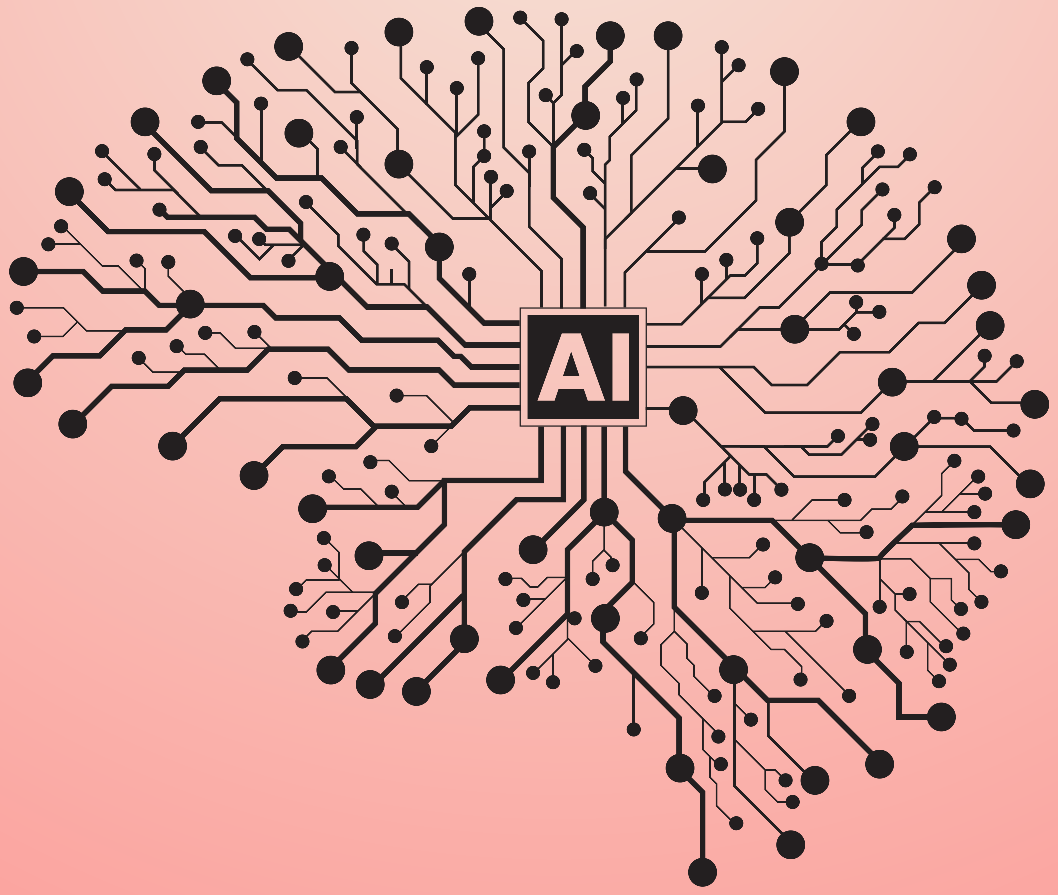
RAG interview questions along with clear answers.

Here are some commonly asked RAG (Retrieval-Augmented Generation) interview questions along with concise and clear answers. These are great for technical interviews, ML roles, or LLM-related positions:

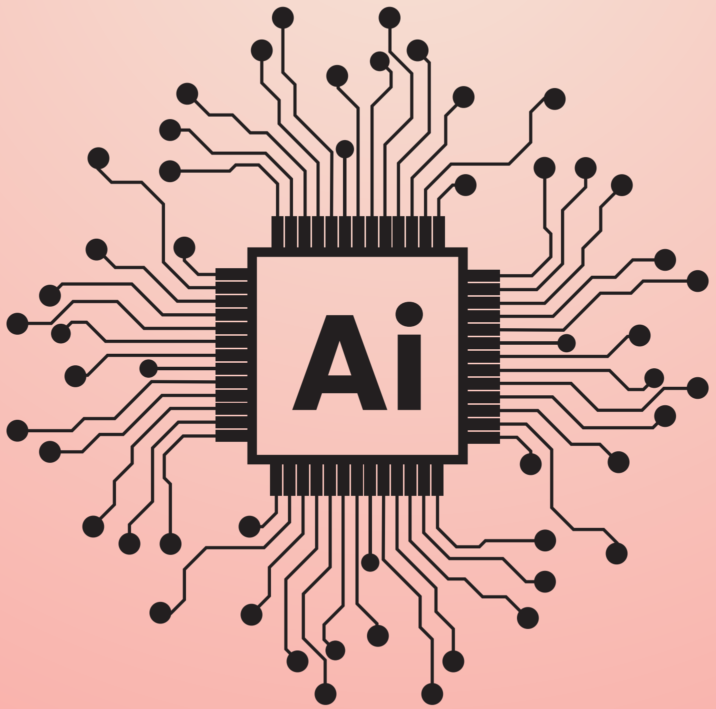
◆ 1. What is RAG in the context of LLMs?

