25 RAG Architectures Cheat Sheet

Unlock the Power of RAG: Explore 25 unique Retrieval-Augmented Generation (RAG) architectures enhancing AI's ability to retrieve info, generate text, and refine answers. Essential for fields like medical diagnosis and customer support.

What is RAG?

Retrieval-Augmented Generation (RAG) is an AI approach that merges the strengths of document retrieval systems with generative language models. It enables AI to access and utilize external knowledge bases, resulting in responses that are more accurate, contextually relevant, and informed compared to standard generative models alone.

The 25 RAG Architectures

1. Adaptive RAG

What It Is: An iterative model that refines answers by dynamically adjusting its information retrieval scope based on confidence scores. Balances accuracy with speed, perfect for evolving datasets.

When to Use

- Medical diagnosis (high accuracy needed).
- Academic research tools (handling changing data).
- Customer support (resolving complex or vague queries).

Workflow

Adaptive Workflow:

➡️ **Input:**User asks a question.

↓

🔍 **Retrieve (Initial):**Gets initial set of documents.

↓

📝 **Generate:**Creates a draft answer.

↓

📊 **Score:**Assesses answer confidence.

↓

🔁 **Iterate/Refine:**If score is low, expands search & repeats generation/scoring.

↓

✅ **Output:**Delivers final answer when confidence threshold is met.

2. Self-RAG

What It Is: Critiques its own generated output using special reflection tokens (like `[Relevant?]`) to check relevance and factuality. Revises uncertain parts before delivering the final response.

When to Use

Legal contract review (precision vital).

Journalistic fact-checking.

Technical documentation quality assurance.

Workflow

Self-Correction Workflow:

➡️ **Input:**User asks a question.

↓

🔍 **Retrieve:**Gathers relevant documents.

↓

📝 **Generate:**Creates an initial response.

↓

🤔 **Self-Critique:**Uses special tokens to evaluate its own response for uncertainty/relevance.

↓

✍️ **Revise:**Corrects flagged sections if needed.

↓

✅ **Output:**Returns the validated and revised response.

3. HybridAI RAG

What It Is: Blends multiple retrieval methods (e.g., keyword search + vector search) to leverage the strengths of each, achieving both broad coverage (recall) and accuracy (precision).

When to Use

Enterprise search (handling diverse query types).

Multi-domain chatbots.

Research analysis (combining strategies for deeper insights).

Workflow

Hybrid Retrieval Workflow:

➡️ **Input:**User asks a question.

↓

🔍 **Parallel Retrieve:**Simultaneously uses keyword & vector search.

↓

🔗 **Fuse Results:**Combines results from both retrieval methods.

↓

📝 **Generate:**Creates response based on fused results.

↓

🏆 **Rank/Refine:**Orders and polishes the final output.

↓

✅ **Output:**Presents the ranked response.

4. Conversational RAG

What It Is: Maintains context by tracking conversation history and using session memory. Improves responses to follow-up questions for a seamless dialogue experience.

When to Use

Therapy chatbots (remembering user's journey).

Personalized sales support agents.

Interactive storytelling applications.

Workflow

Contextual Conversation Workflow:

➡️ **Input:**User asks a question within a conversation.

↓

💾 **Context Encode:**Encodes current query + conversation history.

↓

🔍 **Retrieve:**Gets documents relevant to query and context.

↓

📝 **Generate:**Creates contextually relevant response.

↓

🧠 **Memory Update:**Stores new interaction in session memory.

↓

✅ **Output:**Delivers the contextual response.

5. XAI RAG (Explainable AI)

What It Is: Focuses on transparency by generating human-understandable explanations for its answers, citing the source information used.

When to Use

Healthcare diagnostics (understanding AI's reasoning).

Financial auditing (requiring transparency).

Regulatory compliance (needing traceable decisions).

Workflow

Explainable Workflow:

➡️ **Input:**User asks a question.

↓

🔍 **Retrieve:**Finds relevant documents.

↓

📝 **Generate:**Creates the answer.

↓

📄 **Extract Justification:**Pulls supporting snippets from source documents.

↓

✅ **Output:**Delivers the answer along with its explanation/sources.

6. Corrective RAG

What It Is: Employs a post-generation fact-checking module to verify and correct potential errors in the initial answer, ensuring high accuracy.

When to Use

News generation (accuracy paramount).

Academic writing assistance.

Legal document drafting.

Workflow

Fact-Checking Workflow:

➡️ **Input:**User asks a question.

↓

📝 **Generate Initial Answer:**Creates a response based on retrieval.

↓

🛡️ **Fact-Check:**Verifies the answer against trusted knowledge sources (e.g., databases).

↓

✍️ **Revise:**Corrects any identified inaccuracies.

