# Context-Structure-Telco-RAG: Structured Context Retrieval-Augmented Language Models for Telecommunications

## 1.Abstract

This paper proposes a framework named context-structure-telco-rag (SC-Telco-Rag) that integrates LLMs and RAG technology for the telecommunications domain. This framework encompasses techniques such as the construction of structured knowledge bases, solutions for table question-answering within documents, and a two-stage fine-tuning[1] strategy for large models. Utilizing the aforementioned LLM + RAG framework, the Specializing Large Language Models for Telecom Networks by ITU AI/ML in 5G Challenge achieved an accuracy of 80.7% with the Phi2-2.7B model and 99.7% with the Qwen1.5-7B-chat model.

## 2. Framework Introduction

### 2.1 Framework Overview

The overall architecture of the model is shown in the following diagram:

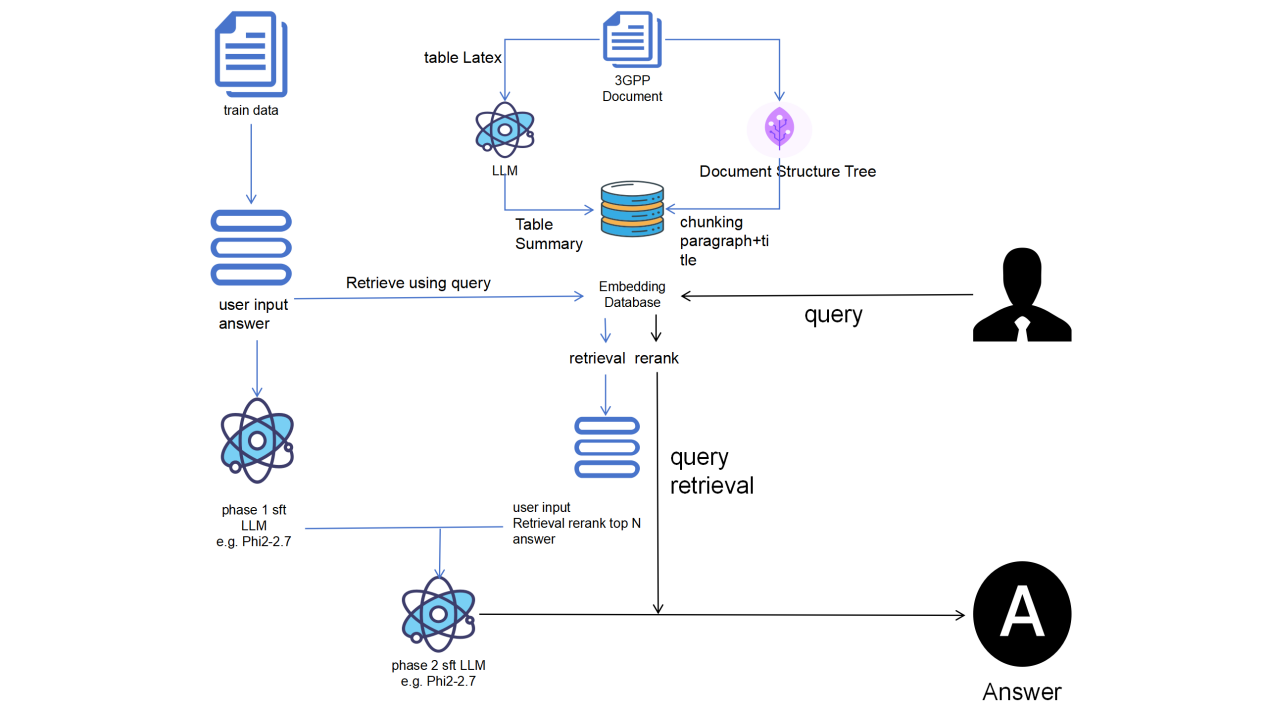


Figure 1. context-structure-telco-rag

The SC-Telco-Rag framework employs a document tree generation method, dividing documents into natural paragraphs and constructing a document tree based on the structure of the document's headings. This ensures that each chunk includes both paragraph information and structural and summary information from the document's headings. For tables and charts within documents, the SC-Telco-Rag framework first converts these elements into LaTeX structures. It then uses advanced large models to summarize the LaTeX representations, incorporating these summaries into the knowledge base. This approach standardizes the format of the knowledge base documents and ensures accurate comprehension of table information by smaller-scale models like Phi2. After identifying several similar chunks based on vector similarity, the SC-Telco-Rag framework uses reranking techniques[2] to enhance the relevance of information retrieval.

