Retrieval-Augmented Generation for Reliable Interpretation of Radio Regulations

Zakaria El Kassimi
KAUST
zakaria.kassimi@kaust.edu.sa

Fares Fourati
KAUST
fares.fourati@kaust.edu.sa

Mohamed-Slim Alouini
KAUST
slim.alouini@kaust.edu.sa

Abstract

We study question answering in the domain of radio regulations, a legally sensitive and high-stakes area. We propose a telecom-specific Retrieval-Augmented Generation (RAG) pipeline and introduce, to our knowledge, the first multiple-choice evaluation set for this domain, constructed from authoritative sources using automated filtering and human validation. To assess retrieval quality, we define a domain-specific retrieval metric, under which our retriever achieves approximately 97% accuracy. Beyond retrieval, our approach consistently improves generation accuracy across all tested models. In particular, while naively inserting documents without structured retrieval yields only marginal gains for GPT-4o (less than 1%), applying our pipeline results in nearly a 12% relative improvement. These findings demonstrate that carefully targeted grounding provides a simple yet strong baseline and an effective domain-specific solution for regulatory question answering. All code and evaluation scripts, along with our derived question–answer dataset, are available at https://github.com/Zakaria010/Radio-RAG.

1 Introduction

Large Language Models (LLMs) have transformed natural language processing, achieving state-of-the-art performance in summarization, translation, and question answering. However, despite their versatility, LLMs are prone to generating false or misleading content, a phenomenon commonly referred to as *hallucination* [4, 10, 22]. While sometimes harmless in casual applications, such inaccuracies pose significant risks in domains that demand strict factual correctness, including medicine, law, and telecommunications. In these settings, misinformation can have severe consequences, ranging from financial losses to safety hazards and legal disputes.

The telecommunications domain presents a particularly challenging case. Regulatory frameworks, and especially the ITU Radio Regulations [12], are legally binding, technically intricate, and demand precise interpretation to ensure compliance. Even small errors can trigger costly service outages, legal disputes, or disruptions to critical infrastructure. Consequently, operators, regulators, and domain experts require reliable tools to assist in interpreting these regulations. As illustrated in Fig. 1, the Radio Regulations corpus exhibits a dense domain-specific vocabulary, underscoring why general-purpose LLMs often struggle in this setting and motivating the need for tailored approaches.

To address these challenges, we introduce a domain-specialized Retrieval-Augmented Generation (RAG) pipeline for regulatory question answering. Our system leverages authoritative external resources to ground LLM outputs, thereby reducing hallucinations and improving reliability [14, 8]. To evaluate this approach, we construct a dedicated dataset of regulatory questions and answers

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

Figure 1: Distribution (word cloud) of key terms in the Radio Regulations corpus, highlighting the domain-specific vocabulary our pipeline must handle

derived from the ITU Radio Regulations [12], validated through both automated checks and human review. To make this system directly accessible to practitioners, we deploy it as a custom *Generative Pre-trained Transformer* (GPT) assistant, which integrates our RAG backend within a conversational interface.

Our contributions are fourfold:

1. We design a RAG pipeline tailored for interpreting and answering regulatory inquiries in the telecommunications domain, with a focus on radio regulations.
2. We create and release a curated dataset of question–answer pairs directly derived from the ITU Radio Regulations, rigorously validated for accuracy and completeness.
3. We present an extensive empirical evaluation across multiple LLMs, showing that our pipeline consistently improves accuracy (e.g., +11.9% absolute for GPT-4o [11]) and achieves up to 97% retrieval accuracy.
4. We deploy **Radio Regulations GPT**, a custom GPT front-end powered by our RAG backend, providing an interactive, reference-backed assistant for querying the ITU Radio Regulations.[1]

By enhancing the precision of automated regulatory interpretation, our work supports more reliable compliance, greater operational efficiency, and improved decision-making across the telecommunications sector.

2 Background

2.1 LLMs

LLMs are transformer-based neural networks [26, 30] trained on massive text corpora to acquire broad language understanding and reasoning capabilities. Recent models such as GPT-4 [11] and GPT-5 demonstrate remarkable progress, yet still exhibit hallucinations and limited long-term memory [4].

