

Build a RAG-Powered Chatbot and Self-RAG Deep Dive



Ankit Khare

DevRel @Rockset

Agenda



Part 1

Diving into Self-RAG

- What is RAG
- Limitations of RAG
- Intro to Self-RAG
- Insights into Self-RAG

20 min

Part 2

Building a Chatbot

Hands-on session on Google Colab building a RAG-powered chatbot using OpenAI embeddings to vectorize data, Rockset database for vector similarity search, and Streamlit for creating an interactive web interface for the chatbot.

40 min

Part 3

Q&A

Ask away all your questions !

15 min and beyond

Output / Outcome

:

RockGPT

Retrieval Augmented Generation based Chatbot using Rockset and OpenAI

Enter your query:

What is the report about?

Submit

Response:

The report is a financial document that provides insights into a company's financial performance over specific fiscal years, namely 2022, 2021, and 2020. It contains detailed financial data, including revenues, net income, operating income, earnings per share, and growth percentages, among other financial metrics.

Key elements discussed include:

1. Revenue and income performance across different product and service offerings like cloud services, gaming, advertising, and devices.
2. Revenue recognition policies.
3. Unearned revenue by segment and changes in unearned revenue for the fiscal year.
4. Information about operating income by segment and geographical distribution.
5. Revenue from various company segments such as Productivity and Business Processes, Intelligent Cloud, and More Personal Computing.
6. Costs associated with revenue, such as product manufacturing and distribution, data center and cloud services, and traffic acquisition.
7. Specifics regarding critical audit matters related to revenue recognition.
8. Tax-related discussions, including impacts of tax benefits and settlements.

Conceptual Overview

LLMs	RAG	Self-RAG
Solely rely on internal knowledge <ul style="list-style-type: none">• Limited• Not current• Expensive training• Not able to handle real-time stuff	Supplementing LLMs with external knowledge <ul style="list-style-type: none">• Reduces Hallucinations but what happens when:<ul style="list-style-type: none">- knowledge base is lacking- prompt wasn't optimal- Retrieved docs not relevant	LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION <ul style="list-style-type: none">• LLMs decides if retrieval is needed• Not a concat but a real critique before generation

RAG and Self-RAG

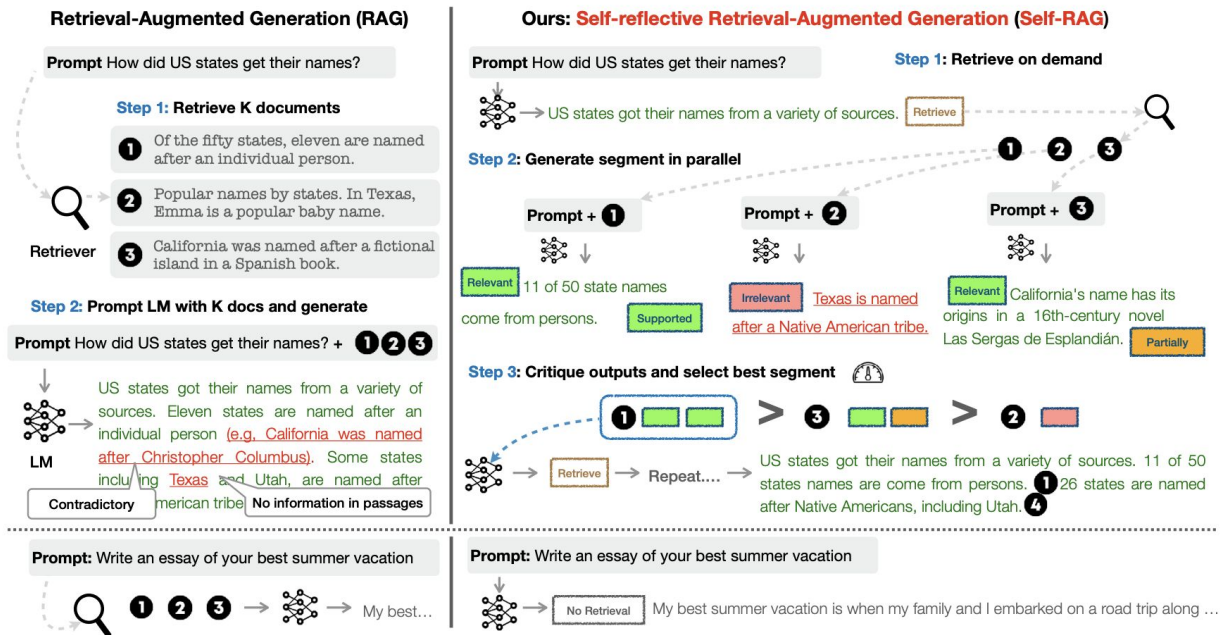


Figure 1: Overview of SELF-RAG. SELF-RAG learns to retrieve, critique, and generate text passages to enhance overall generation quality, factuality, and verifiability.

Follow us for upcoming workshops and more:

- **Workshop resources -**

https://drive.google.com/drive/folders/19Iw8XoO_tpGVDAcQw-cfR5I4mOLhfgfe?usp=sharing

- **Subscribe on YouTube**, turn the notifications on, and get an alert whenever workshop recording is available - https://www.youtube.com/channel/UCy4qLzJ7yuEmsIN2Mm5Pn-w?sub_confirmation=1
- Rockset LinkedIn - <https://www.linkedin.com/company/rocksetcloud>
- My LinkedIn - <https://www.linkedin.com/in/deeplearnerak/>
- Interested in building search, analytics, and AI applications, get your free slot for the **Index conference**, May 16 2024 - <https://rockset.com/index-conf/>
- Questions! Feel free to reach out to me - ankitkhare@rockset.com
- Rockset Product support - support@rockset.com