RETRIEVAL-BASED QUESTION ANSWERING FOR O-RAN DOCUMENTS

My Journey of Experimentation & Refinement

By

Sri Sai Naga Venkata Adithya Swarna

862467837

THE INITIAL PROBLEM

- O-RAN documentation is thousands of pages long, making manual search inefficient.
- Traditional search methods don't capture context—keyword searches fail in technical documents.
- LLMs hallucinate when they don't have factual context, making them unreliable for technical Q&A.
- The solution? A Retrieval-Augmented Generation (RAG) system that finds relevant context first, then generates an answer.

INITIAL APPROACH

- Short Explanation of the First Version
- First, I extracted text from O-RAN PDF and DOCX files.
- Then, I split documents into chunks to preserve context.
- Used BERT-based embeddings for vector representation.
- Stored embeddings in FAISS for fast similarity search.
- Initially tried GPT-2, but it sometimes produced incomplete or irrelevant responses.
- Switched to Llama-2-7B, which gave grammatically correct but contextually weak responses.
- A simple left-to-right flowchart:
- PDF/DOCX → Text Extraction → Chunking → Embeddings (BERT) → FAISS → Retrieval → LLM Answer

CHALLENGES IN INITIAL APPROACH

- Retrieval wasn’t always accurate—some answers lacked full context.
- BERT embeddings didn’t capture long-range dependencies well.
- FAISS had indexing and performance issues.
- GPT-2 produced coherent sentences but lacked factual correctness.

FIRST BIG CHALLENGE – CHUNKING & RETRIEVAL ISSUES

- The Initial Problem with Chunking
- Early version used fixed-size chunking, which caused loss of contextual information.
- Some chunks split important technical details, making retrieval less useful.
- Retrieval sometimes returned incomplete document sections, leading to low-quality answers.
- Solution: Adaptive Token-Based Chunking (Implemented in Later Version)
- Switched to GPT-2 token-based chunking instead of fixed-size character chunking.
- Dynamically adjusts chunk sizes based on the document’s structure.
- Uses overlap of 100 tokens to maintain continuity across chunks.

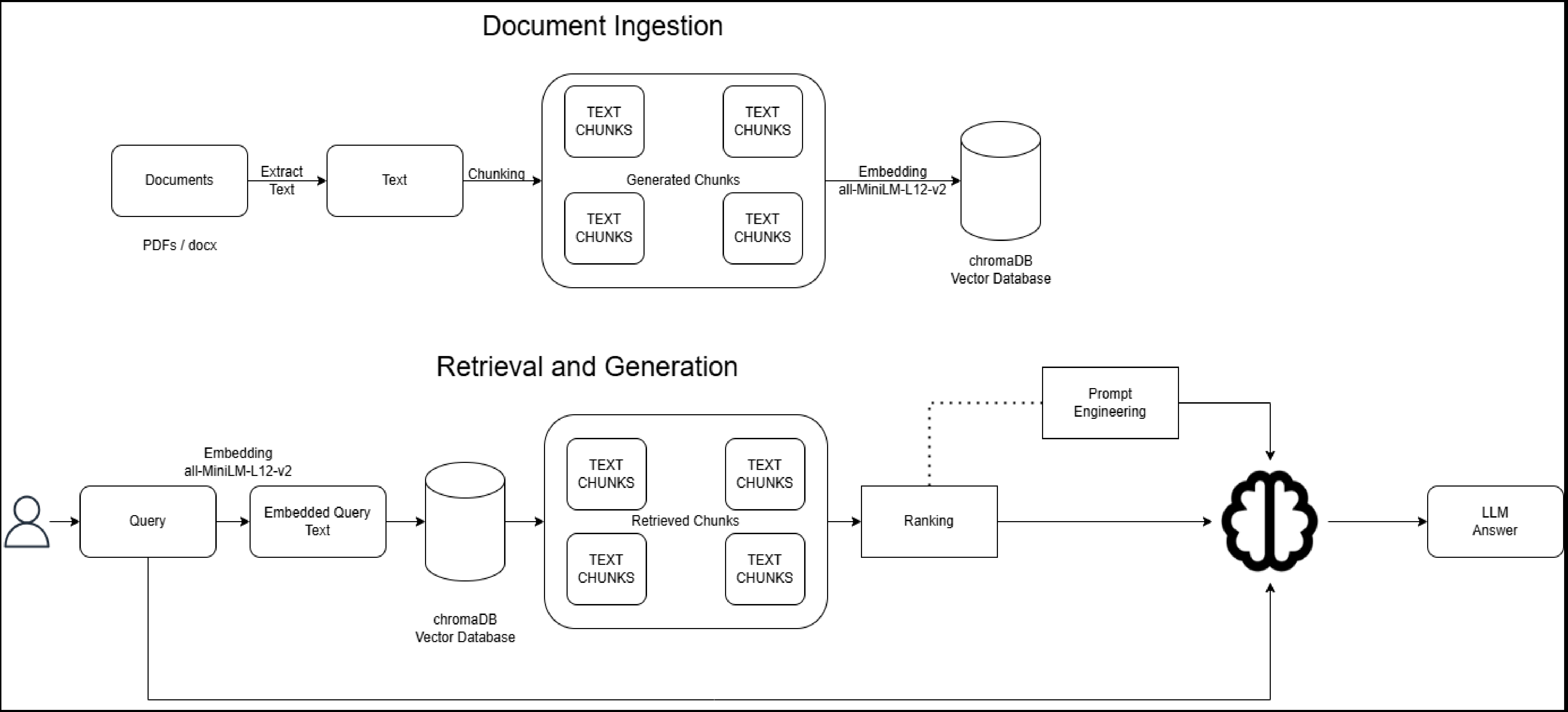
RETRIEVAL STORAGE & METADATA FILTERING (FAISS VS. CHROMADB TRANSITION)

