Advanced Agentic AI RAG Chatbot Assistants for Healthcare and Life Sciences

Ernest Bonat, Ph.D.
Lead Data Science Faculty
Senior GenAI Engineer
ernest.bonat@biogenaisolutions.com

About Ernest?

- Came from Hillsboro, Oregon 5 months ago. Live in North Port now.
- 29 years of experience as a Senior Software Engineer.
- 8 years of experience as a Senior Machine Learning Engineer.
- Develop and teach online AI courses for 3 universities.
- Open a start-up LLC company BioGenAI Solutions in Florida – developing GenAI Assistants for Healthcare and Live Sciences.
- Owner/Organized of Portland Python User Group and Hillsboro Python Machine Learning Meetup in Oregon.

What is Classic RAG Assistant?

A classic RAG assistant is an AI system that answers questions (Q&A) using **Retrieval-Augmented Generation** (RAG) pipeline algorithms. Is an AI chatbot that:

1. The system **retrieves documents or facts** from a knowledge base (**unstructured data**: PDF, JSON, XML, TXT, Words, HTML, etc.), **structured data**: aggregated database engines including MySQL, Oracle, MongoDB, SQL Server, PostgreSQL, etc.)
2. The retrieved information **augments**, or improves, the LLM's answer
3. The LLM **generates** the final answer using the retrieved information.

- Searches your documents files, database, or knowledge base.
- Retrieves the most relevant information.

- Uses an LLM to generate an answer.
- Grounds the answer in the retrieved data.

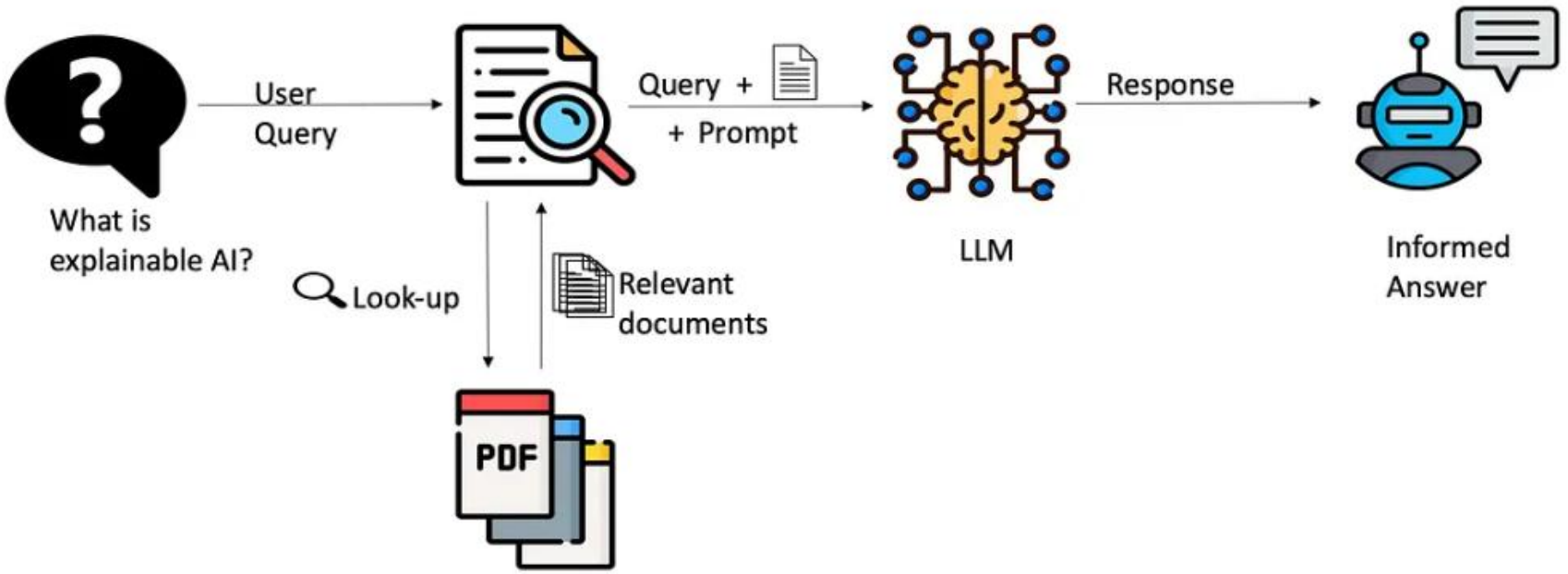

Figure 1. How Retrieval-Augmented Generation Enhances LLMs with Real-World Context

“A classic RAG assistant is an AI chatbot that retrieves real information from your knowledge base and uses an LLM to generate accurate, grounded responses.”

A Classic RAG became an AI Chatbot Assistant in late 2022–2023.

What is Agentic RAG Assistant?

RAG + AI Agent Workflow + Tools + Decision-Making.

It doesn't just *retrieve information and answer questions* — it can **think, plan, choose tools, execute multi-step actions, and solve complex tasks autonomously**.

An Agentic RAG Assistant is an AI system that:

1. Retrieves information from your documents (RAG).
2. Uses an LLM to reason about the task.
3. Plans multiple steps to solve the problem.
4. Chooses the right tools (e.g., text-to-SQL, search engine, calculators)
5. Executes each step autonomously.

