knowledge organized into multiple specification series, with each series covering a particular aspect of the telecommunications system. Beyond this high-level organization, information is further distributed across individual specifications, while the documents are interconnected through cross-references.

This structure makes answering complex questions from 3GPP specifications particularly challenging. The difficulty does not arise solely from the scale of the corpus, but also from the way relevant knowledge is distributed and interconnected across multiple documents [2]. A complex question may require identifying the appropriate specification series, locating relevant documents, and following cross-document references to gather complementary evidence. This process often requires significant effort and domain knowledge, motivating the development of automated approaches that can effectively retrieve and connect relevant information across the 3GPP specification corpus.

Recent approaches have explored automated methods for answering technical questions in telecommunications. Chat3GPP [3] proposes a RAG framework tailored to 3GPP specifications by combining domain-oriented fragmentation strategies with hybrid retrieval. Bornea *et al.* [4] introduce Telco-RAG which implements a dual stage pipeline with query enhancement and retrieval stages, and introduce a neural router for memory-efficient retrieval. Telco-oRAG [5] is the extended version of Telco-RAG that further incorporates a hybrid retrieval strategy combining 3GPP domain-specific retrieval with web search. While these approaches improve access to 3GPP information, they primarily rely on semantic relevance to retrieve individual chunk-level evidence, without leveraging the higher-level context provided by the 3GPP specification hierarchy or the cross-reference relationships connecting related information across specifications. Consequently, they may fail to connect and integrate relevant evidence across multiple levels and specifications, making it challenging to answer complex questions.

To address this challenge, we introduce MUCIRAG, a novel retrieval-augmented generation framework that explicitly leverages the hierarchical organization and citation relationships of the 3GPP specifications to guide evidence retrieval. The key idea is to jointly capture the hierarchical context and cross-document relationships of the 3GPP corpus: MUCIRAG incorporates information from multiple levels of the 3GPP specification hierarchy into a unified representation, which captures both local context and broader specification context, while constructing a citation graph to model relationships among specification documents. Our intuition is that these two complementary sources of information can improve evidence

retrieval: the unified representation provides richer semantic context for identifying relevant evidence, while the citation graph enables the retrieval of related evidence beyond direct semantic similarity.

In particular, MUCIRAG first identifies relevant evidence as reliable anchors and then progressively expands the evidence through citation relationships. MUCIRAG combines series-level, document-level, and chunk-level information into a unified representation for each chunk, enriching its representation with information from higher levels of the specification hierarchy. This representation is consistently used throughout retrieval whenever semantic relevance is required. Starting from the reliable anchors, MUCIRAG traverses the citation graph to discover related specifications and retrieve complementary evidence that may not be directly identified by semantic similarity alone.

We conducted extensive experiments to evaluate the performance of MUCIRAG on a QA dataset widely used in existing telecommunications studies, consisting of 1,840 questions related to 3GPP specifications. The results demonstrate the effectiveness of MUCIRAG in answering complex questions over 3GPP specifications compared with existing approaches.

In brief, this paper makes the following contributions:

1) MUCIRAG: A novel RAG framework that explicitly leverages the hierarchical organization and citation relationships of the 3GPP specification corpus to guide evidence retrieval.
2) An extensive evaluation showing the performance of MUCIRAG over the state-of-the-art approaches.

The rest of this paper is organized as follows. Section II describes the essential basic concepts. Our RAG framework is introduced in Section III. After that, Section IV states our evaluation methodology. Section V presents the experimental results following the introduced methodology and some threats to validity. Section VI provides the related work. Finally, Section VII concludes this paper.

II. BACKGROUND

A. Third Generation Partnership Project (3GPP) specification

The 3rd Generation Partnership Project (3GPP) is a global standardization organization that develops technical specifications for mobile communication systems, providing common standards that enable interoperability among telecommunications vendors and network operators [1]. Established in 1998, 3GPP has continuously evolved its specifications across successive generations of mobile technologies, from 3G and 4G to 5G and beyond. This evolution is organized along a temporal dimension through Releases, with each Release representing a collection of specifications and feature updates developed within a standardization cycle. For example, Release 15 introduced the first major set of 5G specifications, followed by Releases 16, 17, 18, 19, and the ongoing Release 20.

