

Agentic AI

Retrieval-Augmented Generation (RAG): Bridging Knowledge Gaps

Learning Objectives

- Explain the core limitations of traditional Large Language Models (LLMs).
- Define Retrieval-Augmented Generation (RAG) and its purpose.
- Identify the key benefits of using RAG.
- Recognize modern applications of RAG in various industries.
- Describe the three main components of a RAG pipeline.
- Explore use cases for how RAG has improved AI Operations.

What is Retrieval-Augmented Generation?

Motivation: The Challenge With Large Language Models

- **The Power:**

- Human-like text generation across multiple domains
- Massive training datasets enabling broad knowledge
- Versatile capabilities in summarization, translation, and reasoning

- **The Problems:**

- **Hallucinations:** Plausible but factually incorrect outputs
- **Knowledge Cut-off:** Limited to training data timestamp
- **No Source of Verification:** Cannot cite or validate information
- **Domain Gaps:** Struggles with specialized or proprietary knowledge

Motivation: Challenges in Retrieval-Based Systems

- **Lack of Synthesis:** They return multiple relevant sources but do not summarize or interpret the data cohesively.
- **User Burden:** Users must manually review and extract insights from retrieved content.

What is Retrieval-Augmented Generation?

Definition

An AI framework combining information retrieval systems with generative Large Language Models (LLMs).

Core Idea

Before generating responses, the LLM first retrieves relevant information from external knowledge bases.

Analogy

Think of an LLM that can "look up" information in a library before answering questions.

Goal:

Enable LLMs to produce more accurate, contextually relevant, and factually grounded responses.

Why RAG: The Benefits

Improved Accuracy

Grounds responses in verified, external data and significantly reduces hallucinations.

Factual Grounding

Can provide references to retrieved documents, enhancing transparency and verifiability.

Expanded Use Cases

Enables LLMs for specific domains like legal, medical, and internal company knowledge.

Up-to-Date Information

Overcomes LLM knowledge cut-off by querying dynamic sources like live databases and recent articles.

Cost-Efficient

More practical than constant retraining or fine-tuning LLMs for new data.

Increased Trust

Outputs become more reliable and verifiable, building user confidence.

RAG vs. Traditional Search

This file is meant for personal use by renga4599@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. © All Rights Reserved. Unauthorized use or distribution prohibited.

Rag Versus Traditional Search

Retrieval-Augmented Generation (RAG)

Goal: To generate a comprehensive, contextually rich, and factual answer to a user's query.

Process:

- Retrieves relevant information from a knowledge base.
- Augments an LLM's prompt with this information.
- Generates a natural language response.

Output: A synthesized, conversational answer, often with citations. It's designed to *answer* the question directly.

Strengths:

- Reduces hallucinations stemming from LLM use.
- Provides up-to-date information.
- Offers explainability (by citing sources).

Traditional Web Search

Goal: To find and present relevant documents or webpages that might contain the answer to a user's query.

Process:

- Analyzes keywords in the query.
- Matches keywords to indexed content.
- Ranks documents based on relevance (e.g., keyword density, links, authority).

Output: A list of relevant sources. The user must then read and interpret the results to find their answer.

Strengths:

- Supports broad exploration.
- Finding specific documents.
- Provides a variety of sources.

Traditional Search

The screenshot shows a Google search results page for the query "popular polish shoe brands". The top result is a link from Business Insider titled "The 6 best shoe polishes for shining leather shoes", dated Jan 29, 2024. Below the search bar, there's a video section titled "Videos" showing three YouTube videos related to shoe polish reviews. The first video is from Gentleman's Gazette, the second from Kirby Allison, and the third from Robert Powers. Each video thumbnail includes a play button and a "key moments in this video" link.

