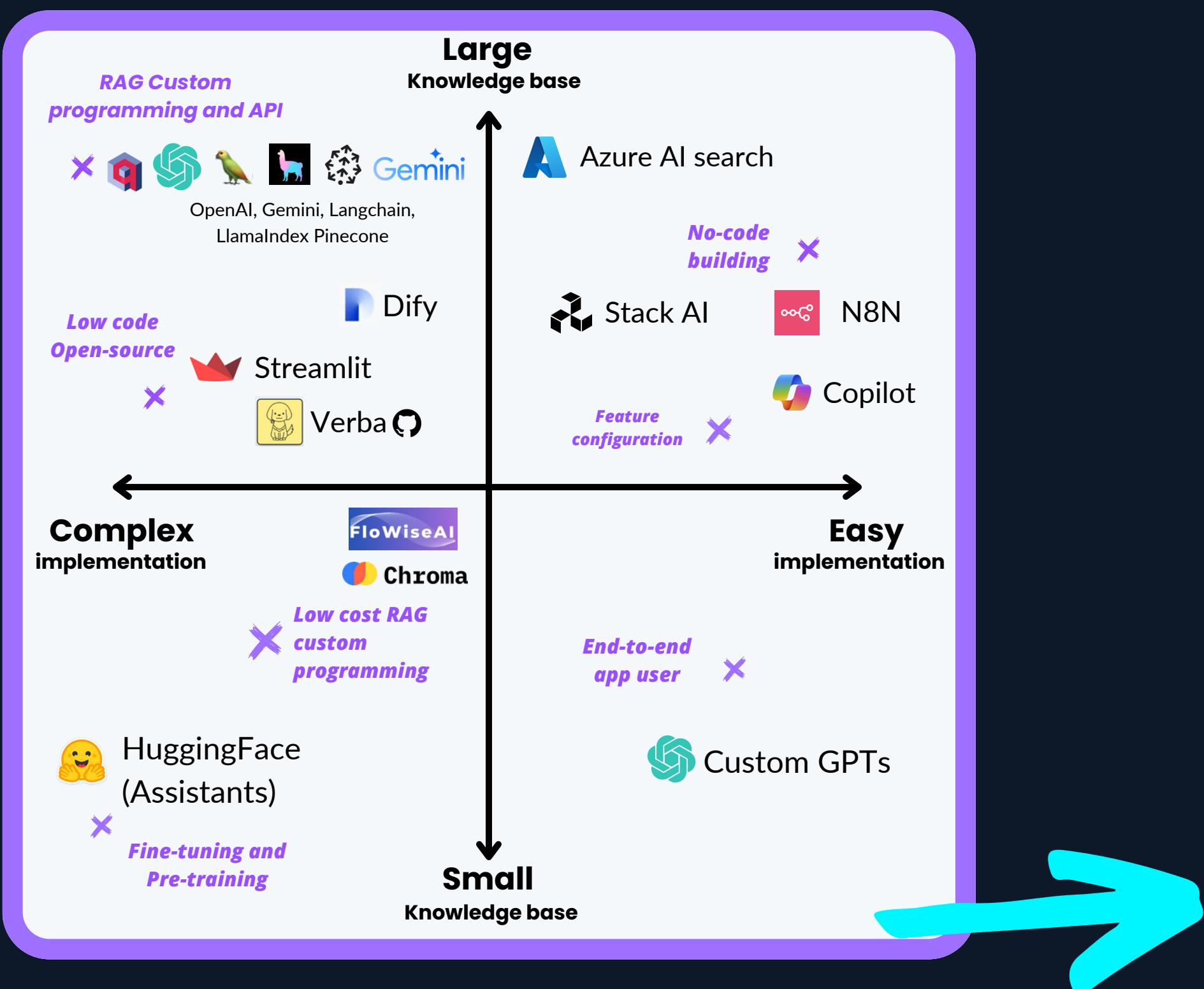




Get started with RAG

BEGINNER TO ADVANCED – COMPLETE GUIDE



**This guide will
help you choose
the best RAG
solution !**



Joanna Stoffregen
@JoannaStoffregen

Quick recap : What is RAG ?