Beyond the temporal dimension, 3GPP specifications are organized according to different technical domains through Technical Specification Groups (TSGs) and Working Groups (WGs). The three main TSGs are RAN (Radio Access Network), SA (Service and System Aspects), and CT (Core Network and Terminals), with each TSG responsible for a set of specialized WGs. Specifications are further classified into technical series, each covering a particular aspect of the mobile communication system; for example, the 23-series focuses on system architecture, while the 38-series covers 5G NR and radio access network specifications.

A distinctive characteristic of the 3GPP corpus is its modular and highly interconnected nature. Technical knowledge is distributed across many specifications, while specifications extensively reuse and reference definitions, procedures, and requirements from other documents. Consequently, answering a technical question often requires combining evidence from multiple documents rather than retrieving an isolated passage.

B. Retrieval-augmented generation

Retrieval-Augmented Generation (RAG) combines information retrieval with language generation to provide large language models (LLMs) with relevant external knowledge [6]. This approach is particularly useful for knowledge-intensive tasks, where relying solely on the knowledge stored in an LLM can lead to limitations such as outdated knowledge and hallucination. Instead of depending only on the model's internal knowledge, RAG retrieves relevant information from an external knowledge base and incorporates it into the input context for generation, thereby improving the factuality and relevance of the generated answer. A typical RAG system consists of a retriever and a generator. The retriever represents the query and documents using a pretrained text embedding model and retrieves relevant documents or passages by comparing their representations in an embedding database. The generator then combines the retrieved information with the original query to construct the context provided to the language model for answer generation. Due to its ability to incorporate external knowledge, RAG has been widely applied to knowledge-intensive tasks such as question answering, text summarization, and content generation.

III. MUCIRAG

This section describes MUCIRAG, our novel citation-aware retrieval-augmented generation framework, which is designed to answer complex questions that require information over 3GPP specifications. A 3GPP expert typically understands a specification not only from its content, but also from its broader context across different levels of the specification hierarchy and its relationships with other specifications. Inspired by this expertise, MUCIRAG captures multi-level specification context to identify reliable initial evidence as anchors, and progressively expands the evidence from these anchors by leveraging cross-document citation relationships. Fig. 1 presents an overview of the MUCIRAG framework, which consists of three main phases: (1) 3GPP Specification Preprocessing, which processes the 3GPP corpus to construct a

