deepset

RAPIDLY BUILD PRODUCTION-READY RETRIEVAL AUGMENTED GENERATION (RAG) APPS

WHAT IS RAG?

Large Language Models (LLMs) are powerful and trained on vast amounts of data, but they can't access your organization's most valuable asset—your proprietary data and expertise. Retrieval Augmented Generation (RAG) solves this challenge, seamlessly connecting your enterprise knowledge to your Generative AI models. With RAG, you can combine the power of LLMs with your business data, delivering accurate and contextualized responses while maintaining data security and cost efficiency.

REAL-WORLD IMPACT WITH DEEPSET

YPULSE

YPulse, a global youth insights firm, created a customer facing GenAI app leveraging its vast database of market research, delivering an enhanced user experience and accelerated value.

+25% increase in customer engagement

5X return on investment within a year

<4 months to production

Learn More About YPulse's Journey ↗

HOW RAG WORKS

When implemented correctly, RAG creates an intelligent application that:

- Builds a knowledge base of your data, allowing the LLM to base its answers on fact-checked and up-to-date information.
- Processes information from multiple data sources of varying formats and complexity.
- Scales efficiently to handle enterprise-level data volumes.

RAG improves usability, reduces research time, and tailors information to the user's request.

BENEFITS OF RAG

01 | **Enhanced Accuracy**

Ensure AI responses are grounded in verified information and mitigate hallucinations.

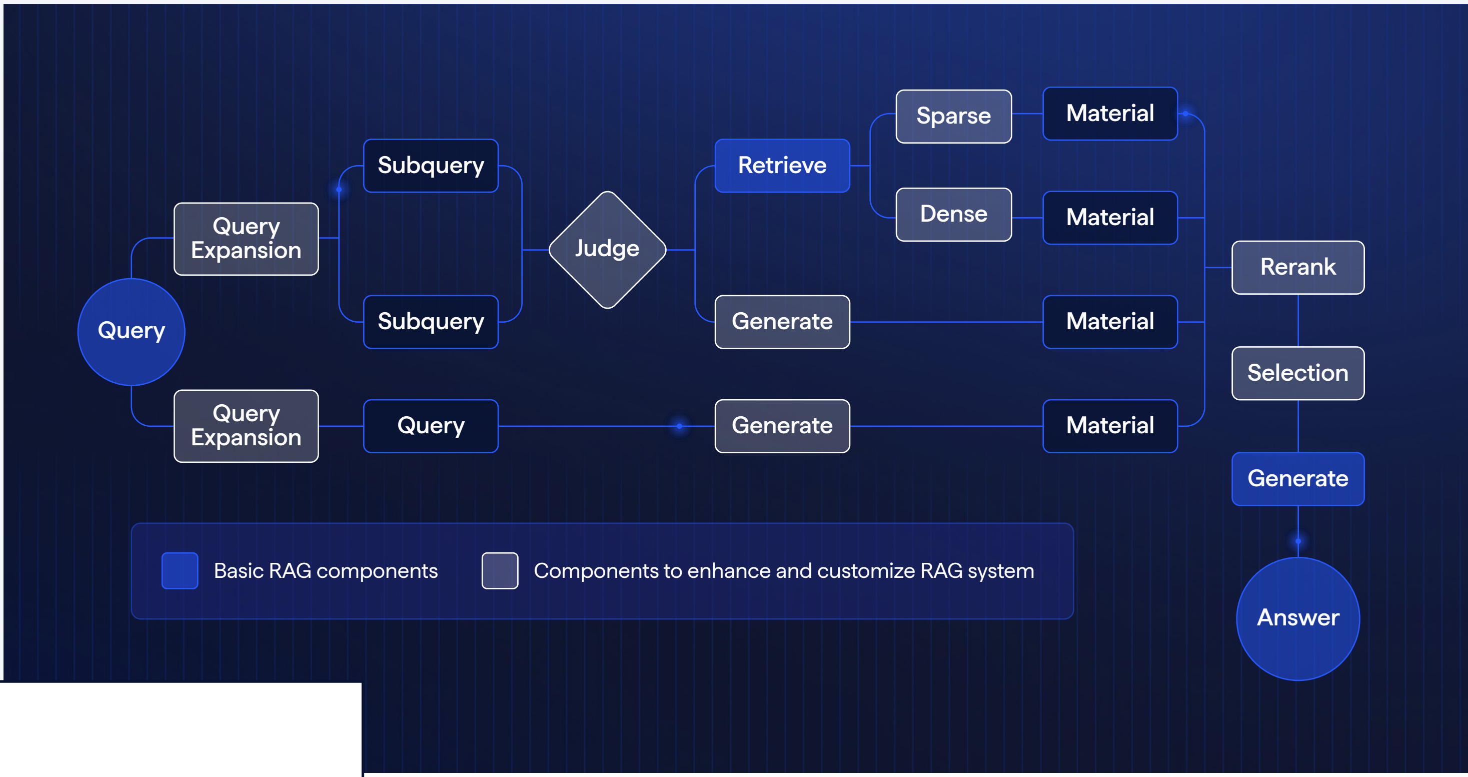
02 | **Cost Optimization**

Smart document retrieval minimizes the need for expensive model training and allows you to change models at any time.

03 | **Knowledge Efficiency**

Turn your LLM into a highly skilled domain expert who can provide reliable information within seconds.

CUSTOMIZING PRODUCTION-GRADE RAG: SAMPLE ARCHITECTURE



WHY DEEPSET FOR RAG?

Technical Expertise

Work with seasoned AI Engineers to customize your RAG application to your business and unlock the potential of your organizational knowledge.

Comprehensive Platform

An end-to-end AI development and deployment platform with built-in security, governance, and compliance.

Rapid Prototyping

Leverage a vast library of pre-built RAG templates to jumpstart your development.

Maximize Adoption

Easily collect end-user feedback to validate accuracy of your RAG application and ensure usability.

Model Flexibility

Choose and switch between any commercial or open-source model without vendor lock-in.

Production-Ready

One-click deployment with automatic scaling.

Multimodal Support

Process any format including text, tables, PDFs, images, audio, and video.

Learn more and request a demo today:

deepset.ai ↗