RAG consists of augmenting an LLM (ie: ChatGPT) with your own data



User

R

A

G

Retrieval

Augmented

Generation

[Prompt]

[Prompt + context]

[LLM Response]

[Context]



Websites, docs, databases...
(proprietary and non proprietary)



Joanna Stoffregen
@JoannaStoffregen

Today you will discover

RAG for beginners : User mindset

Difficulty

RAG for semi-pros : no code

RAG for pros : low code

RAG for developers : code

RAG for experts : production-ready



Joanna Stoffregen
@JoannaStoffregen

Very Easy

RAG for beginners: User mindset - plug & play

Tools

Custom GPT, AskYourPDF,
OpenAI ADA, Dante AI



*Upload
data*



1



Tool

*Start
chatting*



2

Pros Easy, no-code, cost effective

Cons Limited size of docs and formats, no personalized UI, scaling limitations

WHO Individual use and occasional needs

Price \$20/month



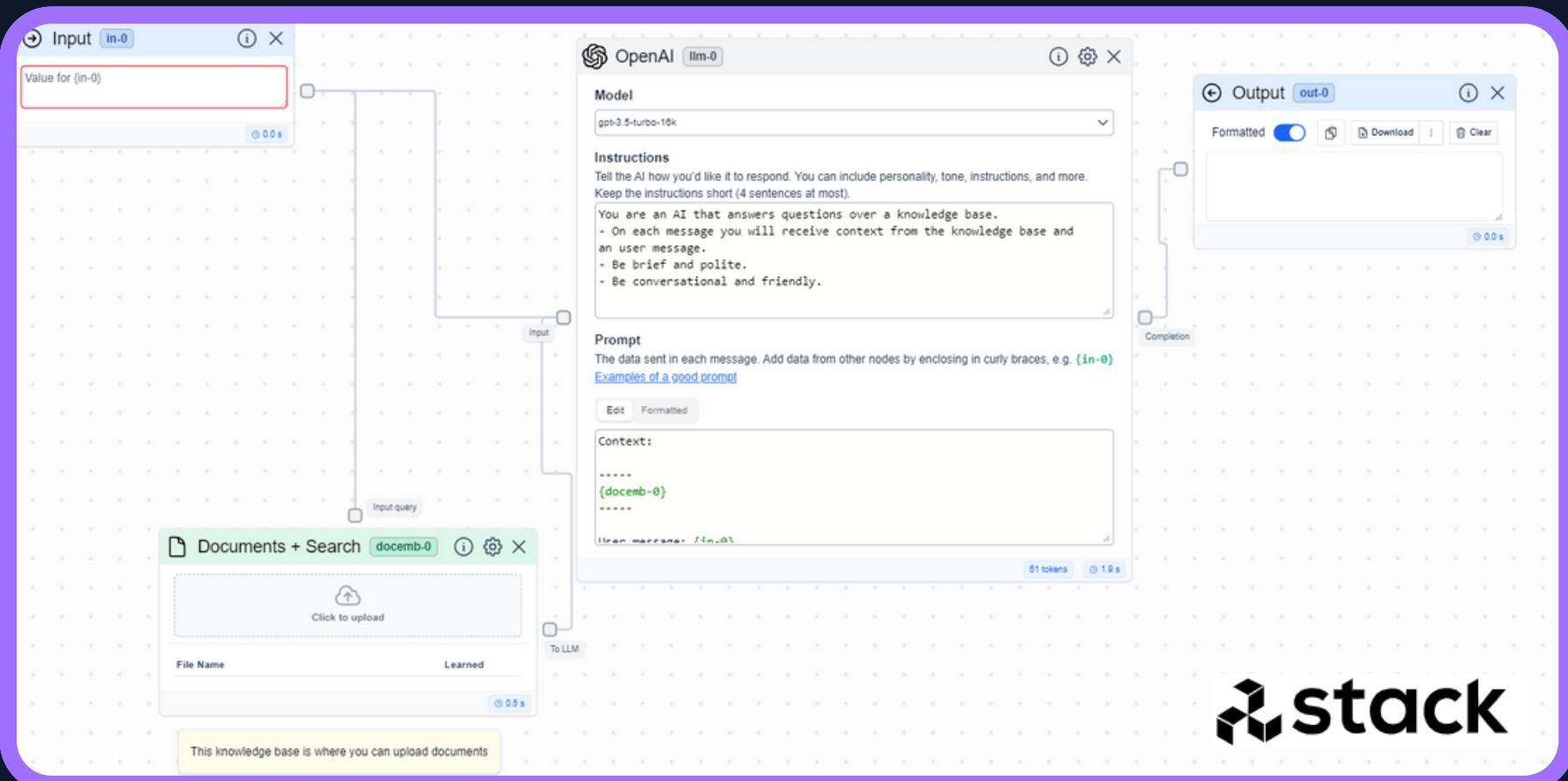
Joanna Stoffregen
@JoannaStoffregen

Easy

RAG for semi-pros: No-code closed source

Tools

Stack AI, AirOps, N8N



Pros Easy, no-code, end-to-end solution (chat interface, embeddings, vector database, deployment)

Cons High costs for individuals and high costs for scaling, UI customization limitations

WHO Ideal for startups, SMEs, solopreneurs who want an end-to-end solution with a **medium budget**

Price Pro plan \$200/m ; Team plan 900\$/month



Joanna Stoffregen
@JoannaStoffregen

RAG for pros: Low-code open source

Tools

Verba AI, Dify, Flowise



The screenshot shows the Verba AI interface. At the top, there's a navigation bar with a yellow icon of a person, the text "Verba", and "The Golden RAGtriever". Below the navigation bar are buttons for "+ Add Document...", "ADAEembedder", "WindowRetrie...", and "GPT4Generat...". A "Blog" button is highlighted in yellow. Below these are several green cards with numerical values: 65, 63, 61, 75, and 73, each associated with a category like "blog-graphql-api-design" or "developers-weaviate".

The main area features a "RAGtriever Chat" window. It displays a message from "Weaviate" about what it is and why it needs an API. Below this, another message discusses GraphQL. At the bottom of the chat window, there's a text input field with the placeholder "What is a vector database?".

At the very bottom of the interface, there are three small buttons: "Search", "Documents", and "Status".

Pros Low code, free, perfect for an MVP, possibility to run locally (data privacy)

Cons UI customization and cloud deployment requires technical knowledge, not ready for production

WHO Ideal for startups, SMEs, solopreneurs who want an MVP solution.