LLMs performance can be further enhanced through several complementary strategies. Supervised fine-tuning (SFT) adapts pretrained models to specific domains or tasks by updating their parameters on curated labeled data [20, 9, 31], while reinforcement learning from human feedback (RLHF) aligns their behavior with human preferences through reward-based optimization [17]. Both methods directly modify the model's parametric memory, embedding new knowledge into its weights, yet this knowledge remains static and can quickly become obsolete, especially in rapidly evolving or highly regulated domains [18].

In contrast, prompt learning [34, 13] and prompt engineering [23] guide model behavior without altering its parameters, using carefully crafted instructions or few-shot exemplars to induce desired task behavior at inference time. Relatedly, chain-of-thought (CoT) [28] prompting encourages models to generate intermediate reasoning steps before producing an answer, which improves multi-step reasoning but does not expand the model's knowledge base. While these approaches enhance task adaptability and reasoning ability, they cannot update or extend the underlying knowledge stored in the model's parameters, leaving LLMs vulnerable to hallucinations [4], especially in specialized or fast-changing domains.

These limitations motivate the use of RAG, which complements LLMs and the above approaches with non-parametric memory by retrieving relevant external documents at inference time, thereby grounding their outputs in accurate, verifiable, and up-to-date information.

[1]Link to Radio Regulations GPT:
`https://chatgpt.com/g/g-687bac267f508191a97daf466eccfa50-radio-regulations-gpt`

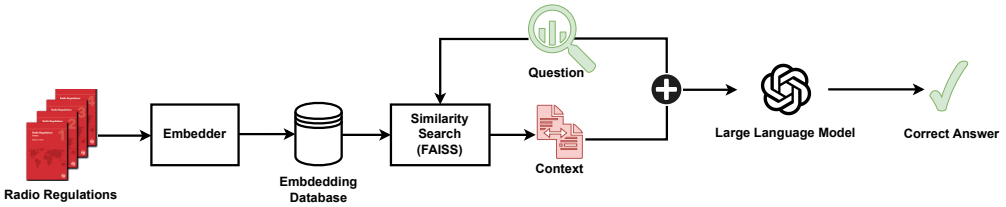

Figure 2: Overview of our Retrieval-Augmented Generation (RAG) pipeline for radio regulations QA, combining FAISS-based retrieval with LLM-based answer generation

2.2 RAG in General

Retrieval-augmented generation (RAG) has emerged as a central paradigm for enhancing LLMs by conditioning them on external knowledge sources rather than relying solely on parametric memory. By retrieving relevant information from corpora, databases, or the web, RAG mitigates key limitations of static LLMs, including hallucination, factual drift, and domain obsolescence.

Recent surveys have mapped the evolution of RAG systems. Gao et al. [6] and Wu et al. [29] categorize approaches, analyzing design dimensions such as retrieval granularity (passages vs. documents), retriever–generator integration (late fusion vs. joint training), and memory augmentation. They also highlight persistent challenges, including reducing hallucinations, handling outdated or incomplete knowledge, and supporting efficient updates in dynamic domains. In parallel, Yu et al. [32] emphasize the need for principled evaluation that captures the hybrid nature of retrieval and generation. Together, these works establish RAG as a promising direction for improving factuality, adaptability, and transparency in LLMs while underscoring unresolved research questions.

2.3 Evaluating RAG

Evaluating RAG remains a challenging open problem due to its hybrid nature. Yu et al. [32] provide a comprehensive survey and propose *Auepora*, a unified evaluation framework that organizes assessment along three axes: **Target**, **Dataset**, and **Metric**.

On the *Target* axis, they distinguish between component-level and end-to-end goals. For retrieval, the focus is on *relevance* (alignment between query and retrieved documents) and *accuracy* (alignment between retrieved content and ground-truth candidates). For generation, they emphasize *relevance* (response–query alignment), *faithfulness* (consistency with supporting documents), and *correctness* (agreement with reference answers). Beyond these, they highlight broader system requirements such as low latency, diversity, robustness to noisy or counterfactual evidence, and calibrated refusal, which are critical for making RAG systems usable and trustworthy in practice.

