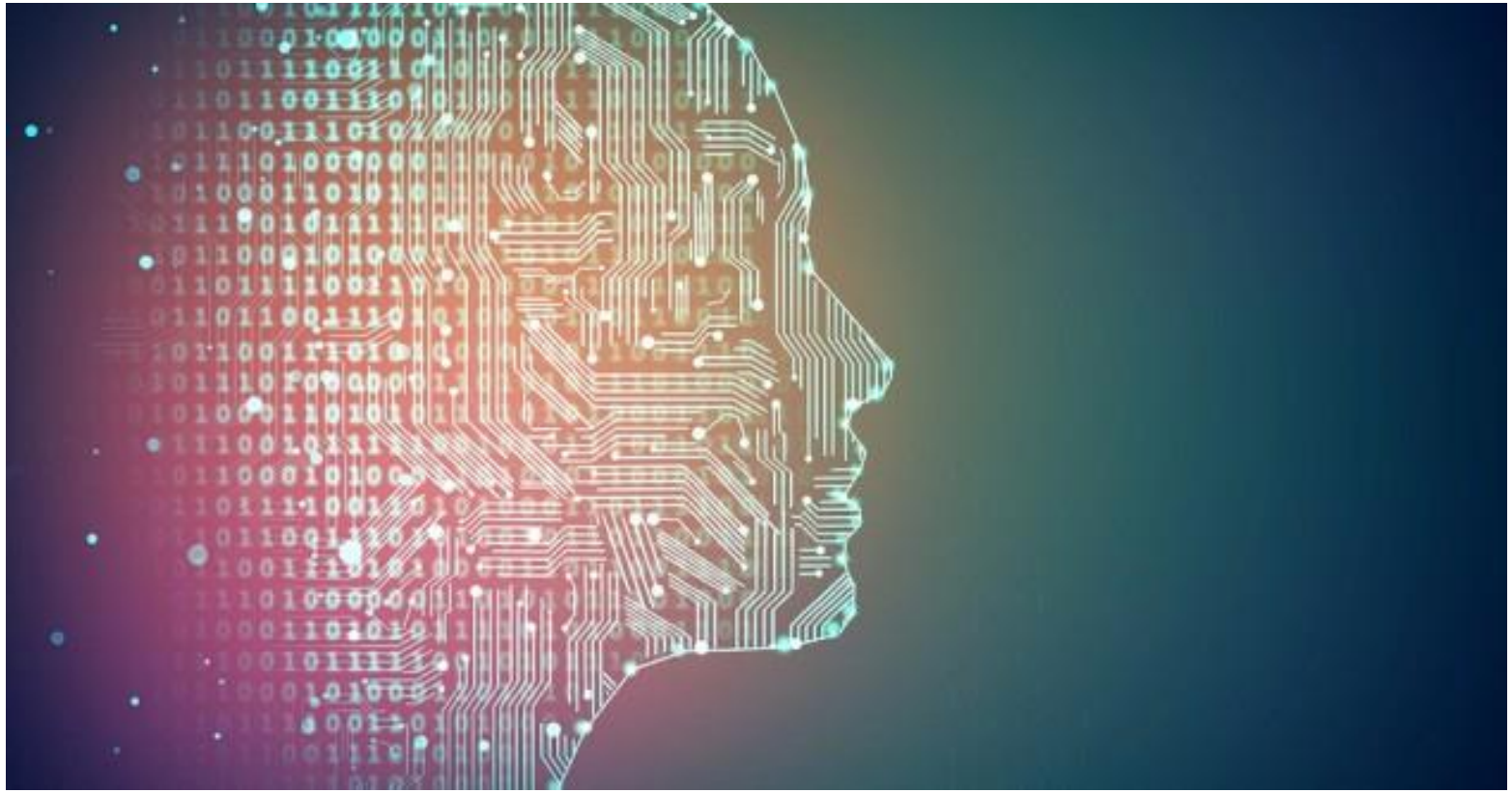
Retrieval-Augmented Generation (RAG)

Thuan L Nguyen, PhD

2: Generative AI: LLM: Retrieval-Augmented Generation (RAG)



AI Deep learning (Source: mindovermachines.com)

