5 Blog Posts To Become a RAG Master

RAG limitations, Tips on performance and production, Applications to specific domains, New methods for retrieval

In this post, I'm providing you with a list of interesting reads about RAGs.

This list doesn't cover introductory concepts nor does it showcase simple demos of 5-line Langchain snippets.

> ***Rather, it discusses the limitations of RAG-based systems, offers valuable insights on deploying them in production, and shares a startup's experience applying RAGs to the insurance sector.***

This curated list is the result of thorough documentation I conducted while building conversational agents for the pharmaceutical industry.

I hope you find it useful to you or your teams.

> PS: if you're new to Retrieval Augmented Generation (RAG), it's a pattern used to develop chat-based applications over custom data. I made an introductory post on RAGs here. Check it out!

Now, get ready to become a RAG master. ♜ 🚀

Building RAG-based LLM Applications for Production

🔗 Link 👨‍💻 Authors: **Goku Mohandas and Philipp Moritz**

> **Disclaimer**: One of the authors (Goku Mohandas) is an excellent technical writer and open-source developer. He's particularly famous for his website (madewithml.com) that taught thousands of developers to productionize ML systems.

This post is a production guide for building RAG-based LLM applications.

It's by far the best piece I've read on this topic. I'm currently using it as a guide to improve my existing RAG-based apps. If you too are getting serious about RAGs, this post is a ***must-read***.

This tutorial is also well-written, detailed, and easy to follow. Plus, the corresponding code is available on Github.

This is a long read but here's what you'll learn if you go through it:

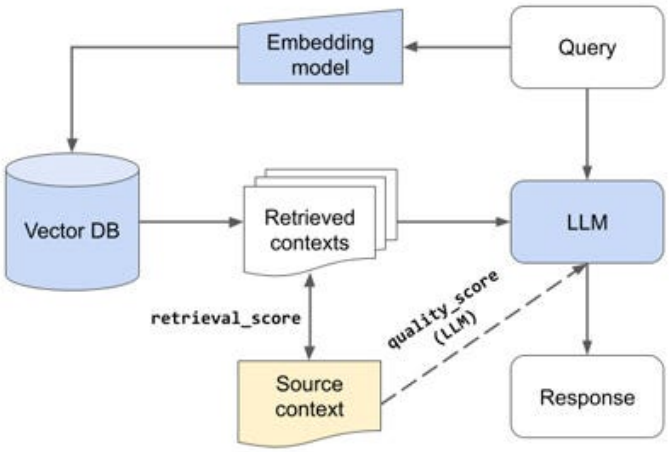
- 💻 Develop a retrieval augmented generation (RAG) LLM app from scratch.
- 🚀 Scale the major workloads (load, chunk, embed, index, serve, etc.) across multiple workers.
- ✅ Evaluate different configurations of our application to optimize for both per-component (ex. retrieval_score) and overall performance (quality_score).
- 🔀 Implement an LLM hybrid routing approach to bridge the gap b/w OSS and closed-source LLMs.
- 📦 Serve the application in a highly scalable and available manner.
- 💥 Share the 1st-order and 2nd-order impacts LLM applications have had on products and organizations

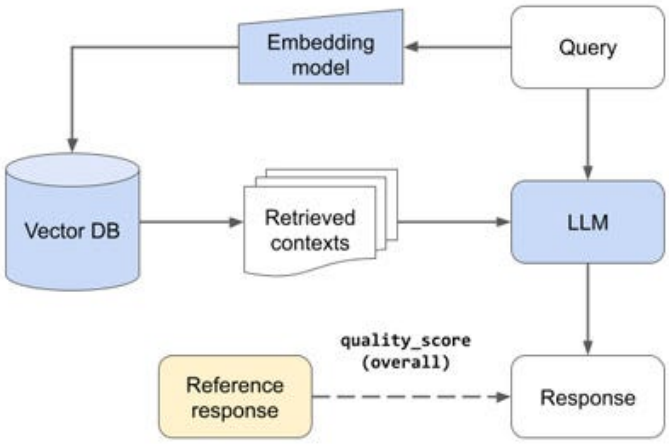
🙌 What did I like the most about this post?

While reading this guide, I learned about a powerful method to evaluate RAGs. It consists of building a synthetic evaluation dataset and then computing two metrics: one for evaluating the retrieval quality and another one for the generation quality. I applied this evaluation framework to my current RAG-based projects and this helped me quickly iterate on my experiments: this was useful for determining the

optimal parameters (chunk size, number of retrieved docs, embedding models, etc) instead of picking them randomly.

Learn more about RAG evaluation here.





Disadvantages of RAG

🔗 Link 👨 Author: **Kelvin Lu**

If you're new to RAGs, it's important to be aware of the limitations these systems hide. This will help you diagnose potential errors and manage unrealistic expectations.

In this timely post, the author provides a refreshing and brutally honest take on RAG's disadvantages. More specifically, he first discusses the semantic limitations of RAGs. Then, he emphasizes the lossy nature of their underlying processes, from chunking and embedding to retrieval and response generation. Interestingly, he also explains why certain types of search like multi-hop Q&A are not suited to RAGs.

With these limitations in mind, the author suggests that hybrid LLMs combined with external knowledge bases like graph databases are the way to achieve harder-to-reach goals.

🙌 **What did I like the most about this post?** Similar to any machine learning system, RAGs come with inherent limitations. Amid the excitement surrounding LLM applications, it becomes crucial to gain a clear understanding of their potential. This post aims to provide precisely that insight.