Price Free



Joanna Stoffregen
@JoannaStoffregen

medium hard

RAG for developers: Code - Do it yourself

Tools

OpenAI API, Gemini API,
Ollama



OpenAI
Assistant API

Knowledge
retrieval function
+
Code
interpreter

Pros Embedding and
vector store included

Cons Low UI customization

WHO Startups, SMEs,
solopreneurs

Price \$0.20 /GB/assistant



Google
Gemini API

Gemini
LangChain
Chroma
Streamlit

Pros Free, Google drive
easy integration

Cons Coding skills
required

WHO Startups, SMEs,
solopreneurs

Price Free



Ollama + Mistral



MISTRAL
AI_



LangChain



Ollama



Chroma

Pros Free, open source,
full data privacy

Cons Coding skills
required

WHO Businesses with high
confidential data

Price Free

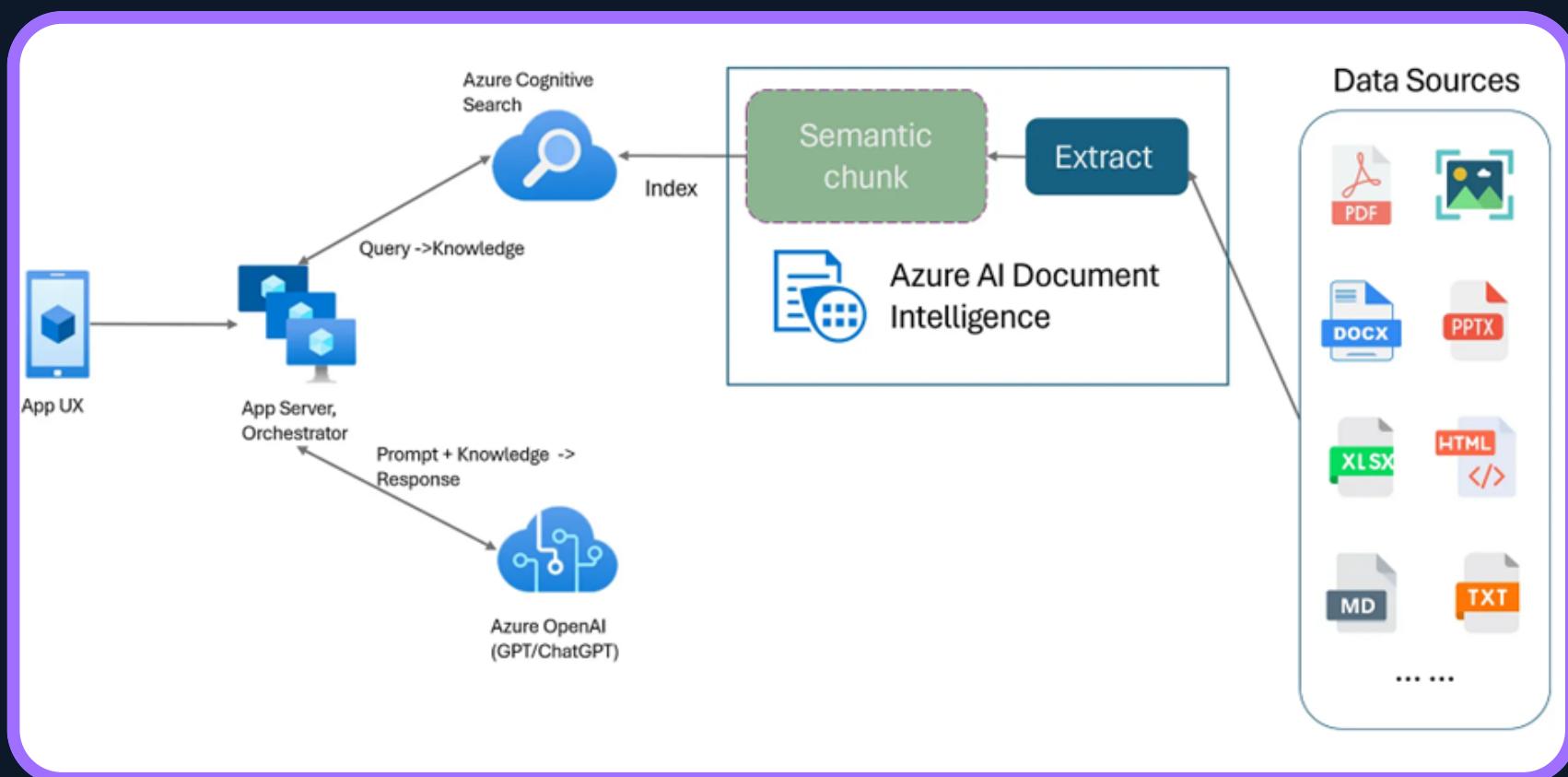


Joanna Stoffregen
@JoannaStoffregen

RAG for experts: Azure end-to-end solution

Tools

Azure AI search



Pros End-to-end solution, low code, can be combined with OCR, variety of formats available

Cons Microsoft enterprise account required, high scaling costs for small businesses and individuals

WHO Ideal for companies who are already using Azure and Microsoft suite

Price \$245/month/user (25 GB) \$5600/month/user (2 TB)