On the *Dataset* axis, they note that widely used Q&A-style benchmarks often capture only narrow aspects of RAG behavior, motivating domain-specific testbeds that reflect dynamic knowledge and application-specific constraints (e.g., legal, medical, or financial domains).

On the *Metric* axis, they collate traditional rank-based retrieval measures (e.g., MAP, MRR [25]), text generation metrics (e.g., ROUGE [5], BLEU [19], BERTScore [33]), and the growing use of *LLM-as-a-judge* for assessing faithfulness and overall quality. While LLM-based evaluation shows promise, they caution that alignment with human preferences and the need for transparent grading rubrics remain unresolved challenges.

Our evaluation choice. In contrast to prior work that primarily reuses general-purpose benchmarks, we construct our own domain-specific testbed: a multiple-choice dataset derived directly from the ITU Radio Regulations [12] using an automated pipeline with LLM generation, automated judging, and human verification (Section 4.5). This design ensures that the benchmark reflects realistic regulatory queries while retaining ground-truth answers. Accordingly, our end-to-end metric is simply *answer accuracy*. To disentangle retrieval from generation, we further introduce a domain-tailored retrieval metric aligned with component-level *relevance* and *accuracy*. Technical details of this metric are developed in Section 4.1.

3 Related Works

The telecommunications domain poses distinctive challenges for language models due to its dense technical standards, highly structured documents, and precise terminology requirements. Recent efforts have begun adapting Retrieval-Augmented Generation (RAG) frameworks to this space. Bornea et al. [1] introduce *Telco-RAG*, a system designed to process 3GPP specifications, while Saraiva et al. [24] propose *Telco-DPR*, which evaluates retrieval models on a hybrid dataset combining textual and tabular inputs from telecom standards. Maatouk et al. [16] contribute *TeleQnA*, a benchmark for assessing LLM knowledge of telecommunications. Subsequently *Tele-LLMs* [15] were presented as a family of domain-specialized models trained on curated datasets (Tele-Data) and evaluated on a large-scale benchmark (Tele-Eval), demonstrating that targeted pretraining and parameter-efficient adaptation yield substantial improvements over general-purpose LLMs. More recently, Zou et al. [35] developed TelecomGPT, a telecom-specific LLM built via continual pretraining, instruction tuning, and alignment tuning, and evaluated on new benchmarks such as Telecom Math Modeling, Telecom Open QnA, and Telecom Code Tasks; TelecomGPT outperforms general-purpose models including GPT-4, Llama-3, and Mistral in several telecom-domain tasks.

Despite these advances, no publicly available benchmark directly addresses radio regulations, a legally binding and technically demanding domain. Existing resources focus on telecom standards, network operations, or numerical reasoning, but none capture spectrum compliance, licensing rules, interference constraints, or jurisdictional variation. This gap underscores the novelty of our contribution: we introduce the first dataset of multiple-choice questions (MCQ) derived from the ITU Radio Regulations [12], constructed through automated generation, LLM-based judging, and human verification. The dataset not only enables systematic evaluation of RAG in this domain but also provides a reusable testbed for future research.

In contrast to prior work such as Telco-RAG and Telco-DPR, which target 3GPP specifications and generic telecom retrieval, our work focuses explicitly on radio regulations. Furthermore, we show that layering our RAG pipeline onto Tele-LLMs yields accuracy gains without any additional pretraining or fine-tuning, highlighting that carefully designed retrieval and grounding, rather than scaling parametric knowledge, is the key driver of performance in this setting.

4 Methodology

Our RAG pipeline, as shown in Fig. 2, comprises two important sequential blocks: *retrieval* and *generation*, to answer radio regulations questions by grounding the LLM in relevant corpus excerpts.

4.1 Retrieval Block

We compute dense sentence embeddings for each corpus segment using the Sentence-Transformers model all-MiniLM-L6-v2 [21, 27], then build a Facebook AI Similarity Search (FAISS) [3] index over these vectors. At inference, we retrieve the top-k most similar segments to the user query, where k is a tunable parameter. This ensures the generator is supplied only with the most pertinent clauses.