- **Problems with FAISS (Before the Transition)**
  - No real-time document updates – required full re-indexing whenever a new document was added.
  - No metadata filtering – retrieval relied only on vector similarity, leading to irrelevant chunks.
  - Limited hybrid search support – couldn't effectively combine keyword-based search with embeddings.
  - Scaling issues – large-scale document storage became inefficient leading to slow performance.
- **Why ChromaDB Solved These Issues**
  - Supports real-time document updates – no need to rebuild the entire index.
  - Allows metadata filtering – can now retrieve text from specific sections of O-RAN documents.
  - Supports hybrid search – combines semantic search (vectors) with keyword-based search.
  - Scales efficiently – better handling of newly uploaded documents and existing stored documents.

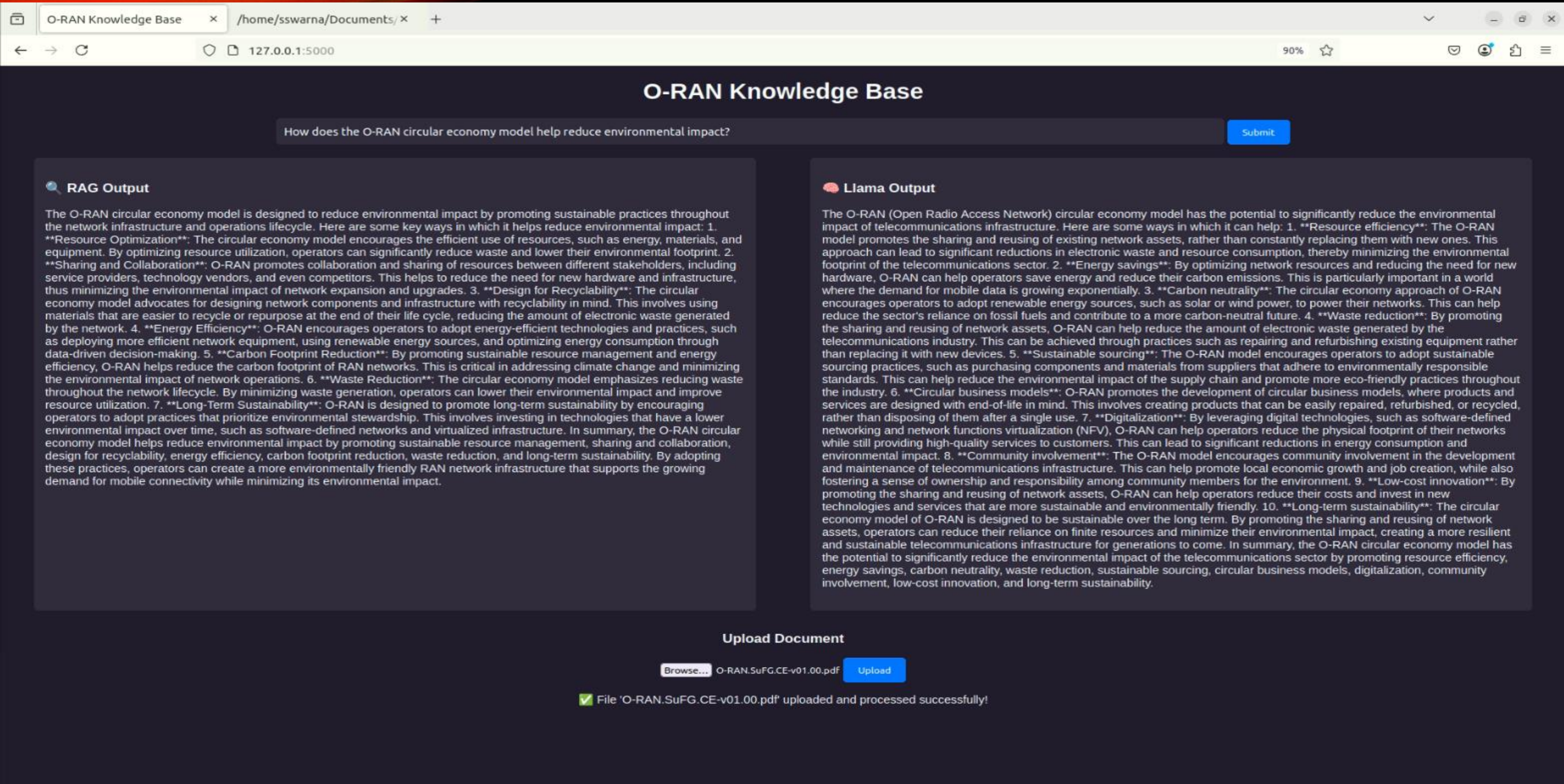
ATTEMPTS TO IMPROVE RETRIEVAL EFFICIENCY

- **Tried Optimizing Retrieval Ranking**
  - **Plan:** Improving ranking should have significantly improved retrieval quality.
  - **Issue:** Despite these optimizations, the improvement was minimal while making the system slower.
  - **Possible Reason:** Retrieval ranking logic may not be prioritizing key metadata properly, leading to inefficient ranking adjustments.
  - **Potential Fix:** Implementing a reinforcement learning-based ranking mechanism to iteratively refine the ranking process.
  - **Challenge:** This would take significant time to test and fine-tune for different types of O-RAN queries.
- **Tried Improving Retrieval Using the LLM Itself**
  - **Plan:** Used dynamic prompt generation and dynamically selected chunks from the top 50 retrieved results.
  - **Issue:** This also showed little improvement but significantly slowed down the system.
  - **Possible Reason:** The overhead of generating multiple prompts and evaluating each chunk dynamically added latency without meaningfully improving precision.
  - **Potential Fix:** Using a two-step retrieval system—first retrieving broader context, then refining ranking using a smaller subset of chunks before querying the LLM.
  - **Challenge:** This would require careful balancing of retrieval depth and speed, making implementation time-consuming.

