Connecting AI models to your business data

Functional Report

Drive Productivity

Value

- Amplify your internal teams' productivity with an AI assistant that knows their data, processes, workflows, inbox, etc

- Automate back-office processing

- Enable decision makers to retrieve the info & insights they need immediately just by asking

AI That Knows Your

- Company SOPS
- Team meetings
- Training materials
- Project documentation

- Quality standards
- Policies
- Contracts

AI That Knows Your Enterprise

How can organizations extend AI models or large language models (LLMs) with their internal company data?

Retrieval Augmented Generation (RAG) is a concept in which leading AI models are augmented with internal data relevant to a specific use case. Take a scenario where you want to deploy an AI assistant that understands all your business's standard operating procedures. Leading AI models have not been trained on your company's SOPs and would not be able to service your internal team out of the box. RAG enables you to supplement an AI model with your own data.
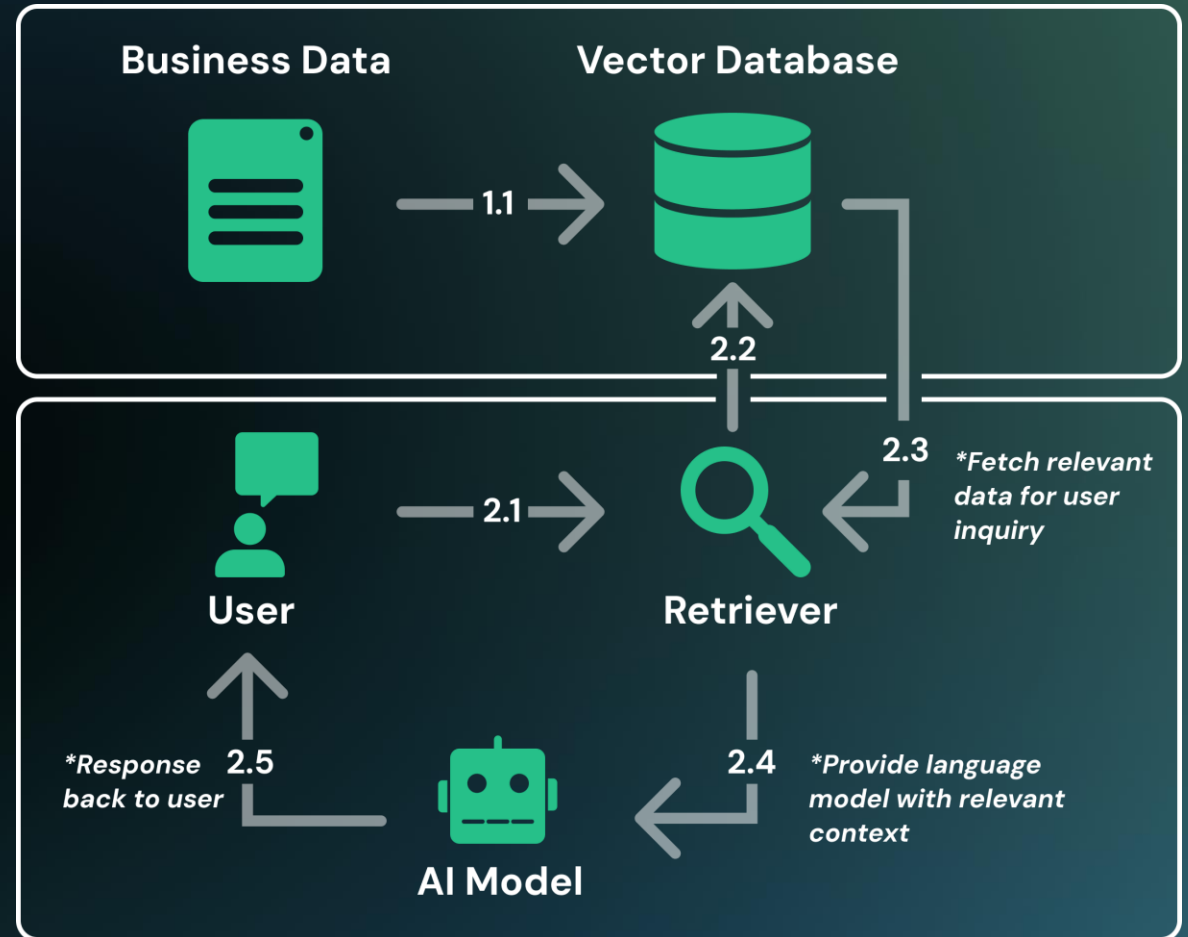
RAG 101

1. Training Data/Preparation

- Study your underlying data
- Develop the appropriate orchestration and ingestion pipeline to process your data
- Create vector embeddings of your data (searchable indexes to the respective data)
- Store vector embeddings in a vector database

2. Retrieval

- User submits inquiry
- Create a vector embedding of the user inquiry
- Search vector database for related vectors
- Pass relevant context to an AI model (LLM)
- LLM responds back to user with appropriate context

1
Business Data → 1.1 → Vector Database

2
User → 2.1 → Retriever
2.2 → Vector Database
2.3 *Fetch relevant data for user inquiry
2.4 *Provide language model with relevant context → AI Model
2.5 *Response back to user → User

* This is only a foundational diagram. Production-ready systems involve a handful of more advanced components and techniques

RAG based systems

4

Moving to Production

eranova

To make a RAG-based system production ready, there is another layer of advanced concepts that must be considered. To name a few…

- Rerankers
- Advanced chunking strategies
- Vision model processing
- Prompt-restructuring

- LLM-routing
- Response moderation
- Custom tool calls (AI Agent Actions)
- Intent classifiers

Common Hurdles Faced

- Underlying training data
- Complex data types
- Advanced prompting requirements

- Dealing with structured vs unstructured data
- Circumventing hallucination
- Properly chunking different docs

Reach Out

eranova

If you want to:

- Avoid complex architectures for one-off projects

- Circumvent overhead over setup, maintenance, and support

- Accelerate beyond hurdles

- Improve existing RAG-based applications

- Expedite standing up production AI models connected to enterprise data and workflows

If you want continuous insights into more advanced-RAG techniques, follow our page or reach out:

Linked in | **Contact us**