To disentangle retrieval quality from downstream generation, we evaluate the retriever against a ground-truth context. For each question i, let R_i denote the concatenation of the k retrieved chunks and C_i the ground-truth supporting context from our dataset. We compute the ROUGE-L [5] F1 score $F1^{(i)}$ between R_i and C_i. A retrieval is considered correct if

$$F_1^{(i)} \geq \alpha, \quad \alpha \triangleq \gamma, F_{1,\max}, \quad F_{1,\max} \triangleq \frac{2, \min(R, , C)}{R + C}, \tag{1}$$

where R and C are the lengths of R_i and C_i, respectively. That is a retrieved context is accepted as correct if it achieves at least a ratio γ of the maximum achievable ROUGE-L overlap with the ground truth, which reflects the practical requirement that an answer does not demand the entire reference passage, capturing a sufficiently overlapping subset is often enough. In particular, the documents contain redundant clauses, so a strict exact-match requirement would underestimate retrieval quality.

Finally, retrieval accuracy is computed as the fraction of correctly retrieved instances:

$$\mathrm{Acc}_{\mathrm{ret}} \triangleq \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big(F_1^{(i)} \geq \alpha\big). \tag{2}$$

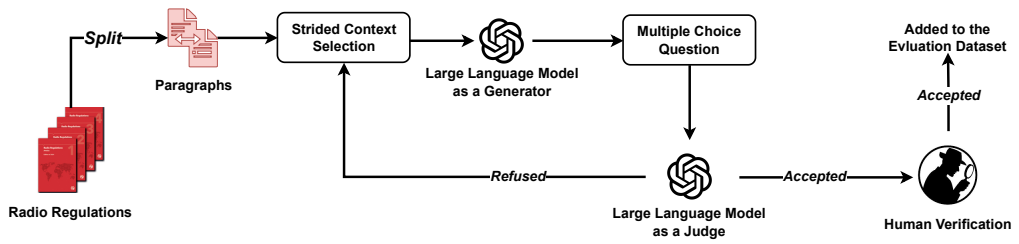

Figure 3: Automated pipeline for generating and validating multiple-choice questions from radio regulations, integrating LLM generation, automated judging, and human verification

4.2 Reranking

Before giving the retrieved chunks to the generator, we optionally use an LLM-based reranker. The goal of this reranker is to provide the generator with the best context in the best order depending on the given query. In our experiments provided in Table 3, reranking yields a modest gain of about 1% absolute accuracy (at the cost of roughly $1.5\times$ higher end-to-end compute time). We disable it by default and recommend enabling it only when compute is abundant and accuracy gains are critical.

4.3 Generation Block

The k retrieved paragraphs plus the MCQ prompt are concatenated into

$$\underbrace{\texttt{Paragraph 1:}\quad r_{i,1}\ \ldots\ \texttt{Paragraph k:}\quad r_{i,k}}_{\text{context}}$$

and prefaced with a system instruction: "You are a radio regulations expert. Answer using the context." We then generate the answer with the chosen model among several models.

4.4 Evaluation

Since the task is MCQ-based, end-to-end performance is measured by standard accuracy:

$$\mathrm{Acc}_{\mathrm{MCQ}} \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big(\hat{a}_i = a_i\big), \tag{3}$$

where $\hat{a}_i$ is the model's selected option and a_i the ground-truth.

By combining these blocks, we can both maximize answer correctness and pinpoint whether any errors arise from retrieval or generation.

4.5 Dataset Construction

We developed a detailed set [2] of questions specifically targeting radio regulations. As summarized in Fig. 3, this process involved extracting clean text from official regulatory documents and automatically generating realistic, domain-relevant questions and answers.

1) Text Extraction and Chunking We extract the full text from telecom-regulation PDF references, then segment it into paragraphs. Let T be the total number of words and M the number of segments.

2) Uniform Sampling of Chunks To maximize coverage, we sample segments in a two-pass strided fashion:

- **First pass:** indices $0, s, 2s, \ldots$ with stride $s = \max\{1, \lfloor M/N \rfloor\}$, where N is the target question count.
- **Second pass:** for each offset $o = 1, \ldots, s-1$, indices $o,\ o+s,\ o+2s, \ldots$ until N accepted questions are reached.