CURRENT SOLUTION



SAMPLE UI


O-RAN Knowledge Base
/home/sswarna/Documents/
127.0.0.1:5000
90%
O-RAN Knowledge Base
How does the O-RAN circular economy model help reduce environmental impact?
Submit
🔍 RAG Output
The O-RAN circular economy model is designed to reduce environmental impact by promoting sustainable practices throughout the network infrastructure and operations lifecycle. Here are some key ways in which it helps reduce environmental impact: 1. **Resource Optimization**: The circular economy model encourages the efficient use of resources, such as energy, materials, and equipment. By optimizing resource utilization, operators can significantly reduce waste and lower their environmental footprint. 2. **Sharing and Collaboration**: O-RAN promotes collaboration and sharing of resources between different stakeholders, including service providers, technology vendors, and even competitors. This helps to reduce the need for new hardware and infrastructure, thus minimizing the environmental impact of network expansion and upgrades. 3. **Design for Recyclability**: The circular economy model advocates for designing network components and infrastructure with recyclability in mind. This involves using materials that are easier to recycle or repurpose at the end of their life cycle, reducing the amount of electronic waste generated by the network. 4. **Energy Efficiency**: O-RAN encourages operators to adopt energy-efficient technologies and practices, such as deploying more efficient network equipment, using renewable energy sources, and optimizing energy consumption through data-driven decision-making. 5. **Carbon Footprint Reduction**: By promoting sustainable resource management and energy efficiency, O-RAN helps reduce the carbon footprint of RAN networks. This is critical in addressing climate change and minimizing the environmental impact of network operations. 6. **Waste Reduction**: The circular economy model emphasizes reducing waste throughout the network lifecycle. By minimizing waste generation, operators can lower their environmental impact and improve resource utilization. 7. **Long-Term Sustainability**: O-RAN is designed to promote long-term sustainability by encouraging operators to adopt practices that prioritize environmental stewardship. This involves investing in technologies that have a lower environmental impact over time, such as software-defined networks and virtualized infrastructure. In summary, the O-RAN circular economy model helps reduce environmental impact by promoting sustainable resource management, sharing and collaboration, design for recyclability, energy efficiency, carbon footprint reduction, waste reduction, and long-term sustainability. By adopting these practices, operators can create a more environmentally friendly RAN network infrastructure that supports the growing demand for mobile connectivity while minimizing its environmental impact.
🧠 Llama Output
The O-RAN (Open Radio Access Network) circular economy model has the potential to significantly reduce the environmental impact of telecommunications infrastructure. Here are some ways in which it can help: 1. **Resource efficiency**: The O-RAN model promotes the sharing and reusing of existing network assets, rather than constantly replacing them with new ones. This approach can lead to significant reductions in electronic waste and resource consumption, thereby minimizing the environmental footprint of the telecommunications sector. 2. **Energy savings**: By optimizing network resources and reducing the need for new hardware, O-RAN can help operators save energy and reduce their carbon emissions. This is particularly important in a world where the demand for mobile data is growing exponentially. 3. **Carbon neutrality**: The circular economy approach of O-RAN encourages operators to adopt renewable energy sources, such as solar or wind power, to power their networks. This can help reduce the sector's reliance on fossil fuels and contribute to a more carbon-neutral future. 4. **Waste reduction**: By promoting the sharing and reusing of network assets, O-RAN can help reduce the amount of electronic waste generated by the telecommunications industry. This can be achieved through practices such as repairing and refurbishing existing equipment rather than replacing it with new devices. 5. **Sustainable sourcing**: The O-RAN model encourages operators to adopt sustainable sourcing practices, such as purchasing components and materials from suppliers that adhere to environmentally responsible standards. This can help reduce the environmental impact of the supply chain and promote more eco-friendly practices throughout the industry. 6. **Circular business models**: O-RAN promotes the development of circular business models, where products and services are designed with end-of-life in mind. This involves creating products that can be easily repaired, refurbished, or recycled, rather than disposing of them after a single use. 7. **Digitalization**: By leveraging digital technologies, such as software-defined networking and network functions virtualization (NFV), O-RAN can help operators reduce the physical footprint of their networks while still providing high-quality services to customers. This can lead to significant reductions in energy consumption and environmental impact. 8. **Community involvement**: The O-RAN model encourages community involvement in the development and maintenance of telecommunications infrastructure. This can help promote local economic growth and job creation, while also fostering a sense of ownership and responsibility among community members for the environment. 9. **Low-cost innovation**: By promoting the sharing and reusing of network assets, O-RAN can help operators reduce their costs and invest in new technologies and services that are more sustainable and environmentally friendly. 10. **Long-term sustainability**: The circular economy model of O-RAN is designed to be sustainable over the long term. By promoting the sharing and reusing of network assets, operators can reduce their reliance on finite resources and minimize their environmental impact, creating a more resilient and sustainable telecommunications infrastructure for generations to come. In summary, the O-RAN circular economy model has the potential to significantly reduce the environmental impact of the telecommunications sector by promoting resource efficiency, energy savings, carbon neutrality, waste reduction, sustainable sourcing, circular business models, digitalization, community involvement, low-cost innovation, and long-term sustainability.
Upload Document
Browse...
O-RAN.SuFG.CE-v01.00.pdf
Upload
✅ File 'O-RAN.SuFG.CE-v01.00.pdf' uploaded and processed successfully!

