5 Questions Based on the Document:

- **How does Retrieval-Augmented Generation (RAG) architecture work in LLMs?**
- **What are the main differences between RAG and traditional language models?**
- **Why is hallucination a common problem in LLMs, and how does RAG help reduce it?**
- **In what real-world use cases is RAG more effective than standard LLMs?**
- **How does the retriever module impact the quality and relevance of generated answers in RAG systems?**