[2] More details about our dataset are available at `https://github.com/Zakaria010/Radio-RAG`.

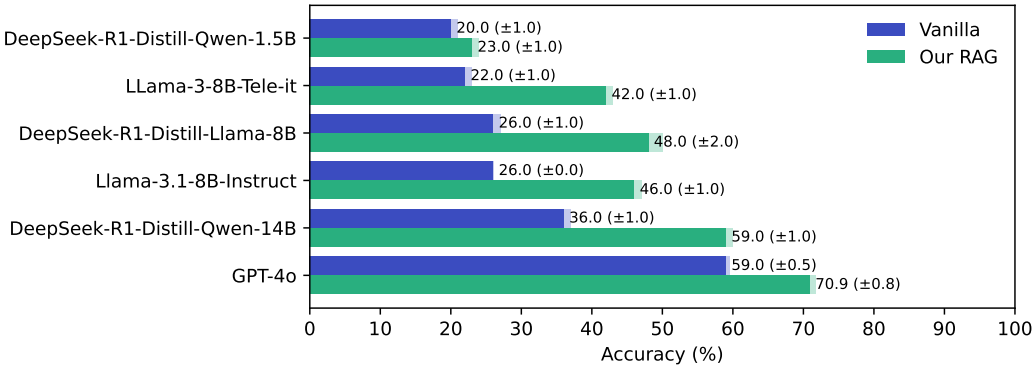

Figure 4: Accuracy comparison of vanilla LLMs versus our RAG-augmented approach, showing consistent gains across models, with GPT-4o achieving the largest improvement

3) Question Generation Each chunk is provided to a text-to-text LLM `google/flan-t5-xxl` [2] with a rigorously defined prompt template enforcing the format:

```
Q: <question>?
Options: A) ... | B) ... | C) ... | D) ...
Answer: <correct option>
Explanation: <justification>
```

4) Quality Filtering Generated Q&A entries are evaluated by a domain-expert judge model `AliMaatouk/Llama-3-8B-Tele` [15]. Only entries judged "Good" are retained; others re-enter the sampling loop.

While this four-step pipeline provides broad coverage of the regulatory corpus, it does not yet include a formal expert-verification stage. We therefore conduct a light human pass to remove obviously illogical questions; incorporating systematic expert review would further improve precision and will become increasingly valuable as the dataset grows.

5 Results and Impact

5.1 Setup

All experiments were run on a Slurm-managed HPC cluster using a single NVIDIA A100 GPU and 200 GB of host memory per job. Retrieval used FAISS-GPU with sentence embeddings from the all-MiniLM-L6-v2 model, indexed over our corpus. Generation used open-source models locally and GPT-4o for comparison with one of the most popular models; the reranker was disabled by default and enabled only in the ablation. We report mean accuracy with standard error across runs, fixing random seeds for repeatability.

5.2 The RAG Results

Our experiments presented in Fig. 4 demonstrate that integrating RAG significantly improves model accuracy across all tested LLMs. Notably, GPT-4o [11] exhibited a +11.9% improvement, suggesting that even sophisticated commercial models significantly benefit from structured retrieval of regulatory contexts. DeepSeek-R1-Distill-Qwen-14B [7] exhibited the largest absolute improvement, +23%. Smaller models like DeepSeek-R1-Distill-Qwen-1.5B [7] showed modest but notable gains of +3%, indicating that retrieval contexts help models of all scales, albeit differently. Examples are provided in the Appendix, Section C.

Interestingly, as shown in Table 1, directly uploading regulatory documents into the prompt, namely GPT-4o + full documents without retrieval, barely improved accuracy, underscoring that structured retrieval is crucial.

5.3 Retrieval-Only Results

Table 4 summarizes results evaluating exclusively the retrieval component, isolating its performance from downstream generation tasks. The table reports retrieval accuracy across various configurations

Table 1: Accuracy of GPT-4o with and without RAG augmentation (mean $\pm$ standard error across runs).