CURRENT SOLUTION

- **Dynamic Document Ingestion**
  - Extracts text from PDFs and DOCX files.
  - Uses adaptive token-based chunking (instead of fixed-size chunking) for better retrieval accuracy.
  - Converts text into embeddings using all-MiniLM-L12-v2 and stores them in ChromaDB.
- **Retrieval & Query Processing**
  - User query is converted into an embedding using all-MiniLM-L12-v2.
  - ChromaDB retrieves the most relevant document chunks based on semantic similarity.
  - Ranking mechanism ensures the best chunks are prioritized before passing to LLM.
- **Generation with LLM**
  - Retrieved chunks are fed into the LLM along with the user query.
  - Some carefully designed prompts were used to structure the input for better response accuracy.

THE CHALLENGE

- The Challenge – Retrieval from Both Old & New Documents
- Terminal-based system: Works only with preprocessed O-RAN documents.
- UI-based system: Allows document uploads, but retrieval doesn’t merge uploaded & stored documents correctly.
- Issue: Queries retrieve from uploaded docs OR stored docs—but not both together.
- Example Problem:
- If a new document is uploaded, retrieval prioritizes it over older documents.
- If retrieval includes stored docs, the uploaded document doesn’t contribute effectively.

PERFORMANCE & EVALUATION

Metric	Why It's Used	How It Works	Score Ranges	What It Tells About a RAG System
BLEU (Bilingual Evaluation Understudy)	Measures similarity to a reference text.	Compares n-grams, penalizes missing words/repetitions.	0.7 – 1.0 = Excellent 0.5 – 0.7 = Good < 0.5 = Needs Improvement	High = Closer to ground truth. Low = Possible poor retrieval or hallucination
ROUGE-1 (Recall-Oriented Understudy for Gisting Evaluation)	Measures recall-based word overlap.	Counts unigram matches with reference text.	> 0.7 = Strong recall 0.5 – 0.7 = Good < 0.5 = Needs Improvement	High = Relevant words retrieved. Low = Missing key information.
Semantic Similarity	Checks if meaning matches, even if wording differs.	Uses cosine similarity between sentence embeddings.	> 0.9 = Excellent 0.5 – 0.9 = Good < 0.5 = Poor	High = Meaning preserved. Low = Incorrect context or hallucination.
KG Score (Knowledge Grounding Score)	Ensures response is fact-based.	Extracts entities/relations and checks alignment with retrieved text.	> 0.8 = Highly Factual 0.5 – 0.8 = Good < 0.5 = Risk of hallucination.	High = Answer grounded in docs. Low = Possible fabricated details.

