# Autonomous RAG

Autonomous RAG is a Retrieval Augmented Generation system where the LLM (or agent) is capable of reasoning, planning, acting, reflecting, and improving — on its own — without manual control over each step.

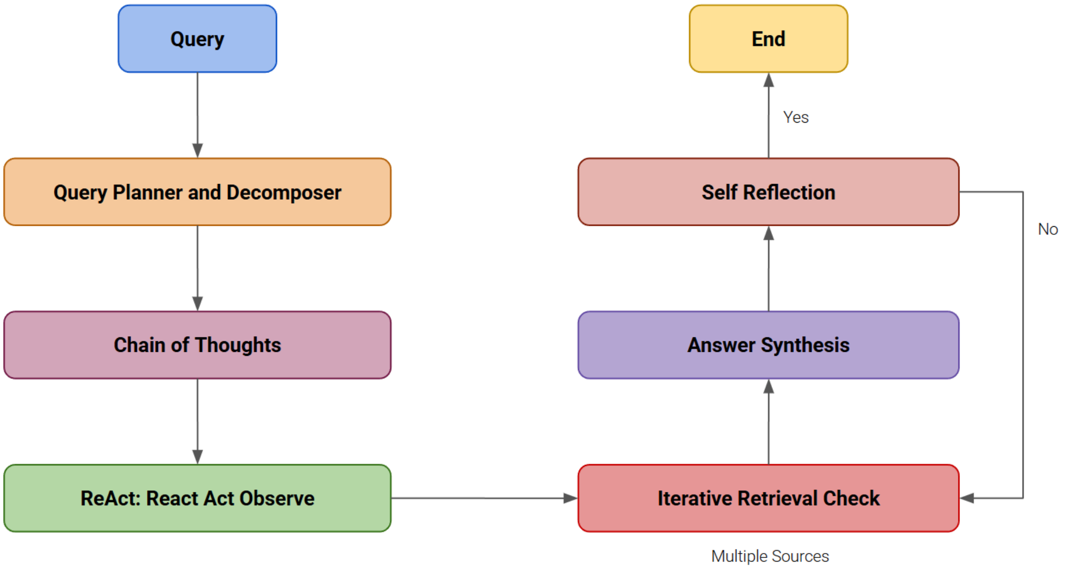
It combines

- Agentic reasoning (like ReAct or LangGraph agents)
- Self-reflection and self-correction
- Dynamic tool selection
- Multi-source retrieval

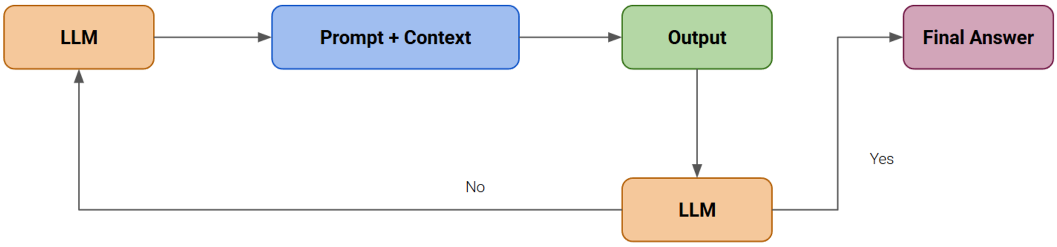
## Core Components of Autonomous RAG

| Component | Role |
|---|---|
| **Planner Agent** | Breaks complex queries into sub-questions |
| **Tool Selector** | Chooses between Wikipedia, ArXiv, vector DBs, APIs, etc |
| **Retriever** | Executes tool calls to retrieve relevant documents |
| **Synthesizer** | Uses LLM to generate the final answer |
| **Reflector** | Verifies whether context or answer is good enough |
| **Retry Loop** | Refines and retries if reflection fails |
| **Memory (opt.)** | Stores feedback, log bad queries, or improves prompts/tools |