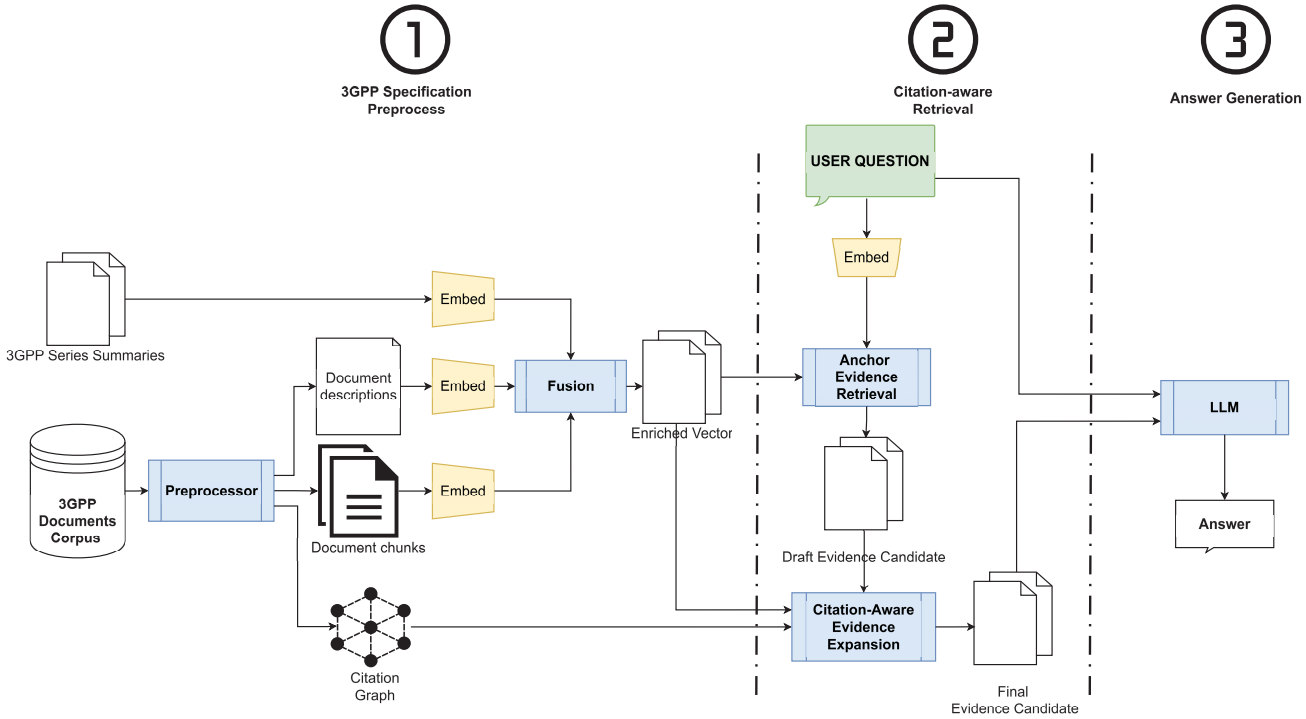


Fig. 1. Workflow of MUCIRAG

unified representation enriched with information from multiple levels and a citation graph that capture cross-document relationships; (2) Citation-aware Retrieval, which identifies reliable anchors and progressively expands the evidence by exploiting the information constructed in the preprocessing phase; and (3) Answer Generation, which produces the final response based on the retrieved evidence. Each component is described in detail below.

A. 3GPP Specification Preprocessing

To facilitate context-aware and citation-aware retrieval, we preprocess the 3GPP specification corpus to construct two complementary types of information: unified multi-level semantic representations and citation graph. The former enriches the representation of each chunk with its broader specification context, while the latter captures explicit relationships among documents for subsequent evidence expansion.

1) Unified Multi-level Representation Construction: A chunk-level representation captures the local semantics of an evidence unit but may not sufficiently reflect its broader context within the 3GPP specification hierarchy. To address this limitation, MUCIRAG incorporates information from three complementary levels: series, document, and chunk. Specifically, for each chunk c, we associate it with its corresponding series summary $s(c)$ and document-level description $d(c)$. For the series level, we reuse the series summaries constructed by Telco-oRAG [5]. For the document level, we use the "Scope" section of the corresponding 3GPP specification. At the lowest level, each specification is segmented using a section-aware chunking strategy, such that the resulting chunk preserves the semantic boundaries and structural context of the original document. We then independently encode these three pieces of information into their respective embedding vectors:

$$\mathbf{e}_s(c) = \mathrm{Emb}(s(c)),\ \mathbf{e}_d(c) = \mathrm{Emb}(d(c)),\ \mathbf{e}_c(c) = \mathrm{Emb}(c)$$

where $Emb(\cdot)$ denotes the embedding function. We then construct an unified representation by weighted fusion of the three embeddings:

$$\mathbf{e}_f(c) = w_s\mathbf{e}_s(c) + w_d\mathbf{e}_d(c) + w_c\mathbf{e}_c(c)$$

where w_s, w_d, and w_c denote the weights assigned to the series-, document-, and chunk-level information, respectively. The resulting $\mathbf{e}_f(c)$ provides a context-enriched representation of the chunk by jointly capturing its local content and broader specification context. This representation is subsequently used throughout the retrieval process whenever semantic relevance is evaluated.