Method	Accuracy	Gain (%)
ChatGPT-4o	$59.0\% \pm 0.5$	—
ChatGPT-4o + full documents (no retrieval)	$59.6\% \pm 0.4$	+0.6
ChatGPT-4o + our RAG	**70.9% $\pm$ 0.8**	**+11.9**

from a comprehensive hyperparameter sweep. Accuracy was measured using a ROUGE-L F_1 similarity threshold set at 0.7. We observe that configurations with chunk sizes of 500 and 700 words and higher values of top-k consistently achieve superior retrieval performance, notably 97% accuracy for the configuration with 700-word chunks and top-7 retrieval. Configurations with smaller chunks, 150 and 300 words, or fewer retrieved contexts, top-k = 3, yielded significantly lower accuracy, demonstrating the importance of sufficient context for capturing regulatory information.

6 Discussion

A key strength of our approach is its robustness to regulatory updates and corpus expansion. Unlike supervised fine-tuning or RLHF, which embed knowledge directly into model parameters and require costly retraining when new information emerges, our RAG pipeline decouples knowledge from the model. The LLM remains fixed while the retriever dynamically accesses an external corpus that can be updated at any time. As new editions of the ITU Radio Regulations are released or additional documents become relevant, they can simply be indexed without retraining, preserving system validity as the regulatory landscape evolves. This modular design also scales naturally to larger corpora, since retrieval performance depends on embedding quality and index structure rather than model size.

Beyond its technical architecture, our work also contributes the first dedicated evaluation dataset for radio regulations, which can serve as a reusable benchmark for future research. By providing standardized, validated multiple-choice questions derived directly from authoritative sources, this dataset enables systematic comparison of retrieval and generation methods in a legally sensitive domain where no prior benchmark existed. It can support the development of domain-specific models, guide fine-tuning or retrieval design, and foster reproducibility by offering a stable testbed for future studies.

Finally, to illustrate the practicality of our approach, we deployed the pipeline as a custom GPT, named Radio Regulations GPT, integrated with ChatGPT via an Actions endpoint. This deployment demonstrates that our RAG architecture can serve as a reliable conversational assistant for practitioners, maintaining grounding in the official ITU corpus while remaining easily updatable as regulations evolve. For more details refer to Appendix B.

7 Conclusion

We presented a domain-specialized RAG framework and the first multiple-choice benchmark for interpreting radio regulations, showing that structured retrieval yields significant gains over vanilla prompting, while naive document insertion provides little benefit. By isolating retrieval from generation, we demonstrated that retrieval can be made reliably accurate. Our findings establish targeted grounding as a simple yet powerful baseline for legally sensitive domains and open directions for advancing generation strategies, refining reranking, and expanding human-verified regulatory datasets. Ultimately, this work underscores RAG's potential to make AI-assisted compliance both more accurate and more trustworthy in high-stakes telecom applications.

References

[1] Andrei-Laurentiu Bornea, Fadhel Ayed, Antonio De Domenico, Nicola Piovesan, and Ali Maatouk. Telco-rag: Navigating the challenges of retrieval-augmented language models for telecommunications, 2024.

[2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, et al. Scaling instruction-finetuned language models, 2022.

[3] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025.

[4] Fares Fourati. A coherence-based measure of agi. *arXiv preprint arXiv:2510.20784*, 2025.

[5] Kavita Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks, 2018.

[6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[8] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions, 2024.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025.

[11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[12] International Telecommunication Union (ITU). Radio regulations (edition of 2024). https://www.itu.int/pub/R-REG-RR, 2024.

[13] Salma Kharrat, Fares Fourati, and Marco Canini. ACING: Actor-critic for instruction learning in black-box LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19086–19113, Suzhou, China, November 2025. Association for Computational Linguistics.

[14] 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. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[15] Ali Maatouk, Kenny Chirino Ampudia, Rex Ying, and Leandros Tassiulas. Tele-llms: A series of specialized large language models for telecommunications, 2025.

[16] Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, Merouane Debbah, and Zhi-Quan Luo. Teleqna: A benchmark dataset to assess large language models telecommunications knowledge, 2023.

[17] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[18] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms, 2024.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002.

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[21] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[22] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models, 2024.

[23] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[24] Thaina Saraiva, Marco Sousa, Pedro Vieira, and António Rodrigues. Telco-dpr: A hybrid dataset for evaluating retrieval models of 3gpp technical specifications, 2024.

