REPORT ON RAG CHATBOT

By Jennifer Teresa George

[jenniferteresageorge@gmail.com](mailto:jenniferteresageorge@gmail.com)

+91 9003145343

Implementation Choices

We built a Retrieval-Augmented Generation (RAG) chatbot that uses context retrieved from hotel booking data to answer user queries more accurately. The core idea was to avoid hallucinations by grounding the responses in real data.

We used FAISS to index and retrieve the most relevant pieces of text based on user queries. This ensures responses are context-aware and not just general language model outputs.

Initially, we used the OPT model (facebook/opt-1.3b), but due to poor output quality and repetition, we switched to Cohere's `command-r-plus`, which significantly improved response relevance, tone, and coherence.

FastAPI was chosen for the backend to build lightweight API endpoints quickly, and we exposed two routes:

`/api/ask` for RAG-based Q&A and `/api/analytics` for data-driven summaries.

Challenges

1. **Low-Quality Output:** The original OPT model output was repetitive and incoherent. We resolved this by switching to Cohere, which delivered faster and more accurate answers.

**2. Prompt Leakage:** At first, the generated responses included the entire input prompt, which looked

messy. We refined the prompt template and trimmed the outputs to fix this.

**3. Slow Inference:** Local models took too long to respond. Cohere helped reduce latency while

maintaining good quality.

**4**. **Duplicate Context**: The same sentence was sometimes retrieved multiple times. We addressed this by

setting `top\_k=3` and removing redundant results before generating the final answer.

Final Thoughts

The system now generates coherent, relevant answers using real data and performs well under typical usage. It can be extended further with conversation memory or a larger dataset, but for its current scope, it is clean, efficient, and works as intended.

The combination of semantic search and a robust language model made this a strong, extensible solution for hotel-related customer queries.