↓

✅ **Output:**Delivers the verified and corrected answer.

7. Cost-Constrained RAG

What It Is: Optimizes for resource efficiency by prioritizing cheaper retrieval methods first, balancing answer quality with computational or API call costs.

When to Use

- Budget-conscious startups.
- High-volume chatbots needing cost control.
- Resource-limited mobile applications.

Workflow

Cost-Efficient Workflow:

➡️ **Input:**User asks a question.

↓

🔍 **Retrieve (Cheap First):**Tries low-cost methods (e.g., keyword search).

↓

💡 **Fallback (If Needed):**Switches to more expensive methods (e.g., vector search) if initial results are poor.

↓

📝 **Generate:**Creates response using the retrieved documents.

↓

✅ **Output:**Delivers the response generated within cost constraints.

8. Multi-Modal RAG

What It Is: Capable of processing queries and retrieving information from diverse data types, including text, images, and audio, generating holistic answers.

When to Use

- E-commerce Q&A (product images + text).
- Interactive museum guides (images, audio, text).

Video/Podcast analysis tools.

Workflow

Multi-Modal Workflow:

➡️ **Input:**User submits query (text, image, audio).

↓

🔍 **Cross-Modal Retrieve:**Finds relevant text, images, audio content.

↓

🔗 **Fuse Modalities:**Combines information from different data types.

↓

📝 **Generate:**Creates an answer incorporating multi-modal information.

↓

✅ **Output:**Delivers the integrated multi-modal response.

9. Graph-Based RAG

What It Is: Utilizes knowledge graphs to represent and retrieve interconnected information, enabling complex reasoning about entities and their relationships.

When to Use

Drug discovery (analyzing compound interactions).

Financial fraud detection (identifying connection patterns).

Supply chain optimization (mapping entity relationships).

Workflow

Knowledge Graph Workflow:

➡️ **Input:**User asks a question.

↓

🔧 **Parse Query:**Translates query into a graph query format.

↓

📈 **Traverse Graph:**Navigates the knowledge graph to find relevant subgraphs/nodes.

↓

📝 **Generate:**Creates response leveraging graph-based knowledge.

✅ **Output:**Delivers the answer derived from graph relationships.

10. Context Cache RAG

What It Is: Speeds up responses and reduces load by caching results for frequently asked queries. Delivers cached answers instantly for repeat questions.

When to Use

High-traffic FAQ chatbots.

Real-time data lookups (weather, stocks).

Systems with repetitive query patterns.

Workflow

Caching Workflow:

➡️ **Input:**User asks a question.

↓

💾 **Cache Lookup:**Checks if a valid answer exists in the cache.

↓

✅ **Cache Hit:**If found, immediately returns cached response.

↓

❌ **Cache Miss:**If not found, proceeds to retrieve & generate.

↓

📝 **Generate (on Miss):**Creates a new response.

↓

📦 **Update Cache:**Stores the new response in the cache.

↓

✅ **Output:**Delivers the response (cached or newly generated).

11. Rule-Based RAG

What It Is: Guides retrieval and generation using predefined rules (e.g., keyword matching, regex patterns). Provides consistent, deterministic responses for specific inputs.

When to Use

- Simple, predictable FAQs.
- Structured database lookups.
- Rapid prototyping (validating basic flows).

Workflow

Rule-Driven Workflow:

➡️ **Input:**User asks a question.

↓

🎯 **Match Rule:**Identifies applicable predefined rule(s).

↓

🔍 **Retrieve/Lookup:**Gets data based on the triggered rule.

↓

📝 **Generate (Template):**Creates response often using a predefined template.

↓

✅ **Output:**Delivers the rule-based response.

12. Iterative RAG

What It Is: Improves responses through multiple rounds of interaction. Refines the query and answer based on user feedback or internal checks, allowing for clarification and exploration.

When to Use

- Collaborative research assistants.
- Creative writing tools (improving drafts).
- Exploratory data analysis (evolving questions).

Workflow

Feedback Loop Workflow:

➡️ **Input:**User asks an initial question.

↓

📝 **Generate Initial Answer:**Provides a first response.

↓

❓ **Seek/Receive Feedback:**User provides clarification or feedback.

↓

🔧 **Refine Query:**Adjusts the underlying query based on feedback.

↓

✍️ **Revise Answer:**Generates an improved answer based on the refined query.

↓

✅ **Output:**Delivers the refined response (can loop).

13. Attention-Based RAG

What It Is: Uses attention mechanisms (common in Transformers) to identify and focus on the most critical parts of the user's query during retrieval and generation.

When to Use

Sentiment analysis (focusing on emotive words).

Generating long, coherent text (maintaining focus).