[25] Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, 2024.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[27] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[29] Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. Retrieval-augmented generation for natural language processing: A survey, 2025.

[30] Tong Xiao and Jingbo Zhu. Foundations of large language models, 2025.

[31] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023.

[32] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer, 2024.

[33] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[34] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022.

[35] Hang Zou, Qiyang Zhao, Yu Tian, Lina Bariah, Faouzi Bader, Thierry Lestable, and Merouane Debbah. Telecomgpt: A framework to build telecom-specfic large language models, 2024.

A Additional Results

Table 2: Impact of retrieval hyperparameters on RAG accuracy (DeepSeek-R1-Distill-Qwen-14B). Accuracy remains robust across FAISS indexing backends, but chunk size and top-k choices significantly affect performance, with insufficient retrieval (large chunks, low k) leading to accuracy drops.

Index	Chunk Size	Overlap	Top-k	Accuracy
flatl2	400	50	5	**59% ± 1**
flatl2	1000	50	5	58% ± 1
innerproduct	400	50	3	57% ± 1
innerproduct	400	50	5	58% ± 0
innerproduct	700	50	3	**59% ± 1**
innerproduct	700	50	5	57% ± 3
innerproduct	1000	50	3	54 ± 0
innerproduct	1000	50	5	**59% ± 1**

According to Table 2, across indexing and context settings, accuracy remains tightly clustered around $57\% - 59\%$, with tree configurations reaching $59\% \pm 1$. Switching the FAISS backend from inner product to flat L2 does not materially change outcomes given overlapping standard errors, indicating that the index choice is not the limiting factor. The only noticeable dip is at chunk size 1000 with top-k=3 ($54\% \pm 0$),suggesting insufficient evidence when k is too small for long chunks. With large chunks, increasing k to 5 recovers performance to 59%, while with medium chunks (700) a smaller k = 3 avoids a redundant or noisy context and performs best.

Overall, the ablation shows that our RAG is robust as long as the generator receives enough context.

Table 3: Accuracy of DeepSeek-R1-Distill-Qwen-14B with and without RAG/reranking (mean $\pm$ standard error across runs).

Method	Accuracy	Gain (%)
DeepSeek-R1-Distill-Qwen-14B (no RAG)	$36.0\% \pm 1.0$	—
DeepSeek-R1-Distill-Qwen-14B + RAG (no reranking)	$59.0\% \pm 1.0$	+23.0
DeepSeek-R1-Distill-Qwen-14B + RAG (with reranking)	**60.0% ± 1.0**	**+24.0**

Table 4: Retrieval-only accuracy across chunk sizes and top-k settings. Larger chunk sizes (500–700 characters) with higher top-k achieve >95% accuracy, while small chunks (<300) fail to capture sufficient context, confirming the importance of retrieval granularity.

Chunk Size	Overlap	Top-k	Accuracy
500	50	5	93%
500	50	7	95%
700	50	3	92%
700	50	5	95%
700	50	7	**97**%
300	30	3	0%
300	30	5	73%
300	30	7	91%
150	10	3	0%
150	10	5	0%
150	10	7	1%

B Deployment as Radio Regulations GPT

To make our system easily accessible to practitioners and regulators, we deploy the proposed RAG pipeline as a custom GPT, denoted *Radio Regulations GPT*. The RAG backend is exposed as a FastAPI service with a simple retrieval endpoint described via an OpenAPI schema:

Prompt

Question: What is the maximum power flux-density produced at the geostationary-satellite orbit by any Earth exploration- satellite service space station?

GPT-4o (vanilla)

Answer: -168 dB(W/m^2) in a 4 kHz band ✗

Explanation: According to the ITU Radio Regulations (specifically Article 21.16.6), the maximum power flux-density (PFD) produced at the geostationary-satellite orbit (GSO) by any Earth exploration-satellite service (EESS) space station is -168 dB(W/m^2) in any 4 kHz band. This is to protect geostationary satellites from harmful interference from non-geostationary satellites transmitting in the direction of the GSO.