Google search results for "popular polish shoe brands":

- Business Insider**
<https://www.businessinsider.com/reviews/style/the-best-shoe-polishes-for-shining-leather-shoes-2024-1>
The 6 best shoe polishes for shining leather shoes
Jan 29, 2024 — Our top picks for the best shoe polishes in 2024 · Best conditioner: Saphir Renovateur · Best wax: Saphir Medaille d'Or Pate de Luxe · Best color: ...
- Ranking Shoe Polishes (19 BEST & WORST Brands) ft ...**
YouTube · Gentleman's Gazette
1 month ago
32:32
20 key moments in this video
- What is the Best Shoe Polish? | Shoe Polish Review**
YouTube · Kirby Allison
Oct 18, 2017
10:18
9 key moments in this video
- Cheap vs High End Shoe Polish - Is it Worth it?**
YouTube · Robert Powers
Oct 21, 2022
15:15
9 key moments in this video

- Traditional search lacks context
- Google has algorithms to correct

Generative Search

list some popular polish shoe brands

Here are some popular Polish shoe brands:

1. **Wojas** - Known for high-quality leather shoes and a wide range of styles for both men and women.
2. **Gino Rossi** - Offers elegant and fashionable shoes, often with a focus on sophisticated designs.
3. **Badura** - Renowned for its stylish and comfortable footwear, including both casual and formal options.
4. **CCC** - A well-known brand offering a variety of shoe styles, including many popular international brands.
5. **Ryłko** - Specializes in comfortable and stylish shoes made from high-quality materials.
6. **Kazar** - Known for its luxurious and trendy shoe collections, often featuring unique designs.
7. **Venezia** - Offers a wide range of fashionable shoes, blending classic and contemporary styles.
8. **Lasocki** - Provides affordable yet stylish shoes, often focusing on comfort and durability.
9. **Bartek** - Specializes in children's footwear, known for its quality and ergonomic designs.
10. **Conhpol** - Offers high-quality men's shoes, combining traditional craftsmanship with modern styles.

- Generative search has context to improve the quality of search.
- Avoids "keyword" based SEO optimization.
- More accurate and relevant.

Generative Search

Generative AI can provide sourcing for information.
Search for “Which planet in the solar system has the most moons?”

 Space.com
<https://www.space.com> › The Universe › Solar System

Jupiter now has the most moons in the solar system ...

Feb 3, 2023 — Jupiter now has the most moons in the solar system, beating Saturn thanks to 12 newfound satellites.



As of now, Jupiter holds the title for the planet with the most moons in the solar system. Jupiter has 95 confirmed moons. This number is constantly updated as new moons are discovered and confirmed by astronomers.

References

- NASA. (2023). Jupiter Overview. Retrieved from <https://solarsystem.nasa.gov/planets/jupiter/overview/>

 KiwiCo
<https://www.kiwico.com> › blog › stem › which-planet-h...

Which Planet Has the Most Moons?

Jul 4, 2023 — Within our solar system, Saturn stands tall as the planet with the most moons, boasting an incredible count of 145.



As of now, Saturn has 83 confirmed moons. This number is constantly updated as new moons are discovered and confirmed by astronomers.

References

- NASA. (2023). Saturn Overview. Retrieved from <https://solarsystem.nasa.gov/planets/saturn/overview/>

 The Planets Today
<https://www.theplanetstoday.com> › which_planet_has_th...

Which Planet is has the most moons?

Which planet has the most moons? Uranus as 27 moons, Neptune has 14 moons, Mars has 2 moons, Earth has 1 moon, Venus and Mercury have no moons.



Memory updated

Why did you say 83 confirmed moons when the NASA website you referenced says there are 146 moons?

You are correct. Saturn has 146 moons, with 83 of them confirmed and 63 still awaiting confirmation. Here is the corrected information:

Generative Search

- After correction, subsequent searches may be updated.

which planet in the solar system has the most moons? Include references in APA format.



As of now, Saturn holds the title for the planet with the most moons in the solar system. Saturn has a total of 146 moons, with 83 confirmed and 63 awaiting confirmation.

References

- NASA. (2023). Saturn Overview. Retrieved from <https://solarsystem.nasa.gov/planets/saturn/overview/>

Evolution of RAG Architectures

Naive RAG (2020-2021)

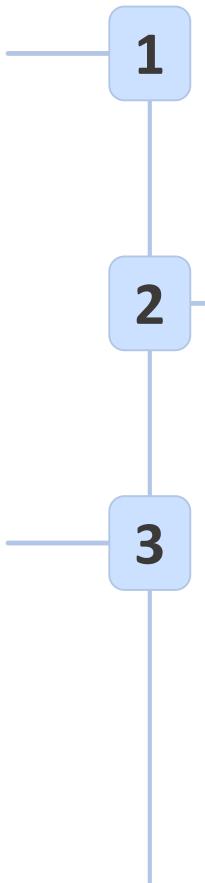
The Foundation: Simple query embedding → retrieve top-k documents → concatenate → feed to LLM