2) Citation Graph Construction: In addition to hierarchical information, 3GPP specifications contain explicit citation relationships that can connect relevant evidence across different documents. We therefore construct a citation graph to support evidence expansion from initially retrieved anchors. Let the citation graph be defined as:

$$G = (V, E)$$

where each node $v \in V$ represents a 3GPP document, and each directed edge $e_{ij} \in E$ represents a citation from document D_i

to document D_j. Each edge additionally preserves the chunks in D_i containing the citation and the target location in D_j.

MUCIRAG distinguishes two types of citation targets. For a citation that explicitly refers to a section in the target document, the edge records the target as (D_j, sec_j). Otherwise, the target is represented only by the document D_j. Formally, a citation edge can be represented as:

$$e_{ij} = (D_i, D_j, C_{ij}, T_{ij})$$

where C_{ij} denotes the source chunks containing citations from D_i to D_j, and T_{ij} denotes the target location, which can be either a specific section sec_j or the entire document D_j. Preserving this target information allows the retrieval process to constrain subsequent evidence search to the referenced section when available, or to the entire target document otherwise.

B. Citation-aware Retrieval

3GPP specifications extensively use cross-references to avoid repeating information across documents. Consequently, critical evidence for complex questions may be found in referenced specifications rather than in text that is directly semantically similar to the query. Relying solely on semantic retrieval over chunks may therefore overlook such complementary evidence. MUCIRAG addresses this limitation by leveraging citation relationships to guide evidence expansion from reliable anchors.

1) Reliable Anchor Retrieval: MUCIRAG first identifies a set of reliable anchor chunks that provide relevant starting points for subsequent evidence expansion. The initial retrieval favors both exact terminology matching and semantic relevance. The initial retrieval combines exact terminology matching with semantic relevance. This is motivated by the nature of 3GPP specifications, which contain many technical terms, abbreviations, and standardized identifiers that can be effectively matched through lexical search.

To capture these complementary signals, MUCIRAG employs a hybrid retrieval function that combines BM25-based lexical retrieval and dense retrieval [7], [8]. Given a query q and a candidate chunk set C, let $r_{\text{Lex}}(q, c)$ and $r_{\text{Dense}}(q, c)$ denote the ranks of chunk c in the rankings produced by lexical retrieval using BM25 and dense retrieval, respectively. We combine the two rankings using weighted Reciprocal Rank Fusion (RRF) [9]:

$$R_{\text{hyb}}(q, c) = \frac{w_{\text{lex}}}{k + r_{\text{Lex}}(q, c)} + \frac{w_{\text{dense}}}{k + r_{\text{Dense}}(q, c)}$$

where w_{lex} and w_{dense} control the contributions of lexical and semantic relevance, respectively, and k is a constant that reduces the impact of the rank position. The dense retrieval uses the unified multi-level representation constructed in the preprocessing stage, thereby allowing semantic relevance to account for the broader specification context of each chunk. The top-N chunks ranked by the hybrid retrieval function are selected as the initial anchor set:

$$A^{(0)} = \text{TopN}\left(R_{\text{hyb}}(q, \mathcal{C})\right)$$

These anchors serve as the starting points for the subsequent citation-aware evidence expansion.

2) Citation-Aware Evidence Expansion: Although the initial anchor chunks provide relevant evidence, complex 3GPP questions may require complementary information from cited specifications. MUCIRAG therefore exploits citation relationships to guide evidence expansion from the retrieved anchors. For each anchor a, its citation target determines the search space: when a specific section sec is referenced, we search only the chunks within sec; otherwise, we search all chunks in the referenced document:

$$\mathcal{C}_a = \begin{cases} \text{Chunks}(D, sec), & \text{if a section is specified,} \\ \text{Chunks}(D), & \text{otherwise.} \end{cases}$$

