A table of information

AI-generated content may be incorrect.

Link : [A Complete Guide to Retrieval-Augmented Generation (RAG): 16 Different Types, Their Implementation, and Use Cases | by Gaurav Nigam | aingineer | Medium](https://medium.com/aingineer/a-complete-guide-to-retrieval-augmented-generation-rag-16-different-types-their-implementation-10d48248517b)

| **RAG Pattern** | **Strengths** | **Use Cases** |
| --- | --- | --- |
| Naive | Simple, fast implementation | FAQ, simple knowledge systems |
| Advanced/Query Rewrite | Better relevance via refined retrieval | Ambiguous query handling |
| Corrective (CRAG) | High factual accuracy via verification | Legal/medical/hallucination-critical apps |
| Speculative | Fast and parallelizable generation | High-throughput chat, multi-lingual Q&A |
| Hybrid | Robust across sparse & dense searches | Enterprise datasets with mixed formats |
| Modular | Very customizable, plug-and-play | Complex pipelines, multi-component systems |
| Graph | Supports relational reasoning | Knowledge graph-driven queries |
| Memory-Augmented | Context persistence across sessions | Personal assistants, support bots |
| Multi-Modal | Handles diverse input types | Media-rich, multimodal applications |
| Agentic | Dynamic task-driven workflow | Research labs, decision-making systems |

User Query ("how to improve model performance")

│

▼

+-------------------+

| Query Rewrite | (LLM rewrites or expands query)

+-------------------+

│

▼

+-------------------+

| Retriever | (vector or keyword-based retrieval)

+-------------------+

│

▼

+-------------------+

| LLM Generator | (generates answer based on docs)

+-------------------+

│

▼

OUTPUT



User Query

│

▼

+----------------------------+

| Retriever → Top‑k Docs |

+----------------------------+

│

▼

+----------------------------+

| Retrieval Evaluator |

| (LLM judge scores relevance) |

+----------------------------+

│

┌───┴───┐

│ │

│ Docs relevant? ≥ threshold?

│ │

No Yes

│ │

▼ ▼

(Optional) Re‑retrieve Use docs as context

│ │

▼ ▼

+----------------------------+

| LLM Generator |

+----------------------------+

│

▼

User Answer

User Query

│

▼

+----------------------------+

| Persistent Memory Store |

| (e.g. prior conversations, |

| session context) |

+----------------------------+

│

▼

+----------------------------+

| Merge Query + Memory |

+----------------------------+

│

▼

+----------------------------+

| Retriever (memory + docs) |

| → fetch relevant memory |

| and external docs |

+----------------------------+

│

▼

+----------------------------+

| LLM Generator |

| → generates answer based |

| on memory + docs |

+----------------------------+

│

▼

+----------------------------+

| Append to Memory |

| → store memory from output |

+----------------------------+

│

OUTPUT

User Query

│

▼

+------------------------+

| Agent Orchestrator | ← analyzes intent, decomposes query

+------------------------+

│

▼

+--------------------+ +----------------+

| Retrieval Agents | | Tool/Tool Calls|

| (vector, API, SQL) | | (web search, |

+--------------------+ | external APIs) |

│ |

└────►--------------►─┘

│

▼

+--------------------------+

| Validation & Planning | ← evaluates results, chooses next step

+--------------------------+

│

▼

+--------------------------+

| Generation Agent |

| (LLM generates final answer)|

+--------------------------+

│

▼

+--------------------------+

| Reply to User |

+--------------------------+

<https://chatgpt.com/share/6880b0bf-3c2c-8000-93fd-d7511d8e9272>