The SC-Telco-Rag framework introduces a two-stage fine-tuning approach that integrates SFT (Supervised Fine-Tuning) and RAG techniques by providing context to the large model. In the first stage, conventional fine-tuning methods are applied to enhance the model’s adherence to instructions. The second stage builds upon this by incorporating context information into the prompt, enabling the model to understand and selectively use the information retrieved by RAG.

### 2.2 RAG Knowledge Base Construction

The construction of the Telecom RAG knowledge base utilizes communication protocol documents provided by the competition.[3] These documents inherently possess a structured format with hierarchical headings and corresponding text under each heading. The documents are semantically segmented into paragraphs, which naturally serve as chunking markers. SC-Telco-Rag segments the documents based on these natural paragraphs and employs a recursive chunking method to control document length, incorporating heading information from various levels into the text. This results in the creation of chunked information.

For models with smaller parameter scales, such as Phi2-2.7B, understanding structured textual information like tables can be challenging. SC-Telco-Rag[4] first converts tables into LaTeX format and then uses the LaTeX representation as input for the large model. The model's output is a summary of the table, which is then added to the knowledge base. This approach standardizes the format of the knowledge base documents, facilitating the model's comprehension of tabular information.[5] Another advantage of this method is that it allows the use of larger models to offline summarize tabular data, thereby improving both efficiency and accuracy.

After constructing the knowledge base, documents are vectorized using the BGE-M3[6] model and stored in a FASIS database. During the retrieval phase, the system first indexes the 50 most similar text vectors from the FASIS database. Then, a reranking method is employed to obtain the top 3 most similar results, which are used as context for the query.

### 2.3 Two-Stage Fine-Tuning Framework

The two-stage fine-tuning framework is inspired by ChatQA[7], a method designed for extracting information from chat interactions. When the chat information exceeds the input limit of the large model, RAG is used to incorporate the chat context into the prompt. Although this competition does not involve chat data, a similar fine-tuning framework can still be applied. This paper uses the Phi2-2.7B model as an example to demonstrate the fine-tuning approach.

## 3. Model Experiments

The model experiments utilize the Phi2-2.7B model and include the following ablation studies:

Table 1

|  |  |
| --- | --- |
| **Prediction Method** | Accuracy Score |
| Phi2 Direct Prediction | 46% |
| Phi2 + RAG | 42% |
| Phi2 Original Train Data Fine-Tuning | 70.76% |
| Phi2 + RAG Direct Fine-Tuning | 76.71% |
| Phi2 + RAG Two-Stage Fine-Tuning | 80.75% |

## 4. References

1. Ovadia O, Brief M, Mishaeli M, et al. Fine-tuning or retrieval? comparing knowledge injection in llms[J]. arXiv preprint arXiv:2312.05934, 2023.
2. Gao, L., Dai, Z., Callan, J. (2021). Rethink Training of BERT Rerankers in Multi-stage Retrieval Pipeline. In: Hiemstra, D., Moens, MF., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science(), vol 12657. Springer, Cham. <https://doi.org/10.1007/978-3-030-72240-1_26>
3. Nikbakht, R., Benzaghta, M., & Geraci, G. (2024). TSpec-LLM: An Open-source Dataset for LLM Understanding of 3GPP Specifications. *arXiv preprint arXiv:2406.01768*.
4. Bornea, A. L., Ayed, F., De Domenico, A., Piovesan, N., & Maatouk, A. (2024). Telco-RAG: Navigating the challenges of retrieval-augmented language models for telecommunications. *arXiv preprint arXiv:2404.15939*.
5. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
6. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
7. Liu, Z., Ping, W., Roy, R., Xu, P., Lee, C., Shoeybi, M., & Catanzaro, B. (2024). Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*.