6.Produces a final answer with citations.

Agentic RAG Assistants appear for the first time in 2024.

Why "Agentic"?

The word **agentic** means **acting like an agent**:

- It decides what to do next
- It takes actions
- It uses tools
- It evaluates results
- It corrects itself

*"An Agentic RAG Assistant is an AI system that **retrieves information, reasons step-by-step, chooses tools, executes actions, and autonomously solves complex tasks**—not just answer questions."*

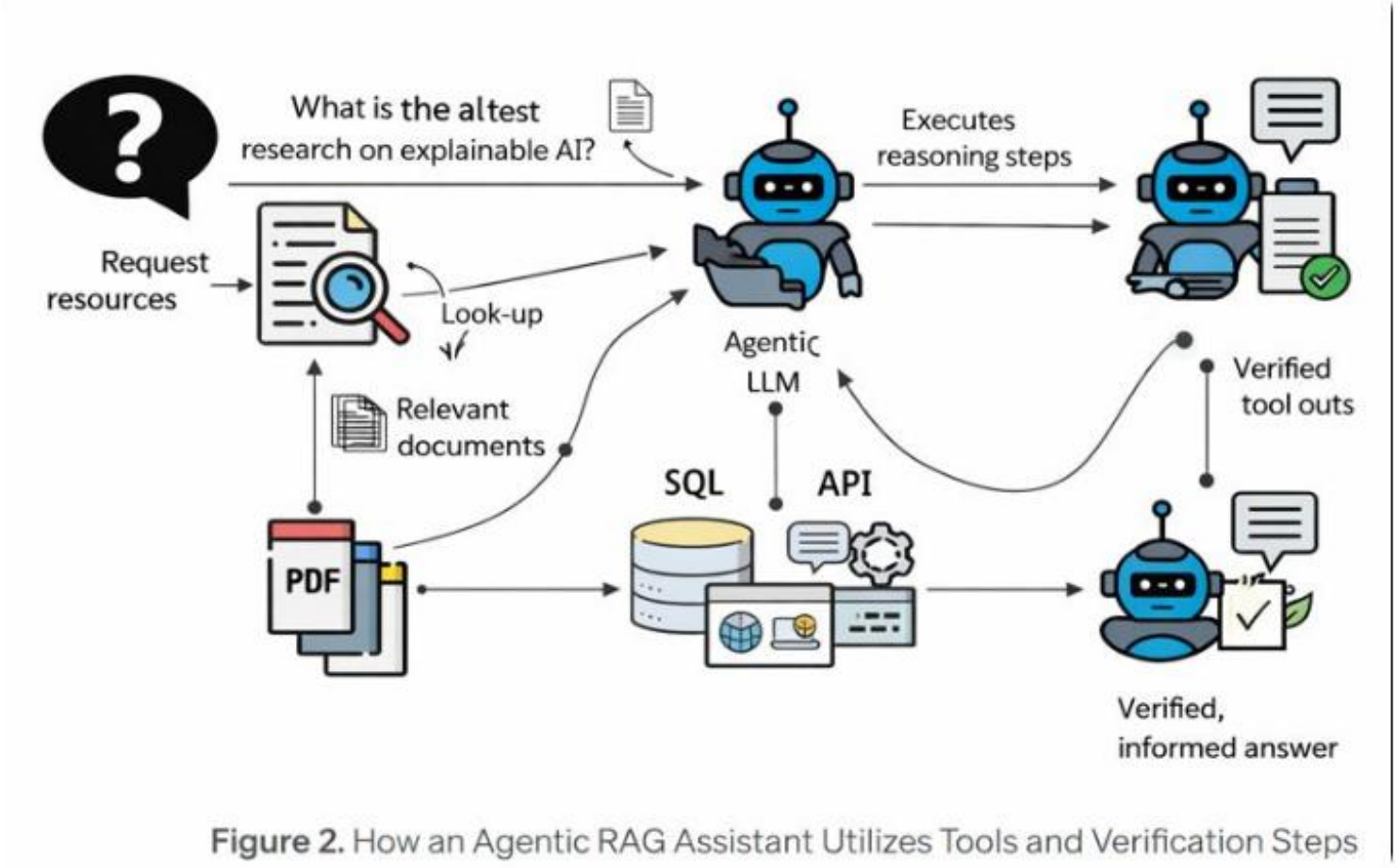

Figure 2. How an Agentic RAG Assistant Utilizes Tools and Verification Steps

Features	Classic RAG Assistant	Agentic RAG Assistant
Retrieve documents	✔ Yes	✔ Yes
Answer questions	✔ Yes	✔ Yes
Plans multi-step workflows	✘ No	✔ Yes
Runs tools (SQL, APIs, ML models)	✘ No	✔ Yes
Corrects mistakes	✘ No	✔ Yes
Handles complex tasks	Limited	Excellent
Behaves like a “SMART EMPLOYEE” – **this is the situation today?**	✘ No	✔ Yes

Every Data Scientist today must know how to design, develop, and deploy Agentic RAG Assistants—this is very important for job security!

Agentic Thinking?

You must understand how to design **multi-step autonomous workflows**:

- Planning → acting → observing → correcting
- Tool selection logic (“Which tool, when, and why?”)
- Failure handling (hallucinations, tool errors, partial results)