RAG or not to RAG

Why Building a Bespoke Retrieval Augmented Generation (RAG) System is Essential

4 min read · 21 hours ago

Yogesh Haribhau Kulkarni (PhD)

(Source: pixabay)

Retrieval Augmented Generation (RAG) based chatbots are increasingly becoming the go-to solution for enterprises looking to deliver accurate, context-aware, and real-time answers. If you've been following the hype around large language models (LLMs), it's tempting to assume that bigger models and larger context windows will solve every problem. But that's far from reality. Even as models scale, RAG remains indispensable, like an open-book exam where the textbook is vast, ever-changing, and are too many to go through.

Why RAG is Non-Negotiable, No Matter How Powerful LLMs Get

At first glance, it's easy to think that once an LLM is powerful enough, there's no need to augment it with external data retrieval. However, the core challenges remain the same and in some cases, they become even more pressing. Here's why:

1. **Access to Private and Specialized Data**
   No matter how expansive an LLM's training corpus is, it will never contain your organization's private data, customer interactions, proprietary research, internal reports, or workflows. Building and curating a fine-tuned model is prohibitively expensive, both in data preparation and computational resources. RAG offers a way to connect the LLM with these private datasets without expensive retraining.

2. **Streaming and Near-Real-Time Information**
   Many business scenarios demand answers based on the latest data, financial reports, supply chain metrics, product availability, or regulatory updates. These datasets evolve continuously, and integrating them into a static model's knowledge base isn't feasible. A RAG system allows real-time or near-real-time querying against updated sources.

3. **Grounding and Reducing Hallucinations**
   LLMs are notorious for hallucinating, confidently fabricating answers when the data isn't sufficient or context is missing. RAG enables grounding by pulling in authoritative references directly from your data sources, improving trustworthiness and reliability.

4. **Cost, Speed, and Efficiency**
   Feeding an entire dataset into an LLM isn't practical. Instead, RAG selectively retrieves relevant chunks from a larger corpus, drastically reducing token consumption and inference time. This approach delivers faster, cheaper, and more efficient responses.

But Building a RAG System is More Than Plug-and-Play

Having implemented RAG at scale for complex enterprise solutions, I can confidently say, production-level RAG is far more than uploading PDFs to a chatbot interface and expecting it to perform. If you're skeptical, upload a finance report, especially one with tables, and ask a question about a specific entry. You'll quickly realize that understanding structure, relationships, and context is non-trivial.

Enterprises can't rely on off-the-shelf solutions because their data is too diverse, noisy, and structured in ways that require custom handling. You HAVE to build bespoke, home-grown RAG systems to meet specific business needs. It's not glamorous work, it's about understanding data deeply and engineering solutions that work at scale.

The Key Pillars of Enterprise-Grade RAG

1. **Parsing and Chunking**
   Documents, reports, logs, and transcripts must be intelligently broken down into meaningful segments without losing context. Poor chunking leads to irrelevant or incomplete answers.

2. **Multimodal Partitioning**
   Data isn't always text-based. Images, videos, tables, graphs, and even audio streams need to be preprocessed and indexed differently. A medical dataset may include MRI scans, lab reports, and structured patient records, all requiring tailored strategies.

3. **Delegation and Workflow Orchestration**
   Not every query should hit the same pipeline. Some require complex reasoning, while others can be answered with a single document lookup. Efficient systems dynamically delegate tasks based on query type, user context, and available resources.

4. **Evaluation and Feedback Loops**
   Continuous monitoring of RAG outputs is essential. Systems must track hallucinations, relevance, and user satisfaction, feeding insights back into data pipelines for iterative improvement.

Exploring New Frontiers: Geometry and Graph Structures

RAG isn't limited to linear text searches. The index-and-search paradigm can extend into unexplored modalities such as geometry and graphs. Imagine querying across architectural blueprints, floor plans, or 3D models. A real estate platform could

instantly pull out similar layouts or highlight safety compliance issues across structures.