RAG (Retrieval-Augmented Generation) is an architecture that combines information retrieval with text generation. It enhances LLMs by retrieving relevant documents from an external knowledge base and then using them as context to generate more accurate and up-to-date responses.



2. Why is RAG important for enterprise use cases?

RAG allows models to access custom or private knowledge without retraining. This enables use cases like customer support, internal document Q&A, and compliance checks using accurate, real-time data.

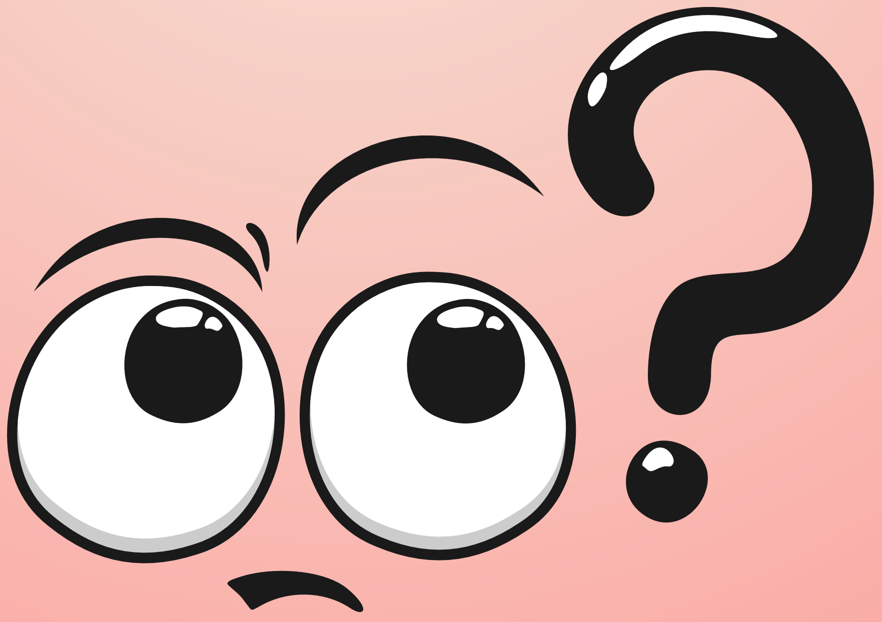


◆ 3. How does RAG differ from traditional fine-tuning?

In fine-tuning, model parameters are updated to fit new data, which is compute-intensive and static. RAG keeps the base model unchanged and adds context via retrieval—making it more flexible, lightweight, and easier to update.

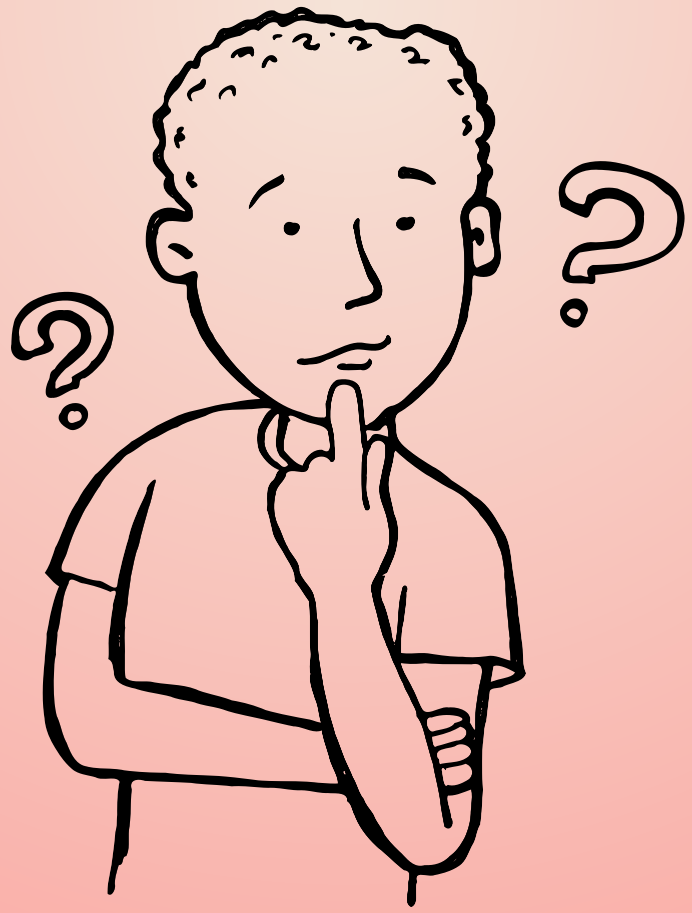
◆ 4. What are main components of a RAG pipeline?