Semantic search (understanding core concepts).

Workflow

Attention-Focused Workflow:

➡️ **Input:**User asks a question.

↓

⚖️ **Calculate Attention:**Determines key phrases/concepts in the query.

↓

🔍 **Retrieve (Weighted):**Finds documents matching high-attention query parts.

↓

📝 **Generate (Focused):**Creates response emphasizing attention-weighted information.

↓

✅ **Output:**Delivers the attention-driven response.

14. Memo RAG

What It Is: Similar to Conversational RAG, it retains session memory (user interactions, history) to personalize future responses, offering a tailored experience.

When to Use

Therapy chatbots needing deep context.

Personalized e-commerce shopping assistants.

Adaptive learning platforms.

Workflow

Personalized Memory Workflow:

➡️ **Input:**User asks a question (potentially mid-session).

↓

💾 **Retrieve Memory:**Accesses past interactions/session data.

↓

🔍 **Retrieve Docs:**Finds documents relevant to query + context.

↓

📝 **Generate Personalized:**Creates answer incorporating query and memory.

↓

📦 **Update Memory:**Stores the latest interaction.

↓

✅ **Output:**Delivers the personalized response.

15. Federated RAG

What It Is: Retrieves information from multiple decentralized data sources without centralizing or exposing the raw data, preserving privacy.

When to Use

Healthcare analysis across hospitals (preserving patient privacy).

Cross-institution financial data analysis.

Secure legal document review across firms.

Workflow

Privacy-Preserving Federated Workflow:

➡️ **Input:**User asks a question.

↓

🔐 **Federated Query:**Sends encrypted/processed query to distributed sources.

↓

🔍 **Local Retrieve:**Each source retrieves relevant info locally.

↓

📊 **Aggregate Results:**Collects processed results/models from sources securely.

↓

📝 **Generate:**Creates response based on aggregated, anonymized results.

↓

✅ **Output:**Delivers answer without exposing raw distributed data.

16. Time-Aware RAG

What It Is: Prioritizes recent or temporally relevant documents, ensuring responses reflect the latest information, crucial for time-sensitive topics.

When to Use

Financial market analysis (latest trends needed).

Real-time news summarization.

Tracking updates in medical guidelines or research.

Workflow

Time-Sensitive Workflow:

➡️ **Input:**User asks a question.

↓

🗓️ **Filter by Time:**Applies time constraints (e.g., "last 24 hours").

↓

🔍 **Retrieve Relevant:**Gets documents matching query and time filter.

↓

📝 **Generate Current:**Creates response based on the latest information.

↓

✅ **Output:**Delivers the up-to-date, time-relevant response.

17. Cross-Lingual RAG

What It Is: Handles queries and retrieves information across multiple languages, often without explicit translation steps, using multilingual models.

When to Use

Global customer support centers.

International news aggregation platforms.

Multilingual knowledge base access.

Workflow

Multilingual Workflow:

➡️ **Input:**User asks question in Language A.

↓

🔍 **Retrieve Multilingual:**Finds relevant documents in Language A, B, C, etc.

↓

📝 **Generate (Target Language):**Creates response in the user's original language (Language A).

↓

✅ **Output:**Delivers the response in the correct language.

18. Diffusion RAG

What It Is: Inspired by diffusion models in image generation, this approach iteratively refines noisy or ambiguous queries/responses by adding and then removing "noise," gradually converging on a clearer result.

When to Use

- Handling very vague or poorly phrased user requests.
- Creative brainstorming or exploration tasks.
- Systems dealing with inherently noisy input data.

Workflow

Iterative Refinement Workflow:

Input: User provides ambiguous/noisy query.

↓

Add Noise (Conceptual): Represents ambiguity.

↓

Iterative Denoising: Gradually refines query/retrieval through multiple steps.

↓

Generate Refined: Creates response based on the clarified query/context.

↓

Output: Delivers the clarified response.

19. Continual Learning RAG

What It Is: Updates its knowledge base and embeddings incrementally as new data arrives, without needing complete retraining. Keeps the model current dynamically.

When to Use

- Real-time news feed analysis.
- Continuously evolving knowledge domains (e.g., scientific research).

Systems needing to adapt quickly to new information.

Workflow

Dynamic Update Workflow:

➡️ **Input:**User asks a question.

↓

➕ **Detect New Data:**System identifies new documents/information added.

↓

⚙️ **Incremental Update:**Updates model/embeddings with new data efficiently.

↓

🔍 **Retrieve (Updated):**Searches the continually updated knowledge base.

↓

📝 **Generate Current:**Creates response based on the latest knowledge.

↓

✅ **Output:**Delivers the response reflecting recent updates.