Graphs, too, open doors to querying relationships, supply chains, organizational hierarchies, and scientific networks. With proper embeddings and search techniques, graph-based RAG systems can deliver insights that traditional text search simply cannot.

A Final Word

Building a RAG system isn't about mimicking open-source tools or replicating tutorials. It's about engineering solutions tailored to real-world problems, grounded, efficient, and scalable. Enterprises that invest in bespoke RAG pipelines unlock the power to leverage their data securely and intelligently.

You can't outsource this problem and expect it to work out of the box. You need to get your hands dirty, chunking, parsing, partitioning, indexing, and constantly refining your approach. Once you've walked the path, you'll see why RAG is not a luxury, but a necessity.

If you're ready to build something that goes beyond demos and delivers real-world value, the time to start is now.

Retrieval Augmented Gen | Artificial Intelligence | Chatbots | Enterprise Technology | Consultant

Following

Published in Technology Hits

3.7K followers · Last published 21 hours ago

We cover important, high-impact, informative, engaging stories on all aspects of technology. Subscribe to our content marketing strategy newsletter: https://drmehmetyildiz.substack.com/ Apply via: https://digitialmehmet.com/contact External: https://illumination-curated.com

Edit profile

Written by Yogesh Haribhau Kulkarni (PhD)

1.8K followers · 2.1K following

PhD in Geometric Modeling | Google Developer Expert (Machine Learning) | Top Writer 3x (Medium) | More at https://www.linkedin.com/in/yogeshkulkarni/

No responses yet

Yogesh Haribhau Kulkarni (PhD)

What are your thoughts?

More from Yogesh Haribhau Kulkarni (PhD) and Technology Hits

In ILLUMINATION Videos and Podcasts by Yogesh Haribhau Kulkarni (PhD)

Summary of “How To BRAINWASH Yourself For Success & Destroy NEGATIVE THOUGHTS!”