The same hybrid retrieval function is then applied to the citation-constrained search space to select the top-M relevant chunks:

$$E_a = \text{TopM}\left(R_{\text{hyb}}(q, \mathcal{C}_a)\right)$$

This process is iteratively performed for up to P hops, allowing MUCIRAG to progressively expand the evidence from the initial anchors:

$$A^{(p+1)} = \bigcup_{a \in A^{(p)}} E_a$$

with the final evidence set obtained by aggregating the evidence across all hops:

$$\mathcal{E}_{\text{final}} = \bigcup_{p=0}^{P} A^{(p)}$$

C. Answer Generation

After retrieving the initial anchors and performing citation-aware evidence expansion, MUCIRAG aggregates the retrieved chunks from all retrieval hops to construct the final generation context. The original question and the aggregated evidence are then provided to a general-purpose large language model, which synthesizes the final answer based on the retrieved 3GPP specifications.

IV. EVALUATION METHODOLOGY

To evaluate our RAG framework, we seek to answer the following research questions:

RQ1: ***Accuracy and Comparison.*** How does MUCIRAG perform in QA compared to state-of-the-art approaches?

RQ2: ***Ablation Study*****.** How does each key component to the performance of MUCIRAG?

A. Dataset

In this work, we benchmark our approach on TeleQnA [10], a question-answering dataset previously used to evaluate LLM-based systems for telecommunications [4], [5]. From the original dataset, we select questions related to 3GPP standards to evaluate the capability of RAG systems in answering over 3GPP specifications. Our evaluation set consists of 1,840 questions, including 1,501 "Standard Specification" questions that require in-depth knowledge of 3GPP technical specifications

and 339 “Standard Overview” questions that focus on broader knowledge and understanding of 3GPP standards.

B. Evaluation Procedure

We compare our approach against two existing 3GPP-oriented RAG systems, Telco-RAG [4] and Telco-oRAG [5]. Telco-RAG is a domain-specific RAG approach designed to improve question answering over telecommunications knowledge by retrieving relevant information from 3GPP specifications and further incorporating glossary-enhanced query refinement and a neural router for memory-efficient retrieval. Telco-oRAG extends Telco-RAG by introducing a hybrid retrieval strategy that combines 3GPP domain-specific retrieval with web search. For a fair comparison, we follow the evaluation settings described in the original papers and use the publicly available source code of both approaches to reproduce and evaluate the baselines under the same benchmark setting. We use the same generation models and text embedding model for all approaches. Specifically, we use four state-of-the-art generation models covering both open-weight and closed-weight LLMs: GPT-4o-mini, Mistral Large 3, GPT-5.6 Luna, and Qwen3.8-27B [11]–[14], and use text-embedding-3-large [15] as the text embedding model.

For **RQ2:** ***Ablation Study***, we examine the contribution of the key components of MuCiRAG. Specifically, we evaluate two variants by removing the unified representation and the evidence expansion, respectively. Both variants use Qwen3.8-27B as the generation model to control for the effect of the LLM and text-embedding-3-large as the text embedding model.

V. Experimental Results

A. Accuracy and Comparison (RQ1)

Table I presents the accuracy comparison of MuCiRAG against existing RAG approaches across different generation models and question types. As shown, MuCiRAG consistently achieves the highest overall accuracy, reaching 82.83% with Qwen3.8-27B. Compared with Telco-RAG, MuCiRAG achieves up to 28.83% relative improvement in overall accuracy with GPT-4o-mini, while achieving up to 28.47% relative improvement over Telco-oRAG with GPT-4o-mini. On “Standard Specification” questions, MuCiRAG achieves up to 28.69% relative improvement over Telco-RAG and up to 29.54% relative improvement over Telco-oRAG, both with GPT-4o-mini. These results demonstrate the effectiveness of MuCiRAG in answering complex questions that require in-depth knowledge of 3GPP specifications.

MuCiRAG also maintains stable performance across different generation models. Its overall accuracy ranges from 79.46% with GPT-4o-mini to 82.83% with Qwen3.8-27B, a difference of only 3.37 percentage points. This variation is considerably smaller than those of Telco-RAG and Telco-oRAG, which span 7.72 and 7.88 percentage points, respectively.