3: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

1. Generative AI: Retrieval-Augmented Generation (RAG): Overview

2. Generative AI: RAG: Why Do We Need It?

3. Generative AI: RAG: A Bit of History

4. Generative AI: RAG: Core Concepts & RAG Process

5. Generative AI: RAG: Main Parameters & Properties

6. Generative AI: RAG: Sequence Models vs Token Models

7. Generative AI: RAG: PROs & CONs

4: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

Artificial Intelligence: Generative AI

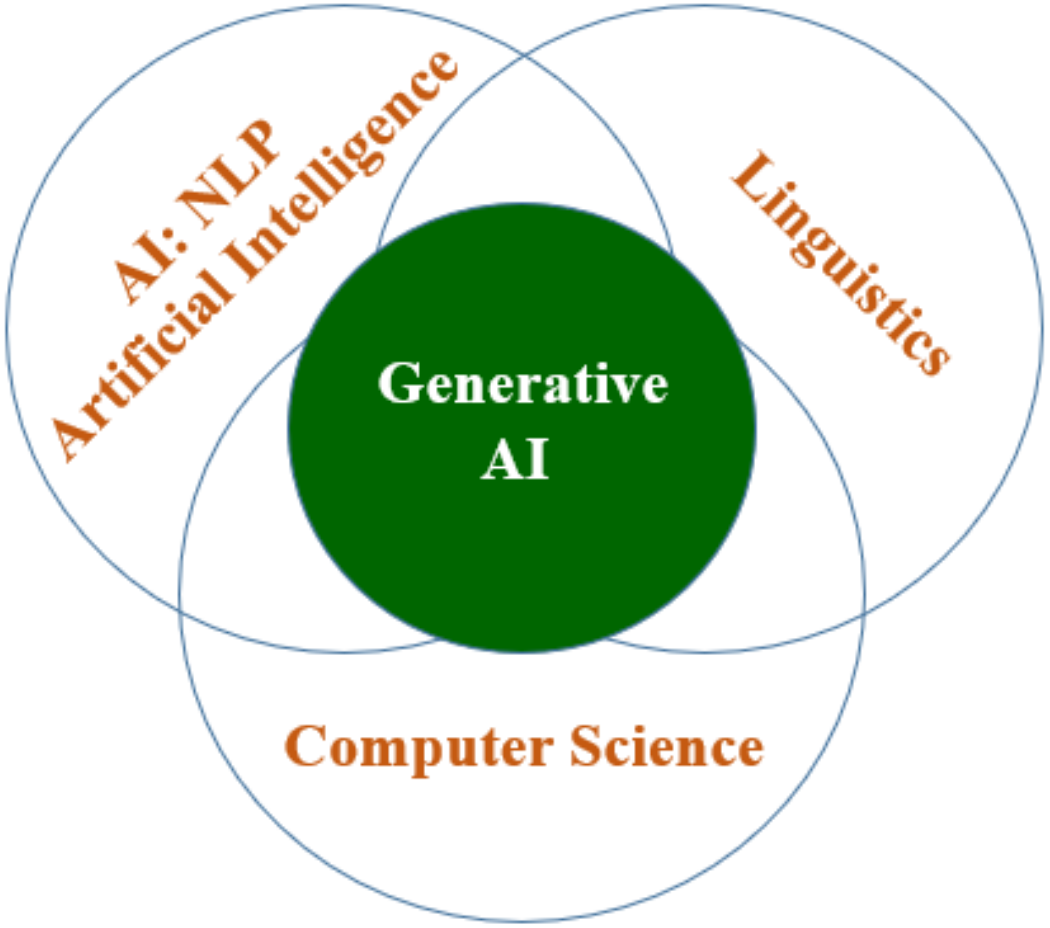
What is It?

Generative AI: A category of artificial intelligence focused on using AI deep learning models to generate new contents, including text, images, audio, video, and more. The contents are novel but look realistic and may be indistinguishable from human-created ones.

5: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

Artificial Intelligence: Generative AI: LLM

What is It?



Generative AI is based on the NLP technologies such as Natural Language Understanding (NLU) and Conversational AI (AI Dialogues) - Those among the most challenging tasks AI needs to solve.