YouTube channel: Tom Bilyeu

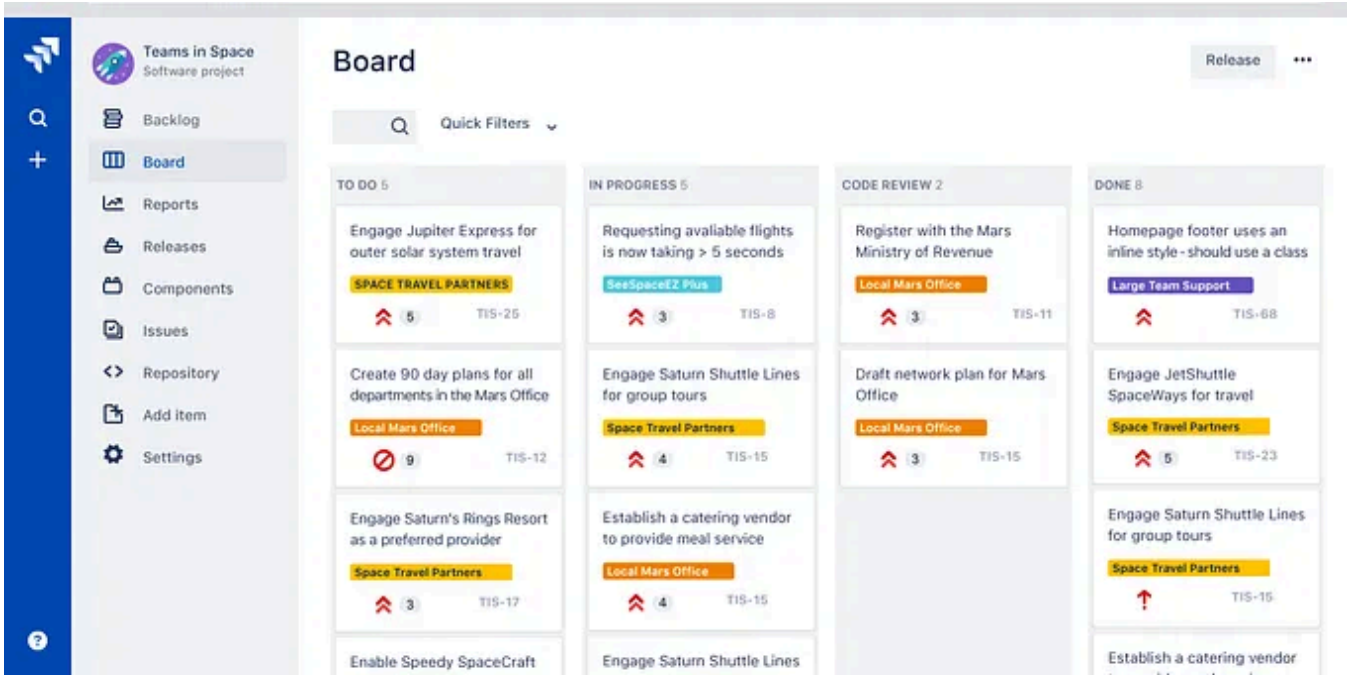
Dec 21, 2022 268 1

In Technology Hits by Brian Iselin

Why Tesla Keeps Slamming Into Fire Trucks While Waymo Doesn’t

Semantic mapping is the invisible backbone of safe autonomy. Without it, “Full Self-Driving” is marketing, not engineering.

Aug 22 930 17



In Technology Hits by Garima Srivastava | Product Mindset

Most Product Managers Are Actually Project Managers in Disguise

Why 70% of Product Managers Are Actually Doing the Wrong Job

Aug 11 211 5

In ILLUMINATION Videos and Podcasts by Yogesh Haribhau Kulkarni (PhD)

The Equation of True Happiness

Based on talks by Arthur C Brooks

Sep 17, 2023 164 1

See all from Yogesh Haribhau Kulkarni (PhD)