We further investigate the questions correctly answered by MuCiRAG and other approaches. Fig. 2 illustrates the sets of questions correctly answered by each approach with Qwen3.8-27B. MuCiRAG correctly answers 291 questions that are not correctly answered by either Telco-RAG or Telco-oRAG. Additionally, MuCiRAG covers 1,100 out of 1,192 questions handled by Telco-RAG (92.28%) and 1,087 out of 1,194 questions handled by Telco-oRAG (91.04%), demonstrating that MuCiRAG preserves the vast majority of the capabilities covered by existing approaches. These results indicate that MuCiRAG not only retains a substantial portion of the questions successfully handled by existing approaches but also provides considerable complementary coverage.

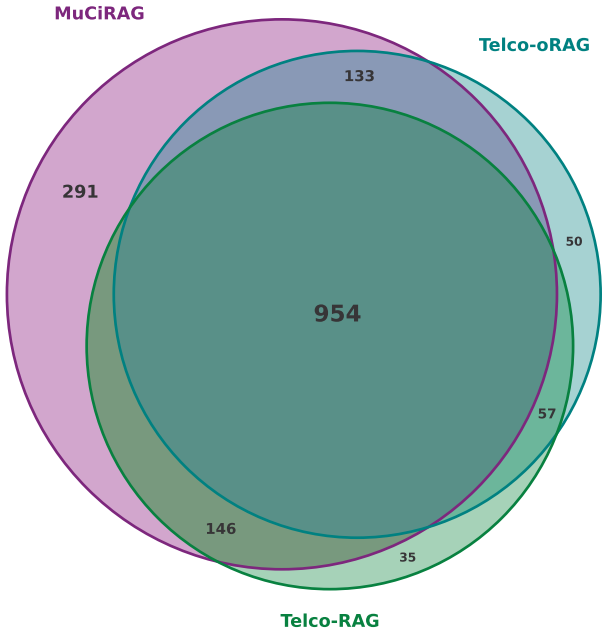


Fig. 2. The sets of questions correctly answered by MuCiRAG and other approaches

Overall, these results confirm the effectiveness of MuCiRAG in retrieving relevant evidence for 3GPP standard-related questions. Its advantage is particularly pronounced for “Standard Specification” questions and for questions that cannot be correctly answered by either baseline, supporting the effectiveness of explicitly incorporating hierarchical and citation-aware information into the retrieval process.

B. Ablation Study

Table II presents an ablation study in answering “Standard Specification” questions, evaluating the contribution of two key components of MuCiRAG: the unified representation and the evidence expansion. As shown in the results, removing the evidence expansion leads to a more pronounced drop in accuracy, from 82.31% to 79.92%, highlighting its importance in answering “Standard Specification” questions. Similarly, removing the unified representation also results in a performance degradation, although less severe than removing the

TABLE I. ACCURACY COMPARISON OF MUCIRAG WITH EXISTING RAG APPROACHES ACROSS DIFFERENT GENERATION MODELS AND QUESTION TYPES. BEST RESULTS FOR EACH GENERATION MODEL ARE SHOWN IN BOLD

Approach	Backbone LLM	Standard Specification Accuracy	Standard Overview Accuracy	Overall Accuracy
Telco-RAG	GPT-4o-mini	60.97	64.95	61.68
	Mistral Large 3	66.53	72.81	67.66
	GPT-5.6 Luna	68.99	71.30	69.40
	Qwen3.8-27B	64.41	66.47	64.78
Telco-oRAG	GPT-4o-mini	60.57	67.67	61.85
	Mistral Large 3	67.79	72.21	68.59
	GPT-5.6 Luna	68.72	74.32	69.73
	Qwen3.8-27B	63.95	69.18	64.89
MUCIRAG	GPT-4o-mini	**78.46**	**83.99**	**79.46**
	Mistral Large 3	**79.85**	**87.61**	**81.25**
	GPT-5.6 Luna	**80.52**	**88.82**	**82.01**
	Qwen3.8-27B	**82.31**	**85.20**	**82.83**

TABLE II. ACCURACY OF DIFFERENT MUCIRAG CONFIGURATIONS ON THE STANDARD SPECIFICATION QUESTIONS

Method	Accuracy
w/o Unified Representation	81.84
w/o Evidence Expansion	79.92
MUCIRAG	**82.31**

evidence expansion, with accuracy decreasing from 82.31% to 81.84%. This suggests that while the evidence expansion plays a more important role in improving the retrieval of relevant evidence, the unified representation also contributes to the overall performance of MUCIRAG by providing richer specification context for retrieval.