## Complete Flow of Autonomous RAG

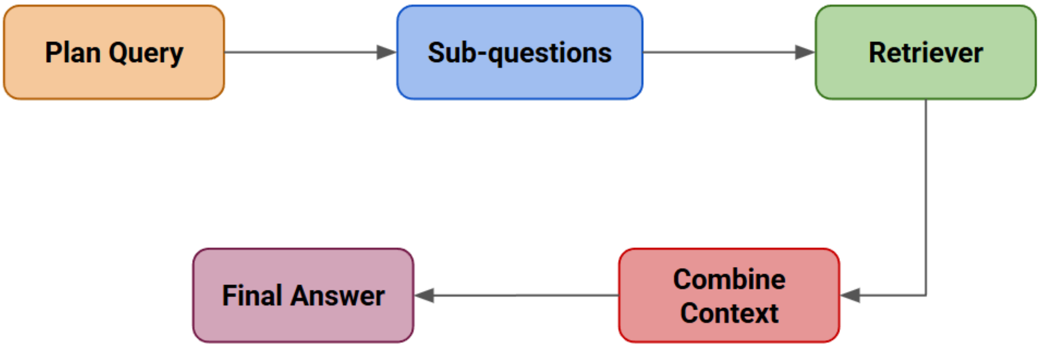


## Self Reflection RAG

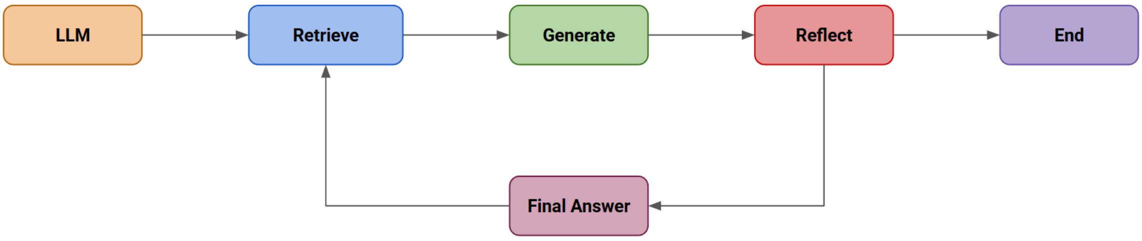


## Query Planning and Decomposition RAG

| | Chain of Thoughts | Query Planning and Decomposition |
|---|---|---|
| **Purpose** | Let the LLM reason step-by-step | Break a complex query into structured sub-queries |
| **Style** | Natural language reasoning path | Explicit sub-queries or formal question segments |
| **Inspiration** | Human-like scratchpad thinking | Structured task planning or modular Q&A |
| **Agent Behavior** | Think → Retrieve → Think → Answer | Plan all → Retrieve all→ Answer once |



## Iterative Retrieval with Self-Reflection



### Benefits

| Feature | Advantage |
|---|---|
| Reflection | Reduces hallucination or overconfidence |
| Query refinement | Improves recall and relevance |
| Agentic behavior | Mimics human-like research process |

## Agentic RAG vs Autonomous RAG

| Concept | Agentic RAG | Autonomous RAG |
|---|---|---|
| **Definition** | A RAG system that uses an agentic approach — where an LLM reasons, plans, and acts using tools | A RAG system that operates independently, with full self-management of planning, retrieving, reflection, and improvement |
| **Focus** | Structured reasoning and tool use (ReAct, LangGraph, etc) | Complete autonomy in task execution, retry, and learning |
| **Behavior** | Think → Act → Observe → Answer | Think → Act → Reflect → Retry → Learn → Answer |
| **Retry Logic** | Optional — usually static agent plans | Built-in retry/refine strategies (context + answer reflection) |
| **Self-Reflection** | May include it optionally | Core feature: reflects on retrieval and answers before finalizing |
| **Tool use** | Uses tool via agents (e.g., Wikipedia, SQL, ArXiv) | Selects and adapts tools dynamically based on reasoning |
| **Planner** | Often present (manual or LLM-generated plans) | Always present — triggers multi-step workflows adaptively |
| **Learning Loop** | Not always present | May log feedback, improve over time |