See all from Technology Hits

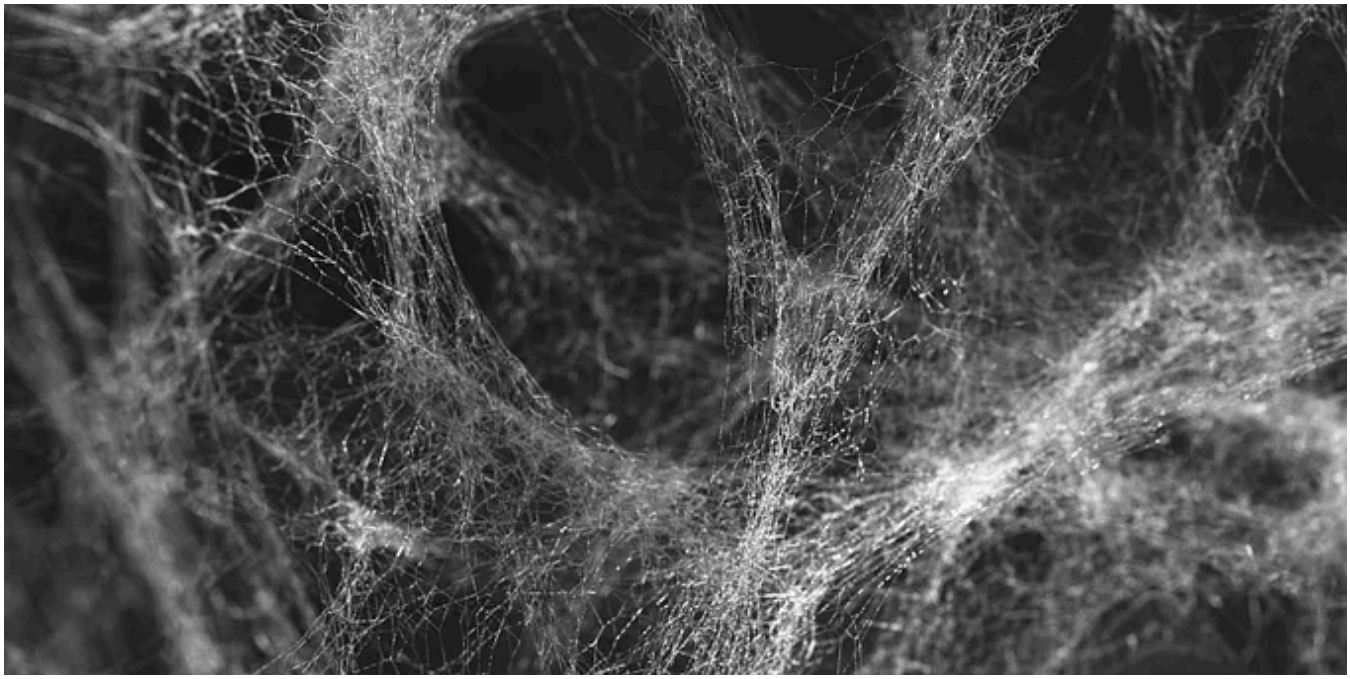
Recommended from Medium

In GoPenAI by John Wong

Part 6: Power up RAG Chatbot with Hybrid Search and Filtering

Boost your RAG chatbot. Learn hybrid search, metadata filtering, and LLM-powered query rewriting for smarter, more precise retrieval.