C. Thread to Validity

The main threats to the validity of our work consist of internal, construct, and external threats.

Threats to internal validity mainly concern the correctness of our implementation and experimental procedure. One threat may arise from the construction of the multi-level representations and citation graph used in our retrieval process. To reduce this threat, we carefully reviewed our implementation and made it public [16] so that other researchers can double-check and reproduce our experiments. Another threat may come from the choice of retrieval parameters, such as the fusion weights, the number of initial anchors, and the number of citation-expansion hops. To mitigate this threat, we use consistent configurations across experiments.

Threats to construct validity relate to the suitability of our evaluation procedure. We use answer accuracy as the primary evaluation metric, which directly measures the correctness of the generated answers. However, accuracy may not fully capture other aspects of answer quality, such as the quality of retrieved evidence. Besides, a threat may come from the adaptation of the baselines. To reduce this threat, we use the original implementations and configurations whenever available and follow the settings described in their original papers. We also conduct an ablation study how key component contribute the performance of MUCIRAG.

Threats to external validity mainly concern the representativeness of our evaluation dataset. Our experiments are conducted on 1,840 3GPP-related questions selected from TeleQnA. Although these questions cover both standard-specification and standard-overview questions, they may not fully represent all types of questions about 3GPP specifications. Moreover, our evaluation focuses specifically on the 3GPP domain. Thus, the results may not generalize to other technical specification domains. In our future work, we plan to conduct more experiments on larger and more diverse datasets and other telecomunication technical standards.

VI. RELATED WORK

The application of LLMs to telecommunications has attracted growing research interest. Maatouk *et al.* [17], Zhou *et al.* [18], and Soman and Ranjani [19] examined the opportunities and potential research directions for applying LLMs in telecommunications. Bariah *et al.* [20] investigated the use of fine-tuned LLMs for identifying relevant 3GPP working groups. Zou *et al.* [21] further investigated adapting general-purpose LLMs to the telecommunications domain through domain-specific training and Maatouk *et al.* [22] introduced the Tele-LLMs, a family of models with sizes ranging from 1B to 8B parameters.

Several RAG-based systems have been developed specifically for telecommunications. Chat3GPP [3], Telco-RAG [4], and Telco-oRAG [5] employ techniques such as hybrid retrieval, query reformulation, and domain-specific retrieval to improve question answering over 3GPP specifications. While these approaches improve the retrieval of relevant specification content, they primarily rely on semantic relevance at the chunk level. As a result, they do not explicitly incorporate the higher-level context provided by the hierarchical organization of 3GPP specifications or the citation relationships that connect related information across specifications. This limitation can be particularly challenging for complex questions that require relevant evidence to be identified and connected across mul-

tiple specifications. In contrast, MUCIRAG jointly leverages multi-level specification information and citation relationships to identify context-enriched evidence and progressively expand it across related specifications.

VII. CONCLUSION

In this paper, we introduce MUCIRAG, a novel citation-aware retrieval-augmented generation framework for answering complex questions over 3GPP specifications. By incorporating multi-level specification information into a unified representation and exploiting citation relationships among specifications, MUCIRAG enriches semantic relevance with broader specification context and enables citation-aware evidence expansion. These features address the challenge of connecting and integrating distributed evidence across the hierarchical and interconnected 3GPP specification corpus. The experimental results on 1,840 3GPP-related questions show that MUCIRAG significantly outperforms state-of-the-art RAG approaches by 28.47–28.83% in accuracy. Furthermore, MUCIRAG correctly answers 91.04%–92.28% of the questions solved by the other approaches.