6: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

Artificial Intelligence: Generative AI: LLMs

Large Language Models

Large Language Models (LLMs) are revolutionary AI **Deep Learning neural networks** that excel in natural language understanding (NLU) and content generation.

- “LARGE" in LLMs refers to the vast scale of data and parameters used to train them, allowing LLMs to develop a comprehensive understanding of language.
- Being particularly transformer-based models trained on massive text datasets using deep learning techniques,
- Able to learn complex language patterns, capture nuances like grammar and tone, and generate coherent and contextually relevant text

7: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

Artificial Intelligence: Generative AI: LLMs

Large Language Models

Large Language Models (LLMs) are revolutionary AI **Deep Learning neural networks** that excel in natural language understanding (NLU) and content generation.

- “LARGE" in LLMs refers to the vast scale of data and parameters used to train them, allowing LLMs to develop a comprehensive understanding of language.
- Being particularly transformer-based models trained on massive text datasets using deep learning techniques,
- Able to learn complex language patterns, capture nuances like grammar and tone, and generate coherent and contextually relevant text

8: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

RAG: Overview

- Retrieval-Augmented Generation (RAG) is a technique that **augments** traditional language model generation with the ability to retrieve and integrate information from a knowledge base (like Wikipedia or a company-specific database). This technique improves the **factual accuracy** and **consistency** of the generated text.
- **RAG** cleverly integrates information retrieval with powerful generative language models to enhance **reliability** and **accuracy**.
- The **key idea** behind RAG is to **augment** the **text generation process** with relevant facts and information retrieved from a **knowledge base**. This knowledge base could be a structured database, Wikipedia, or a collection of documents.

***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*

9: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

RAG: Main Parameters and Properties

- **Knowledge Base**: RAG can use various structured (e.g., databases) or unstructured (e.g., Wikipedia, text documents) knowledge bases. The choice depends heavily on the nature of the task.

- **Retrieval Model**: The performance of RAG is often dependent on the quality of the retrieval model. Common techniques include traditional keyword-based search, dense vector representations (like BM25), and neural retrievers.

- **Generative Model**: RAG is flexible and can work with different generative language models (e.g., LLMs like Gemini, GPT, Claude, and other transformer-based models like BART or T5).

- **Sequence vs. Token Models**: RAG has two primary variations:
  - RAG-Sequence: Generates the entire output sequence in one go.
  - RAG-Token: Generates the output one token at a time, allowing the retrieval model to provide new context after each generated token.

10: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*

RAG: Sequence Models (RAG-Sequence)

- Focuses on a single retrieved document for the entire generation process.
- Generates the entire output sequence based on the context of that single document.

- **Advantages**:
  - Simpler and potentially faster processing.
  - Maintains consistency in the generated text by using information from a single source.

- **Disadvantages**:
  - Might be less flexible in incorporating information from diverse sources.
  - May struggle with tasks requiring references from multiple documents.