Aug 19 1

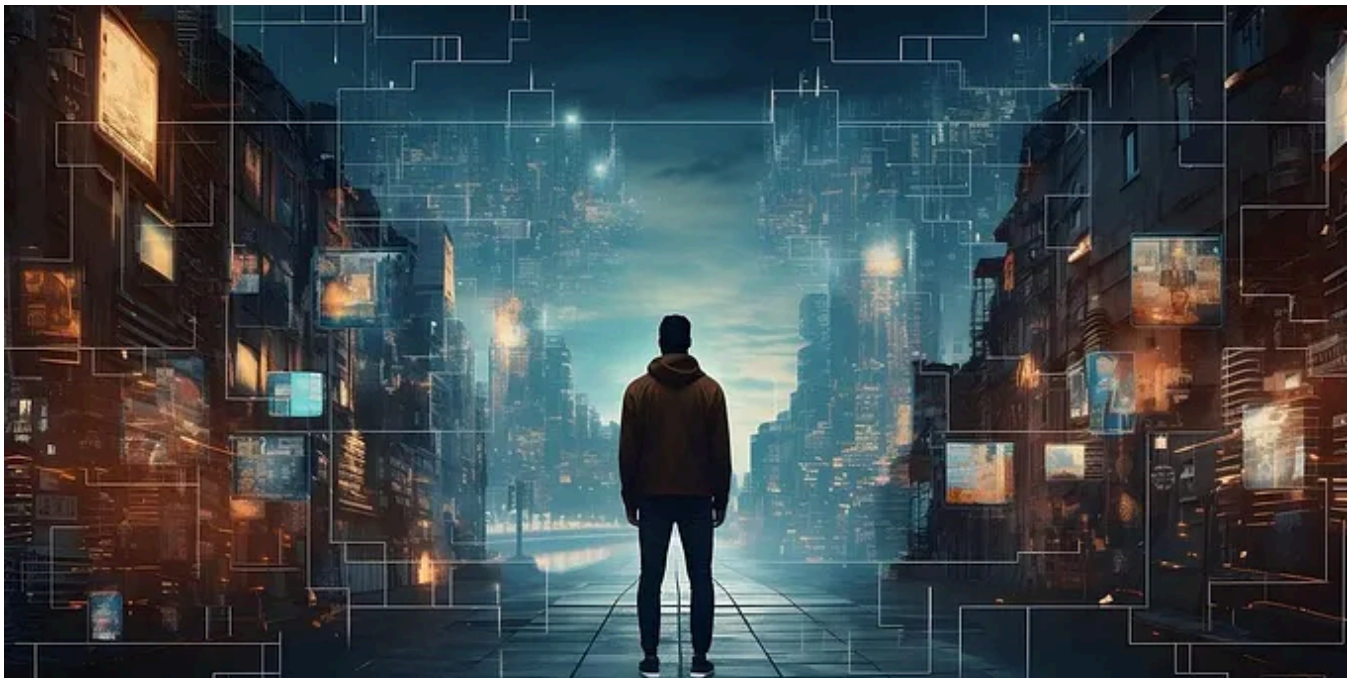


Bahadır AKDEMİR

RAG Notes 3: Extending RAG through Knowledge Graphs

A base guide to building retrieval-augmented generation systems that leverage graph structures for better reasoning and attribution

Sep 2 6

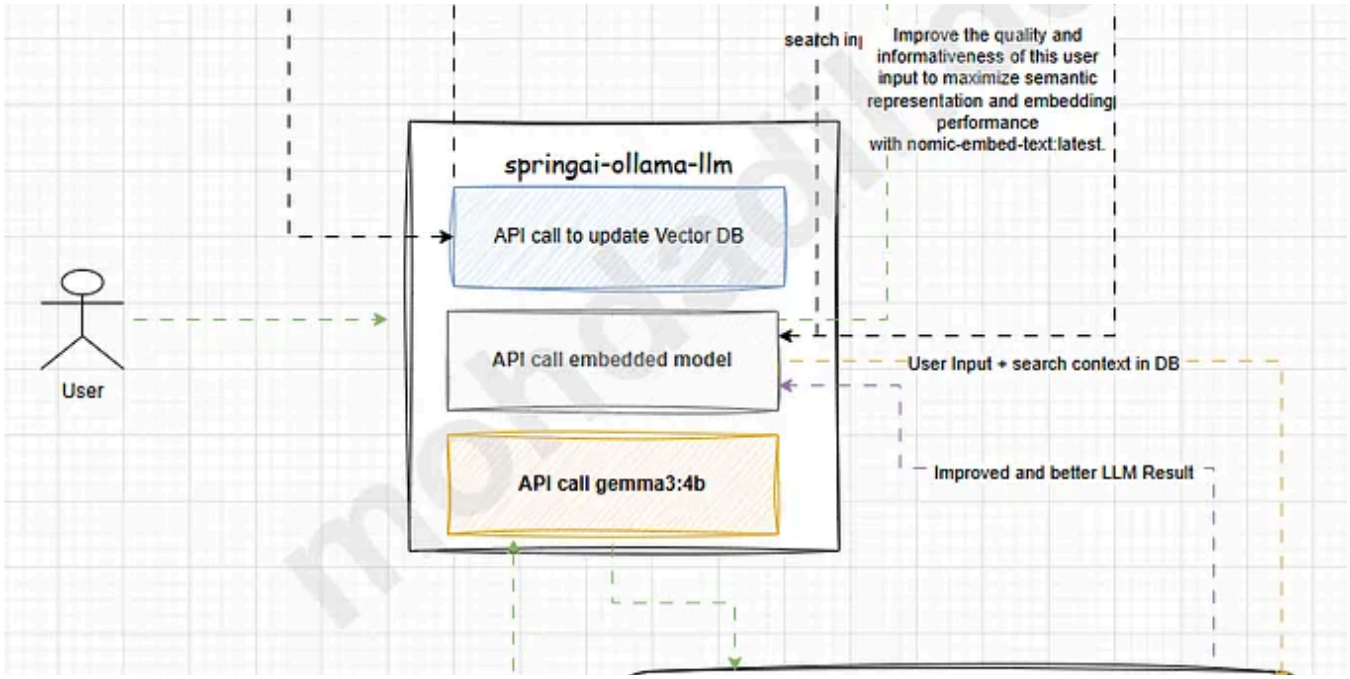


Paradigma Digital

Deep Learning about Spring AI: RAG, Embeddings and Vector Databases

As we've already discussed, LLMs are models pre-trained up to a certain date and with public data (it's assumed they don't have access to a…