10 Ways to Improve the Performance of Retrieval Augmented Generation Systems

🔗 Link 👨 Author: **Matt Ambrogi**

After addressing some of RAG's shortcomings, it's useful to explore practical tips to improve their performance.

In this post, the author outlines 10 strategies to implement to bridge the gap between getting RAGs to work and getting them to work very well.

Here's a one-sentence summary for each tip:

1. Inspect and clean the data your RAG relies on to ground an answer
2. Explore different index types: embedding-based, key-word-based and hybrid
3. The chunk size is a hyper-parameter that must be tuned properly
4. Tweak the system prompt to customize your RAG, control its answers, and avoid hallucinations
5. Retrieval is not limited to the k nearest neighbors: you can leverage chunks' metadata for additional filtering
6. Create specialized indexes (each one for a specific task) and route your query to one of them
7. Rerank the retrieved documents based on relevance
8. Transform your query to help the LLM provide a better answer
9. Fine-tune the embedding model to align it further with your context
10. Debug LLMs by using developer tools like Rivet

Some of these tips are easy to implement directly while others are already integrated in LlamaIndex.

🙌 **What did I like the most about this post?** In a 9-minute read, the author condenses a wealth of insights that would typically take any ML practitioner weeks (if not months) of experimenting to uncover.

Gen AI: why does simple Retrieval Augmented Generation (RAG) not work for insurance?

🔗 Link 👷 Authors: **Damien Philippon, Christophe Bourguignat**

In this article, the authors, who are the founders of a French startup operating in the insurance sector, claim that RAG, on its own, doesn't fully satisfy consumer expectations or meet regulations.

They explain this through multiple reasons, the most interesting one being: the inability of RAGs to understand the structure of insurance documents.

In fact, in their default setting, RAGs cannot accurately answer questions like ***"Am I covered if a storm damages the trees in my garden?"*** because this question needs to be addressed from multiple angles, each one backed with its related policy in a different section of a +70-page contract.

A generic RAG would answer this question incorrectly because it retrieves a quick mention of the trees somewhere on page 12 (*'your trees and ground plantations' if they were planted at least two years before the disaster.') *without connecting this information with the other conditions of the policy that are on page 11, 66, 13, 52, etc.

More juicy details in the post. 👆

Answer (Generic RAG) ❌	Answer (Insurance RAG) ✅
yes, you are covered if they were planted at least two years before the disaster	1. Eligibility for Outdoor Installations Reinforcement: First and foremost, it's crucial to check whether the policyholder has subscribed to the "Outdoor Installations" reinforcement as mentioned in the contract. This subscription is necessary for the trees and plantations to be covered. 2. Age of Trees: According to the contract, to be eligible for compensation, the trees and plantations must have been planted at least two years before the incident. 3. Guarantee Limit: The contract stipulates a guarantee limit of $355 per tree. It's important to clarify to the policyholder that this amount represents the maximum sum that can be claimed per damaged or destroyed tree. 4. Size of the Land: It is mentioned that trees and plantations located on land larger than 5 hectares are not insured. Therefore, the size of the policyholder's land needs to be checked to ensure it is eligible for coverage. 5. Estimation and Compensation: If the policyholder has subscribed to the outdoor installations reinforcement, the compensation will be equal to the cost of replanting the destroyed trees, upon presentation of supporting documents. The costs associated with replanting, such as stump removal, logging, cutting, and removal of the damaged trees, are included in the replanting cost. Compensation will be paid progressively as replanting takes place, which must be carried out within **two years following the incident.** 6. Public Subsidies: If the policyholder receives a subsidy from Public Authorities, like the National Forestry Fund or other public bodies, on the occasion of the incident, this subsidy will be deducted from the compensation amount.

**👐 What did I like the most about this post? **It's uncommon to get tangible and honest feedback from companies, especially when evaluating over-hyped technologies. This post provides a fresh perspective that anyone applying RAGs can benefit from.

Azure Cognitive Search: Outperforming vector search with hybrid retrieval and ranking capabilities

🔗 Link 👨‍💻 Author: Alec Berntson

This post comes from Microsoft's tech blog and presents a new method to improve retrieval in RAG applications. This method uses **hybrid retrieval + semantic ranking**.

The author claims that it's the most effective approach for improved relevance out of the box.

👉 **Hybrid retrieval (L1)**: Performs both keyword and vector retrieval and applies a fusion step to select the best results from each technique.

👉 **Semantic ranking** is performed by Azure Cognitive Search's L2 ranker which utilizes multi-lingual, deep learning models adapted from Microsoft Bing. The Semantic ranker can rank the top 50 results from the L1.

Retrieval comparison using Azure Cognitive Search in various retrieval modes on customer and academic benchmarks. See §6.1 How we generated the numbers in this post and §6.2 Search and Dataset configuration for Table 1 for the setup and measurement details.

Search Configuration	**Customer datasets** *[NDCG@3]*	**Beir** *[NDCG@10]*	**Multilingual Academic (MIRACL)** *[NDCG@10]*
Keyword	40.6	40.6	49.6
Vector (Ada-002)	43.8	45.0	58.3
Hybrid (Keyword + Vector)	48.4	48.4	58.8
Hybrid + Semantic ranker	**60.1**	**50.0**	**72.0**

**👐 What did I like the most about this post? **Instead of applying the same retrieval method over and over, this post provides a new approach that beats the existing state-of-the-art.

Thanks for reading 🙏

This is a new format I'm experimenting with on Medium. Please tell me what you think. (was it helpful? boring? too short?)

Anywho, I hope you learned one thing or two and explored the articles I shared with you.

Have a good day and until next time! 👋

Ahmed