Limitations: Retrieved irrelevant information, struggled with complex multi-hop reasoning

Modular/Agentic RAG (2024-Present)

Current Frontier: Customizable toolkits, knowledge graph integration, multi-modal capabilities, iterative reasoning

Future: Autonomous systems that can plan, use tools, and reason before generating responses



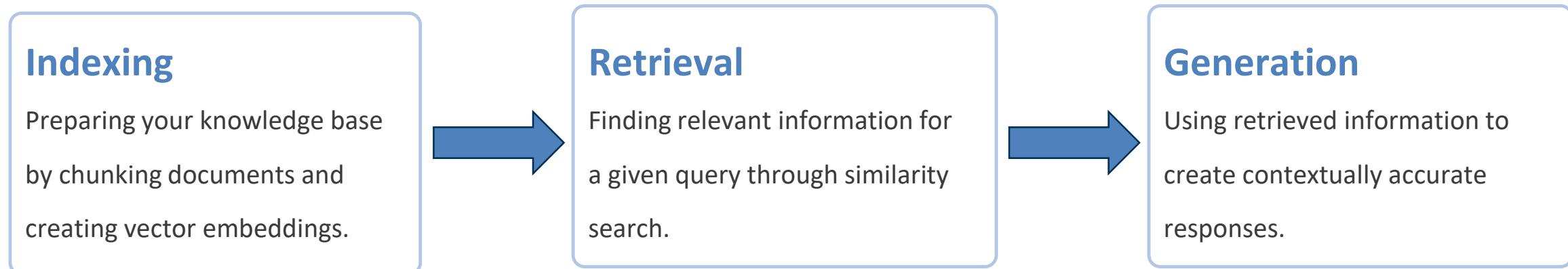
Advanced RAG (2022-2023)

Key Optimizations: Fine-tuned retrievers, semantic search, dense embeddings, re-ranking systems, context compression

Impact: Significantly improved relevance and reduced noise in retrieved content

The Components of a Rag Pipeline

How RAG Works: Pipeline Overview



This three-step process transforms static LLMs into dynamic, knowledge-aware systems.

Component 1: Knowledge Base and Indexing

- **The Foundation Layer:**
 - **Knowledge Base:** Your external data sources—documents, databases, web pages, proprietary company information, real-time feeds
- **The Preparation Process:**
 - **Chunking:** Breaking large documents into manageable segments (paragraphs, sentences, logical sections)
 - **Embedding Generation:** Converting text chunks into numerical vector representations that capture semantic meaning
 - **Vector Storage:** Storing embeddings in specialized databases optimized for similarity search
- **Quality Considerations:**
 - Chunk size, overlap strategies, and embedding model selection critically impact retrieval quality and downstream generation performance.

Component 2: The Retriever

Query Processing

User query is converted into a vector embedding using the same model that indexed the knowledge base.

Ranking and Selection

Retrieved candidates are scored and ranked, with top-N most relevant chunks selected for generation.

Similarity Search

The vector database performs high-speed similarity calculations to find semantically relevant document chunks.

Context Assembly

Selected chunks are formatted and prepared as context for the language model input.

- The retriever's effectiveness depends on embedding quality, search algorithms, and ranking strategies that determine which information reaches the generation stage.

Component 3: The Generator (LLM)

- **The Synthesis Stage:**
 - **Augmented Prompt Construction:** The original user query is enhanced with retrieved document chunks, creating a rich, contextual prompt.
- **LLM Processing:**
 - The language model processes this **augmented prompt** using its pre-trained knowledge combined with the newly provided external context.
- **Intelligent Response Generation:**
 - Synthesizes information from multiple sources.
 - Maintains factual accuracy through grounding.
 - Generates coherent, contextually appropriate responses.
 - Can provide source attribution and confidence indicators.

The Result:

Factual, up-to-date, and contextually rich responses that combine the creativity of LLMs with the accuracy of retrieved knowledge.