SAMPLE METRICS SCORES

📊 **Final Evaluation Summary** 📊

Category	Query	BLEU	ROUGE-1	Semantic Similarity	KG Score
Factual Questions (Basic Retrieval)	What are the security measures in O-RAN?	0.0014	0.0891	0.7388	0.5885
Factual Questions (Basic Retrieval)	Describe the architecture of O-RAN Near-RT RIC.	0.0013	0.0468	0.6616	0.5757
Comparative Questions (Complex Retrieval)	Compare the roles of Near-RT RIC and Non-RT RIC in O-RAN.	0.0032	0.1250	0.6982	0.7842
Comparative Questions (Complex Retrieval)	How does O-RAN differ from traditional RAN architectures?	0.0009	0.0307	0.6452	0.6422
Summarization Questions	Summarize the O-RAN security framework.	0.0116	0.0870	0.7976	0.6732
Summarization Questions	Provide a high-level overview of O-RAN architecture.	0.0015	0.0739	0.5043	0.6245
Ambiguous Queries (Robustness Check)	Tell me about RIC.	0.0014	0.0435	0.5796	0.5988
Ambiguous Queries (Robustness Check)	What's new in O-RAN?	0.0014	0.0490	0.5003	0.6009
Misleading Queries (Hallucination Prevention)	Does O-RAN use blockchain for security?	0.0031	0.0816	0.7905	0.6075
Misleading Queries (Hallucination Prevention)	What is the O-RAN 6G standard?	0.0021	0.0459	0.6290	0.6413

- Factual & Comparative Queries → High similarity (~0.7+), showing retrieval is effective.
- Summarization Queries → Slightly lower scores, meaning some key details were lost.
- Ambiguous Queries → Lower similarity (< 0.6), indicating retrieval ranking needs improvement.
- Misleading Queries (Hallucination Test) → KG Score ~0.6, meaning most responses stay fact-based, but can improve.
- Takeaway: The system retrieves relevant information well, but ranking needs optimization for certain queries.

WHY SOME SCORES ARE LOW

- **LLM Provides More Detailed Responses Than Reference Answers**
  - The model may expand on the topic with additional relevant details.
  - Example: If the reference answer is "O-RAN uses encryption for security," and the model responds with "O-RAN employs encryption, authentication, and access controls," the BLEU score may be low, but the model's response is more informative.
- **Reference Answers May Not Be Fully Optimized**
  - The test answers were written manually, but a domain expert might phrase them differently or provide additional context.
  - This doesn't mean the LLM response is wrong—just that it used different words or expanded on the concept.
- **Short Reference Answers Penalize the LLM for More Comprehensive Responses**
  - If the reference answer is short, but the model provides a detailed and accurate response, BLEU/ROUGE scores drop because the extra words don't match exactly.
  - Solution: Use more comprehensive reference answers during evaluation.

FUTURE WORK & NEXT STEPS

- Improve Retrieval Merging for Uploaded & Stored Documents.
- Fine-Tune Retrieval Ranking for Ambiguous Queries.
- Resolve Confusion Between Similar Terms (e.g., Near-RT RIC & Non-RT RIC).
- Experiment with Alternative Embedding Models.
- Improve Evaluation with Human Annotation.
- Enable Real-Time ChromaDB Updates Without Full Re-Indexing.

CONCLUSION

- **What Was Achieved?**
  - Developed a Retrieval-Augmented Generation (RAG) pipeline for O-RAN documents.
  - Implemented document chunking, embedding, retrieval, and response generation using ChromaDB & Llama-2-7B.
  - Optimized retrieval accuracy, document merging, and hallucination prevention.
- **Major Challenges Overcome**
  - Reduced hallucinations by grounding responses in retrieved documents.
  - Enhanced system scalability by transitioning from FAISS to ChromaDB.
- **Future Impact & Applications**
  - Can be expanded to multi-document context awareness for better technical document retrieval.
  - Potential to fine-tune the LLM for O-RAN-specific question-answering.
  - Can evolve into a fully functional web-based tool for real-time document querying.
- **Final Takeaway**
  - This project built a solid foundation for scalable, document-aware AI retrieval.
  - Further refinements can make it a robust, industry-level RAG system.


THANK YOU!