Yogesh Haribhau Kulkarni • You
AI Advisor (Helping organizations in their AI journeys) | PhD (Geometric Modeling) | Tech Colum...
now •

Retrieval Augmented Generation (RAG) based chatbots are in high demand. RAG is like an open-book exam, except the “books” can be far many, and the system has to find answers quickly.

No matter how powerful the LLMs (Large Language Models) become or how much their context window grows, building a RAG system remains essential. The reasons are:
➡️ The need for answers from private data, which isn’t part of any LLM’s training. And we can't get that 'in' LLM as we often lack curated datasets, budget, or AI expertise to build a pretrained or fine-tuned model.
➡️ The need for answers from streaming or near real-time data.
➡️ The need for grounding and minimizing hallucinations.
➡️ The need for cost, speed, and efficiency by retrieving only the most relevant context for the LLM to answer.

Having worked on some non-trivial RAG implementations, I’ve realized that production-level enterprise RAG is far more than what some compare it to, ie just uploading documents to ChatGPT and asking questions. Don’t believe it? Upload a finance report and ask about an entry in a table, you’ll see what I mean.

You HAVE to build custom/bespoke, home-grown RAG systems. Parsing , Chunking, Multimodal Partitioning, Delegation, and Evaluation are the keys. Unless you’ve gone through this yourself, it’s hard to grasp just by watching YouTube videos.

RAG works across diverse data, text, images, videos, code, tables, etc. It can be sequential or graph-based. At its core is index-and-search functionality, and for those who are research-oriented there is one unexplored modality, geometry. Imagine querying across house floor plans or 3D shapes!

I’m very much interested in this topic and upgrading myself every day. One way to grow is by discussing others’ problems and learning from them.

If you have any questions or would like a sounding board for your RAG implementations, feel free to comment below or contact me at yogeshkulkarni at yahoo do com. These interactions will be informal, voluntary, non-monetary, and happen on Saturdays, as my weekdays are packed with other commitments. 🚀 📂 🤖

#chatbots #retrievalaugmentedgeneration