Summary

- RAG is a critical component supporting evolution toward more intelligent, autonomous systems that can reason, plan, and interact with complex knowledge environments in real-time.
- **Key Takeaways:**
 - RAG bridges critical gaps in standalone LLM capabilities by providing external, verifiable information
 - Real-world impact across enterprise, research, and customer service applications
 - Continuous evolution from naive implementations to sophisticated agentic systems
- **Next Steps for Users:**
 - Experiment with embedding models and chunking strategies
 - Explore multi-modal RAG implementations
 - Focus on evaluation metrics and user feedback loops

Evaluating RAG Outputs: Ensuring Quality and Trust

Importance of Evaluation

- Without rigorous evaluation, RAG systems can:
 - Misinterpret retrieved information
 - Generate responses unfaithful to source documents
 - Retrieve irrelevant contexts for user queries
 - Produce incomplete or verbose answers
- Continuous evaluation ensures your RAG system maintains trust and effectiveness in real-world applications where decisions matter.

Key Dimensions for RAG Evaluation

1

Accuracy

Does the generated response accurately reflect the retrieved source documents without hallucinating beyond the provided context?

2

Answer Relevance

Does the generated answer directly address what the user asked? Is it on-topic and helpful for their specific need?

3

Context Relevance

Are the retrieved documents themselves relevant to the user's query? Poor retrieval inevitably leads to poor generation.

4

Completeness

Does the answer cover all aspects of the query that could be addressed by the retrieved context without omitting important information?

Additional Evaluation Considerations

Conciseness

Is the answer brief and to the point without unnecessary verbosity or redundancy?

Harmfulness/Bias

Does the answer contain any inappropriate, biased, harmful, or misleading content?

Readability

Is the answer well-written, easy to understand, grammatically correct, and logically structured?

LLM-As-A-Judge

- **Idea:** Use a well-instructed LLM to assess RAG outputs.
- **Result:** Automated evaluation that can handle the open-ended nature of responses while maintaining consistency across thousands or millions of evaluations.

How LLM-As-A-Judge Works

Input

Provide the judge model with:

- Original user query
- Retrieved documents/context
- RAG-generated answer
- Evaluation criteria and scoring rubric

Instruction

The LLM judge is prompted to assess the RAG output:

- "Rate faithfulness on a scale of 1-5"
- "Identify any factual errors"
- "Score the relevance to the query"

Output

The judge provides:

- Numerical scores per dimension
- Detailed justification
- Suggested improvements
- Overall quality assessment

LLM-As-A-Judge: Considerations

Advantages

- Reduces human effort
- Enables continuous monitoring of millions of outputs
- Ideal for understanding complex relationships
- Enables consistent evaluation
- Cost-effective for ongoing, large-volume assessment