GPT-4o with RAG context

Answer: -174 dB(W/m^2) in any 4kHz band ✓

Explanation: From the provided context: "...the maximum power flux-density produced at the geostationary-satellite orbit by any Earth exploration-satellite service space station shall not exceed -174 dB(W/m^2) in any 4 kHz band."

This directly matches the value. The negative sign is correct—PFD values are often expressed in negative dB(W/m^2) because they are very small.

Figure 5: Qualitative comparison of vanilla GPT-4o versus our RAG-augmented approach on a regulatory question, where RAG retrieves the rule and yields the correct answer.

`POST /query`: takes a JSON payload with the user question (`query`) and a fixed number of retrieved chunks (`top_k`), and returns the most relevant snippets from the indexed ITU Radio Regulations.

This API is integrated into ChatGPT through the Actions interface, so that the Custom GPT can call our backend directly. We configure Radio Regulations GPT with explicit system instructions that: (i) detect when a user request concerns the ITU Radio Regulations or related spectrum-policy questions; (ii) in such cases, call the `/query` endpoint with the full user message as input; (iii) treat the returned snippets as the primary evidence for the answer and synthesize a concise response grounded in these passages; (iv) include precise references to the relevant RR articles, numbers, and tables; and (v) fall back to standard LLM behavior without calling the API for clearly out-of-scope queries. This deployment ensures that answers remain faithful to the official ITU text while providing a convenient conversational interface.

C Question-Answer Examples

The figure 5 shows an example of how RAG efficiently helped GPT-4o [11] to get the correct answer.

Figures 6 and 7 show qualitative MCQ cases; bringing-into-use date for GSO assignments and the HAPS identification band; where our RAG-augmented GPT-4o retrieves the rule and answers correctly, unlike vanilla GPT-4o.

Prompt

Question: What is the date of bringing back into use of the frequency assignment to a space station in the geostationary-satellite orbit?

Options

A. The date on which the recorded assignment is brought back.
B. The day on the commencement of 90-day period defined below.
C. The end of 30 days from the end.
D. The time on 15 days after the completion of.

GPT-4o (vanilla)

Answer: A. The date on which the recorded assignment is brought back. ✗

Explanation: According to the ITU Radio Regulations, specifically the definitions in Article 1 and procedures outlined in Article 11, the "date of bringing back into use" of a frequency assignment to a space station in the GSO is the actual date on which the frequency assignment is resumed and the station begins transmitting again as per its notified characteristics.

GPT-4o with RAG context

Answer: B. The day on the commencement of 90-day period defined below. ✓

Explanation: From the provided context: "The date of bringing back into use of a frequency assignment to a space station in the geostationary-satellite orbit shall be the commencement of the 90-day period defined below." This language (e.g., in sections 20*bis*, 24*bis*, 3211.49.1) explicitly defines the start of the 90-day period as the official date of bringing back into use.

Figure 6: Multiple-choice question with vanilla vs. RAG answers (bringing back into use date).

Prompt

Question: What is the frequency band that is identified for use by high-altitude platform stations (HAPS) in accordance with the provisions of Resolution 166 (Rev.WRC-23)?

Options

A. 25.25–27.5 GHz.
B. 27.0–27.5 GHz
C. 27.5–29.0 GHZ
D. 29.5–31.75 Hz

GPT-4o (vanilla)

Answer: A. 25.25–27.5 GHz ✗

Explanation: According to Resolution 166 (Rev.WRC-23) of the ITU, the frequency band 25.25–27.5 GHz is identified for use by High-Altitude Platform Stations (HAPS). This band is part of the fixed service spectrum and is suitable for broadband connectivity via HAPS systems, which operate in the stratosphere to provide telecommunication services.

GPT-4o with RAG context

Answer: B. 27.0–27.5 GHz ✓

Explanation: According to the provided context, Resolution 166 (Rev.WRC-23) identifies the following bands for HAPS:

- 25.25–27.0 GHz for ground-to-HAPS direction.
- 27.0–27.5 GHz for HAPS-to-ground direction.

This makes Option B (27.0–27.5 GHz) the correct answer.

Figure 7: Multiple-choice question with vanilla vs. RAG answers (HAPS identification band).