REFERENCES

[1] 3rd Generation Partnership Project, "Specifications and technologies," 2026. [Online]. Available: https://www.3gpp.org/specifications

[2] F. A. Rodriguez Y, "Technical language processing for telecommunications specifications," *Applied AI Letters*, vol. 6, no. 2, p. e111, 2025.

[3] L. Huang, M. Zhao, L. Xiao, X. Zhang, and J. Hu, "Chat3gpp: An open-source retrieval-augmented generation framework for 3gpp documents," in *2025 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2025, pp. 492–497.

[4] A.-L. Bornea, F. Ayed, A. De Domenico, N. Piovesan, and A. Maatouk, "Telco-rag: Navigating the challenges of retrieval augmented language models for telecommunications," in *GLOBECOM 2024-2024 IEEE Global Communications Conference*. IEEE, 2024, pp. 2359–2364.

[5] A.-L. Bornea, F. Ayed, A. De Domenico, N. Piovesan, T. S. Salem, and A. Maatouk, "Telco-orag: Optimizing retrieval-augmented generation for telecom queries via hybrid retrieval and neural routing," *IEEE Journal on Selected Areas in Communications*, 2025.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[7] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3," in *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 1995, pp. 109–126.

[8] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and trends® in information retrieval*, vol. 4, no. 1-2, pp. 1–174, 2009.

[9] G. V. Cormack, C. L. Clarke, and S. Buettcher, "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*, 2009, pp. 758–759.

[10] A. Maatouk, F. Ayed, N. Piovesan, A. De Domenico, M. Debbah, and Z.-Q. Luo, "teleqna: A benchmark dataset to assess large language models telecommunications knowledge ", 2023," *URL https://arxiv.org/abs/2310.15051*.

[11] OpenAI, "Gpt-4o mini: advancing cost-efficient intelligence," 2024. [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[12] Mistral AI, "Mistral large 3," 2025. [Online]. Available: https://mistral.ai/news/mistral-large-3/

[13] OpenAI, "Gpt-5.6 luna," 2026. [Online]. Available: https://developers.openai.com/api/docs/models/gpt-5.6-luna

[14] Qwen Team, "Qwen3.8-27b," 2026. [Online]. Available: https://huggingface.co/Qwen/Qwen3.8-27B

[15] OpenAI, "New embedding models and api updates," Jan. 2024. [Online]. Available: https://openai.com/index/new-embedding-models-and-api-updates/

[16] "Mucirag." [Online]. Available: https://github.com/QuangHari1/MuCiRAG

[17] A. Maatouk, N. Piovesan, F. Ayed, A. De Domenico, and M. Debbah, "Large language models for telecom: Forthcoming impact on the industry," *IEEE communications magazine*, vol. 63, no. 1, pp. 62–68, 2024.

[18] H. Zhou, C. Hu, Y. Yuan, Y. Cui, Y. Jin, C. Chen, H. Wu, D. Yuan, L. Jiang, D. Wu *et al.*, "Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 3, pp. 1955–2005, 2024.

[19] S. Soman and R. HG, "Observations on llms for telecom domain: capabilities and limitations," in *Proceedings of the Third International Conference on AI-ML Systems*, 2023, pp. 1–5.

[20] L. Bariah, H. Zou, Q. Zhao, B. Mouhouche, F. Bader, and M. Debbah, "Understanding telecom language through large language models," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 6542–6547.

[21] H. Zou, Q. Zhao, Y. Tian, L. Bariah, F. Bader, T. Lestable, and M. Debbah, "Telecomgpt: A framework to build telecom-specific large language models," *IEEE Transactions on Machine Learning in Communications and Networking*, 2025.

[22] A. Maatouk, K. C. Ampudia, R. Ying, and L. Tassiulas, "Tele-llms: A series of specialized large language models for telecommunications," *IEEE Access*, 2026.