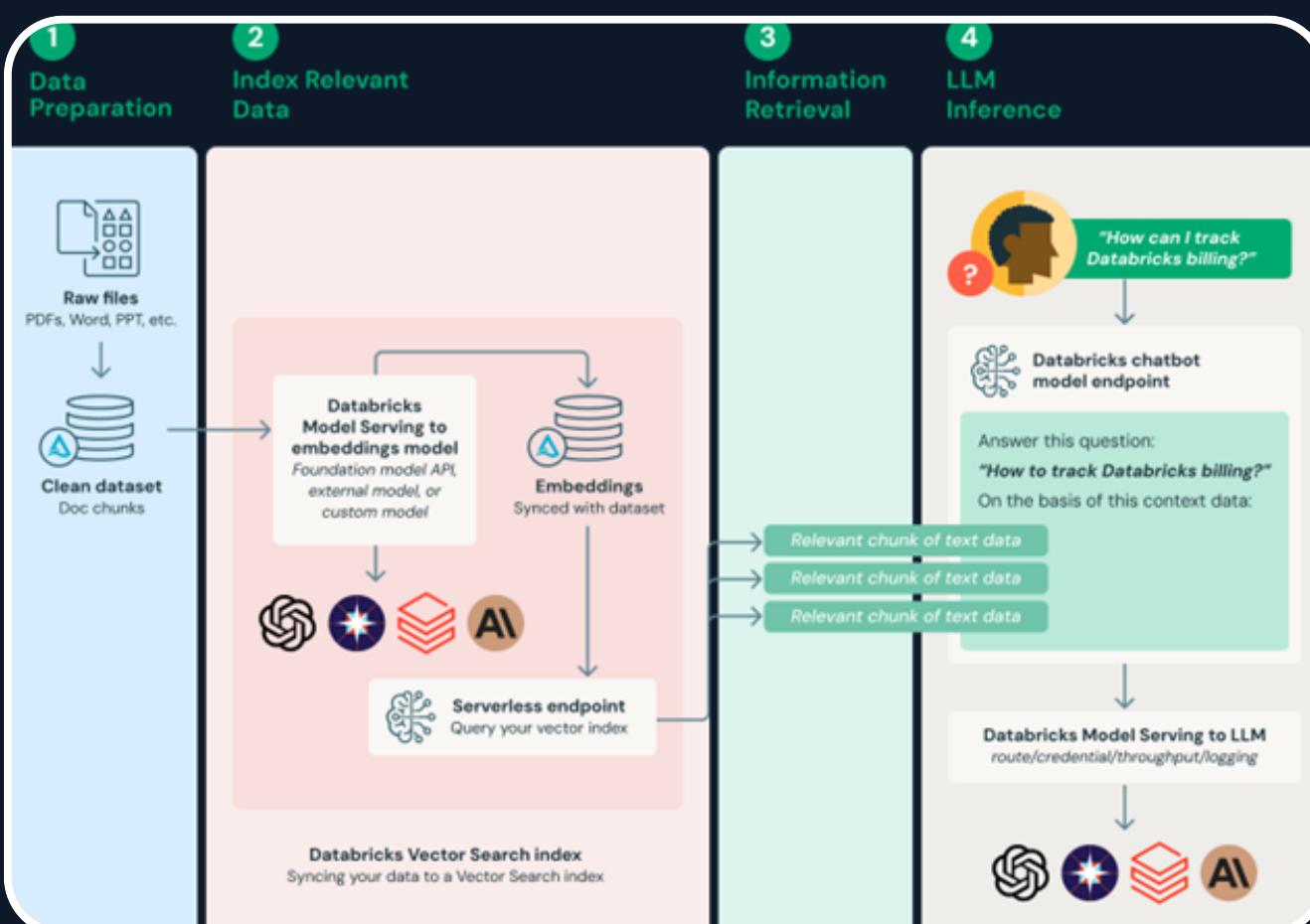
Joanna Stoffregen
@JoannaStoffregen

Very hard

RAG for experts: Databricks – for production

Tools

Databricks



Pros Wide choice of cloud providers, End-to-end solution, serverless endpoint , variety of formats available

Cons Requires deep technical skills, high scaling costs and complexity for small and medium businesses

WHO Ideal for very large companies (Fortune 500)

Price Pay-as-you go (complex)



Joanna Stoffregen
@JoannaStoffregen

Hey, I'm Joanna from Labsbit.ai

An AI/ML Product Development
& Automation Company

Have trouble with RAG
implementation ?

Reach out to learn more about
how we can help.



Joanna Stoffregen
@JoannaStoffregen