Sep 2 51

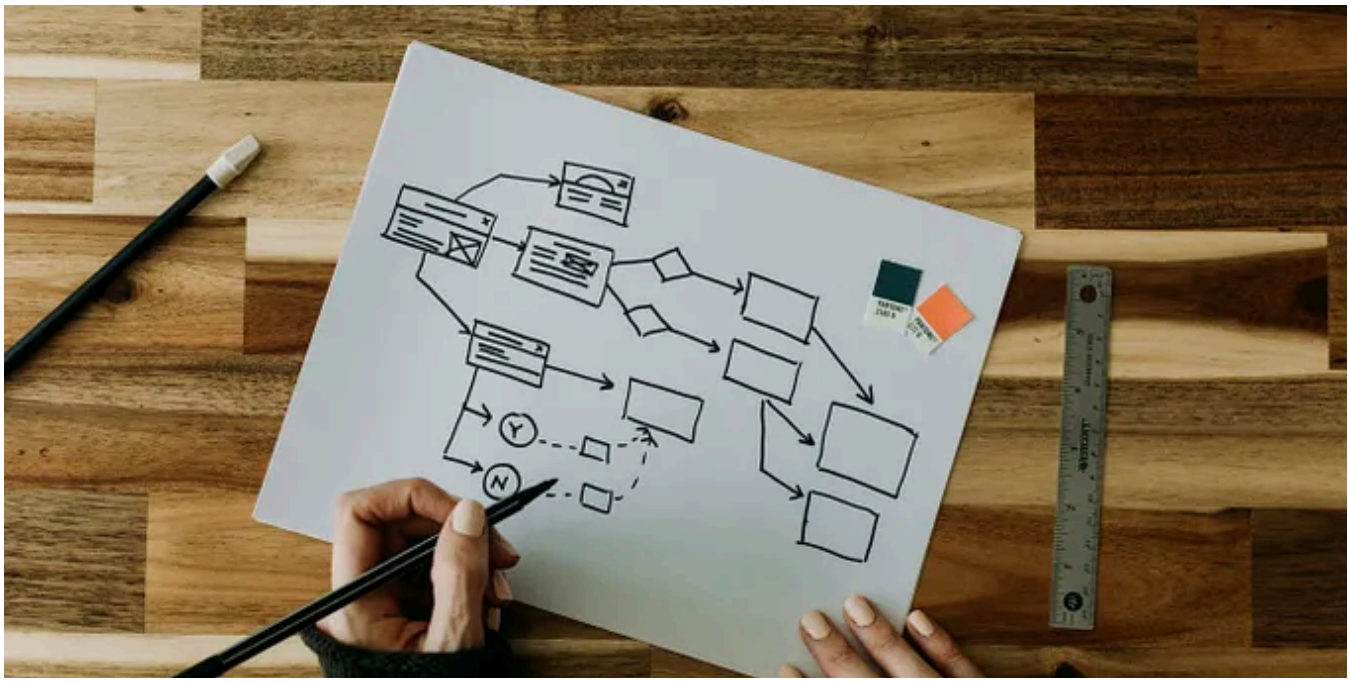


In Towards AI by Adil

SpringAI Retrieval Augmented Generation (RAG) With PgVector Part 2

Welcome back, folks! 👋 This is the second part of our series on using SpringAI RAG with an embedded Ollama model.

Sep 2



In Stackademic by Jonathan Cardoz

n8n: workflow building made easy

Over the course of the last two months, I have been working on learning about and building chatbots and ‘agents’, in order to better…

Jun 16 3

In Artificial Intelligence in Plain English by Shanmuga priya

15 Best AI Website Builders in 2025 (Ranked & Compared)

Looking for the best AI website builder? Discover 2025 rankings of the top 15 tools with features, pricing & reviews.

2d ago 252 4

See more recommendations