- Retriever: Fetches relevant documents (usually using vector search).
- Generator: Uses an LLM to generate answers based on retrieved context.
- Embedding model: Converts text into vector format for semantic search.



◆ 5. What retrievers are commonly used in RAG systems?

Popular options include FAISS, Weaviate, Pinecone, Elasticsearch, or open-source tools like LangChain + Chroma for in-memory retrieval.



◆ 6. What are some limitations of RAG?

- Retrieval quality can heavily affect generation.
- May still hallucinate if context is weak.
- Requires a well-structured and updated knowledge base.
- Latency increases due to the retrieval step.

◆ 7. How can you evaluate a RAG system?

Evaluate both retrieval and generation:

- Retrieval: Precision@K, Recall, MRR.
- Generation: BLEU, ROUGE, factual correctness, human review.
- Also, task-specific metrics like exact match or semantic similarity scores.

CONGRATULATIONS

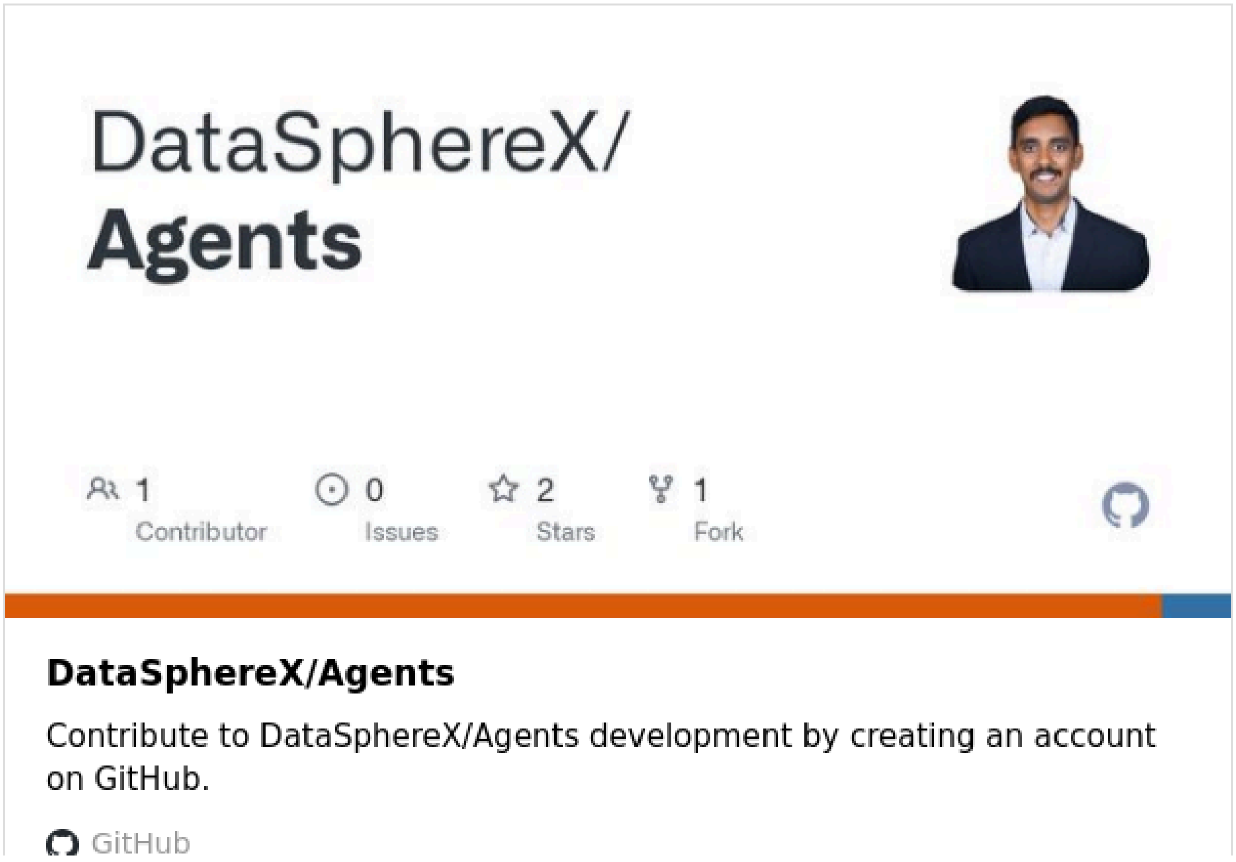
You have reached the end, now

If you want to help your network

REPOST THIS

Sarveshwaran R

https://github.com/DataSphereX/Agents



DataSphereX/
Agents

1 Contributor

0 Issues

2 Stars

1 Fork

DataSphereX/Agents

Contribute to DataSphereX/Agents development by creating an account on GitHub.

GitHub