20. Tiny RAG

What It Is: Optimized for resource-constrained environments like mobile phones or IoT devices. Uses compressed models and efficient algorithms for on-device processing.

When to Use

On-device AI features in mobile apps.

Smart devices with limited connectivity/compute power.

Offline-first applications requiring local intelligence.

Workflow

Edge/Device Workflow:

➡️ **Input:**User query on edge device.

↓

🔍 **Retrieve (Compressed):**Uses lightweight local models/indexes.

↓

📝 **Generate (Efficient):**Uses compressed generative model locally.

✅ **Output:**Delivers the locally computed response on the device.

21. Constitutional RAG

What It Is: Incorporates predefined ethical rules or "constitutions" (e.g., fairness, harmlessness, privacy) directly into the retrieval and generation process to ensure outputs align with ethical standards.

When to Use

Chatbots discussing sensitive topics (government, health).

Educational tools for children (ensuring safe content).

AI systems where bias or harm mitigation is critical (HR, moderation).

Workflow

Ethical Guardrail Workflow:

➡️ **Input:**User asks a question.

↓

📜 **Check Constitution:**Evaluates query and potential responses against ethical rules.

↓

🔍 **Retrieve (Filtered):**Gets documents, potentially filtering unethical sources.

↓

📝 **Generate (Guided):**Creates response ensuring adherence to the constitution.

↓

🚫 **Redact/Modify (If Needed):**Alters or blocks content violating rules.

↓

✅ **Output:**Delivers the ethically compliant response.

22. Bayesian RAG

What It Is: Quantifies uncertainty in its responses by providing confidence scores or probability distributions, helping users gauge the reliability of the information.

When to Use

Financial risk assessment (estimating probabilities).

Medical prognosis (indicating likelihood of outcomes).

Situations where understanding confidence level is crucial.

Workflow

Uncertainty-Aware Workflow:

➡️ **Input:**User asks a question.

↓

🔍 **Retrieve:**Finds relevant documents.

↓

🔢 **Estimate Confidence:**Calculates probability/confidence for docs and potential answers.

↓

📝 **Generate + Score:**Creates response and appends uncertainty score (e.g., 85% confident).

↓

✅ **Output:**Delivers response with its confidence level.

23. Adversarial RAG

What It Is: Designed to be robust against malicious or manipulative inputs (adversarial attacks) intended to trick the system or elicit harmful responses.

When to Use

Public-facing chatbots vulnerable to misuse.

Content moderation systems filtering harmful inputs.

Security applications analyzing potentially malicious text.

Workflow

Robust/Secure Workflow:

➡️ **Input:**User submits query (potentially adversarial).

🚨 **Detect Attack:**Identifies signs of malicious intent or manipulation.

↓

🔧 **Sanitize Input (If Needed):**Neutralizes or modifies harmful query aspects.

↓

🔍 **Retrieve (Safe):**Gets documents based on sanitized/safe query.

↓

📝 **Generate Secure:**Creates a safe and uncompromised response.

↓

✅ **Output:**Delivers the secure, adversarial-resistant response.

24. Sparse-Expert RAG

What It Is: Uses a mixture-of-experts approach, routing queries to specialized sub-models ("experts") trained on specific domains or tasks for more accurate retrieval and generation.

When to Use

Large enterprise knowledge management (routing to domain experts).

Multi-domain technical support systems.

Complex Q&A requiring specialized knowledge areas.

Workflow

Expert Routing Workflow:

➡️ **Input:**User asks a question.

↓

🧭 **Route to Expert:**Selects the best expert model based on query topic.

↓

🔍 **Expert Retrieve:**Selected expert retrieves domain-specific documents.

↓

📝 **Expert Generate:**Expert model generates the specialized response.

↓

🧩 **Aggregate (If Needed):**Combines outputs if multiple experts were consulted.

↓

✅ **Output:**Delivers the expert-generated response.

25. Replug RAG

What It Is: Builds in resilience by automatically retrying failed retrieval attempts using backup data sources or alternative strategies, ensuring system robustness.

When to Use

- Mission-critical applications demanding high availability.
- Systems relying on potentially unreliable data sources or APIs.
- Ensuring graceful degradation during partial system failures.

Workflow

Fallback/Retry Workflow:

➡️ **Input:**User asks a question.

↓

🔍 **Retrieve (Primary):**Attempts retrieval using the main method/source.

↓

❓ **Detect Failure:**Checks if the primary retrieval failed (e.g., timeout, error).

↓

🔄 **Fallback Retrieve:**If failed, automatically tries backup method/source.

↓

📝 **Generate:**Creates response once retrieval succeeds (primary or fallback).

↓

✅ **Output:**Delivers the response obtained through the successful retrieval path.