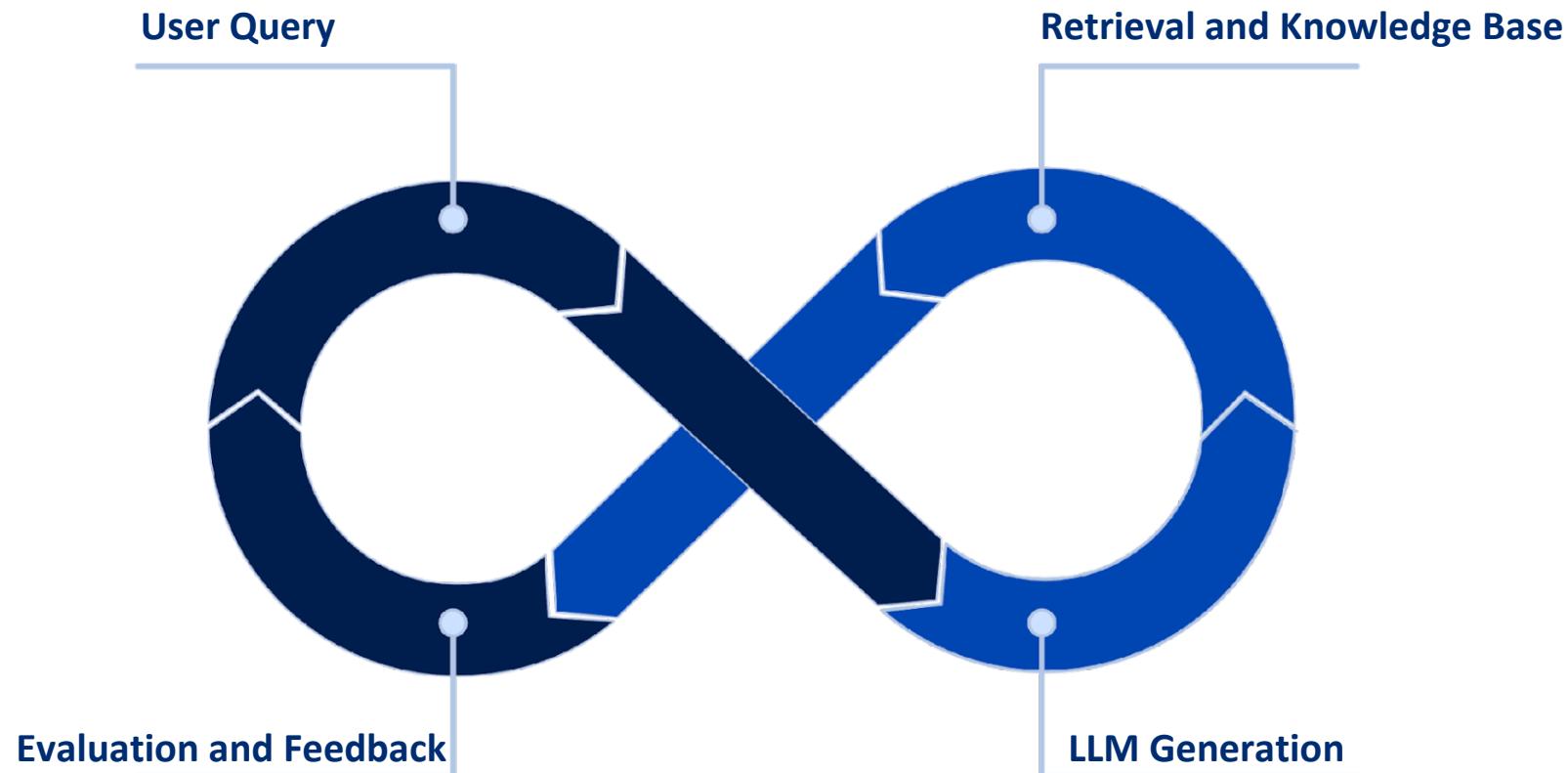
Limitations

- The LLM judge is susceptible to biases and hallucinations
- May not catch subtle domain-specific errors
- Requires careful prompt engineering
- Needs validation against human judgments to ensure alignment

Best Practices

- Use a more powerful model as the judge than the one being evaluated
- Include clear, explicit criteria in judge prompts
- Validate judge outputs against human evaluations periodically
- Use as a filter to prioritize outputs for human review

The Continuous Improvement Loop



Real-World Business Impact of RAG

Enhanced Customer Support and Service

- **Business Problem:** Traditional LLM-based customer support can provide generic or incorrect information, especially for complex products, rapidly changing policies, or company-specific FAQs.

- **Solution:**

Integrate the LLM with a RAG system connected to:

- Company knowledge bases (FAQs, product manuals, troubleshooting guides)
- Customer relationship management systems for customer history
- Real-time policy documents

- **Observable Metrics:**

Average handle time
for complex queries

Percentage of first
contact resolution

Percentage of
escalated tickets
requiring human
intervention



Personalized Wealth Management

- **Business Problem:** Generic financial advice from LLMs is insufficient. Clients need recommendations tailored to their unique portfolios, risk tolerance, and real-time market conditions.
- **Solution:**
 - Integrate the LLM with a RAG system connected to:
 - Real-time market data feeds (stocks, bonds, commodities)
 - Client investment portfolios and historical performance
 - Risk assessment profiles
 - Financial product databases and regulatory guidelines
- **Observable Metrics:**



Client engagement with personalized financial insights

Streamlined onboarding process for new clients by assessing needs faster

Reduced advisory costs by automating personalized, client communications

Advanced Medical Diagnostics and Research

- **Business Problem:** Medical knowledge is vast, constantly evolving, and highly specialized. Both LLMs and humans may struggle with the latest clinical trial data, drug interactions, or patient-specific Electronic Health Records (EHRs).
- **Solution:**
 - Integrate the LLM with a RAG system connected to:
 - Up-to-date medical journals and research papers
 - Electronic Health Records (EHRs) for patient history
 - Drug databases and clinical trial results
 - Genomic data repositories
- **Observable Metrics:**

Rate of identification
of promising drug
candidates in early
research phases

Improved accuracy in
predicting drug
interactions

Enhanced ability to
correlate patient
genomic data with
treatment outcomes