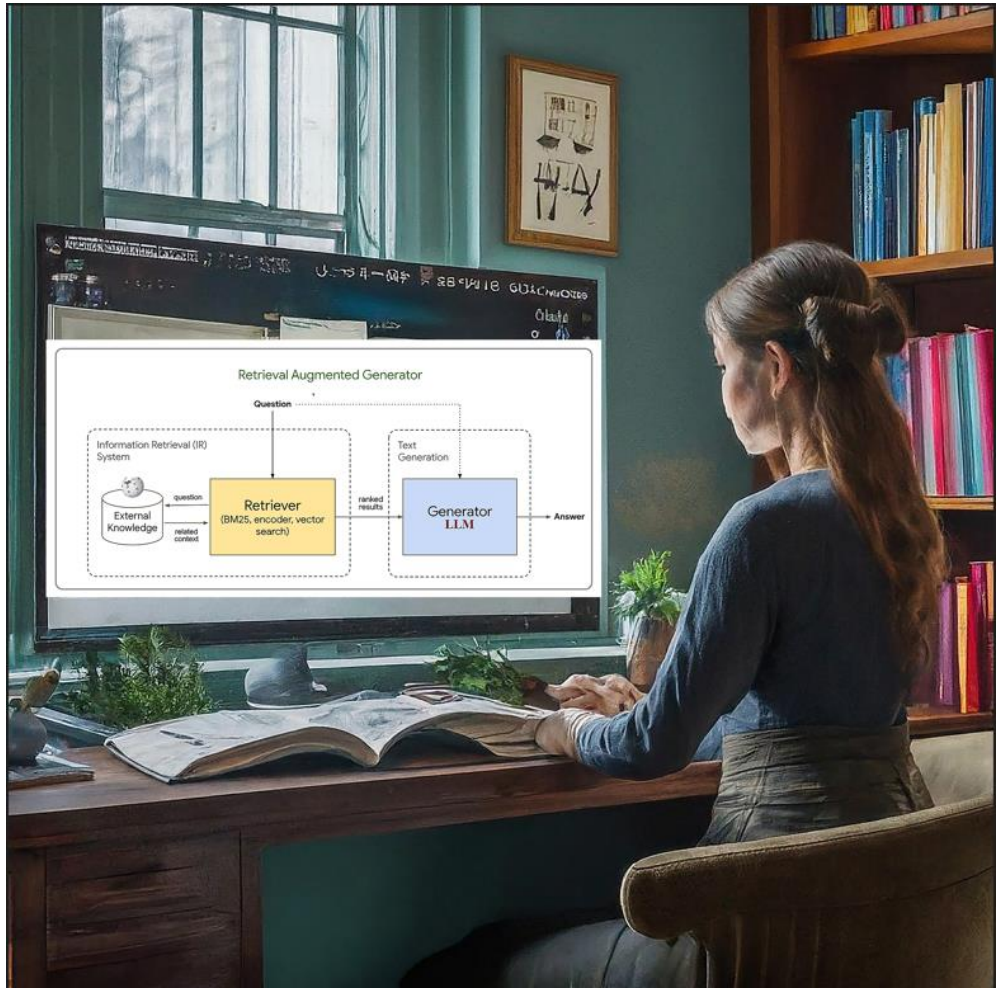
11: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

RAG: Token Models (RAG-Token)

- Allows for different "latent documents" to be drawn for each target token in the generated sequence.
- Provides more flexibility as each word can be informed by the most relevant context.

- **Advantages**:
  - Can potentially generate more informative and nuanced outputs by leveraging multiple knowledge sources.
  - Well-suited for tasks requiring references from various documents.

- **Disadvantages**:
  - Can be computationally more expensive compared to RAG-Sequence.
  - Might introduce inconsistencies in the generated text due to switching between document contexts.



***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*

12: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*

RAG: RAG-Sequence – When to Use

More **likely** if the **priorities** are efficiency and consistency:

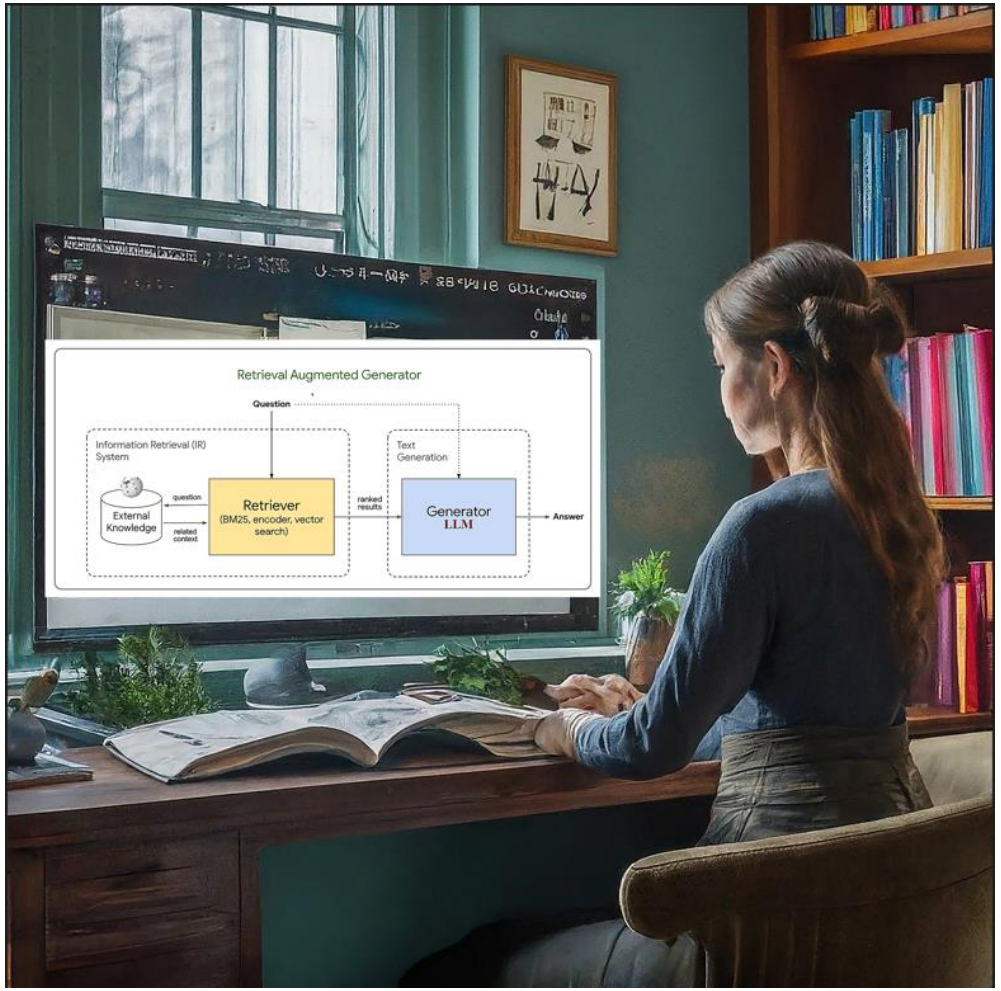
- If the primary focus is on generating coherent and consistent text based on a single source document, **RAG-Sequence** would be a good choice.
- This could be applicable for tasks like question answering or summarization where a single relevant document is sufficient for the output.

13: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

RAG: RAG-Token – When to Use

More **likely** if the **priorities** are flexibility and information richness:

- If the primary focus is on generating informative text that can leverage knowledge from various sources, **RAG-Token** would be a better fit.
- This could be beneficial for tasks like creative writing or information retrieval where incorporating diverse perspectives is crucial.



***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*

14: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

RAG: Sequence vs Token – The Choice for Implementing RAG with LangChain

The **ideal choice** depends on the **specific goals** and **use cases** of the LangChain gen AI platform.

- If factual consistency and efficiency are top **priorities**:
  - Select **RAG-Sequence**.
  - This ensures all generated text aligns with a single source, reducing potential factual inconsistencies.
  - The simpler architecture also translates to faster processing.

- If flexibility and incorporating diverse information are **crucial**:
  - Choose **RAG-Token**.
  - This allows LangChain to leverage the strengths of multiple retrieved documents, potentially leading to richer and more informative outputs.
  - This is particularly beneficial for tasks where comprehensive understanding and diverse perspectives are desired.

15: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

RAG: Sequence vs Token – The Choice for Implementing RAG with LangChain

The **ideal choice** depends on the **specific goals** and **use cases** of the LangChain gen AI platform.

Additional Considerations:

- Computational Resources: RAG-Token requires more resources due to its complex nature. Consider LangChain's available hardware and processing power before making a decision.

- Task-Specific Evaluation: Ultimately, the best approach is to evaluate both models on your specific LangChain tasks. This will provide concrete data on which model performs better for your desired outcomes.

16: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*

RAG: Similar & Comparable Techniques

RAG shares some similarities with other techniques, but also has distinct characteristics:

- **Knowledge-Grounded Dialogue Systems**:
  - Both RAG and knowledge-grounded dialogue systems use external knowledge.
  - However, RAG focuses on single-turn text generation tasks, while dialogue systems handle multi-turn conversations over an extended period.

- **Open-Domain Question Answering**:
  - These systems also retrieve information before answering.
  - RAG differs by focusing on generation tasks, not just direct answer extraction.

17: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*

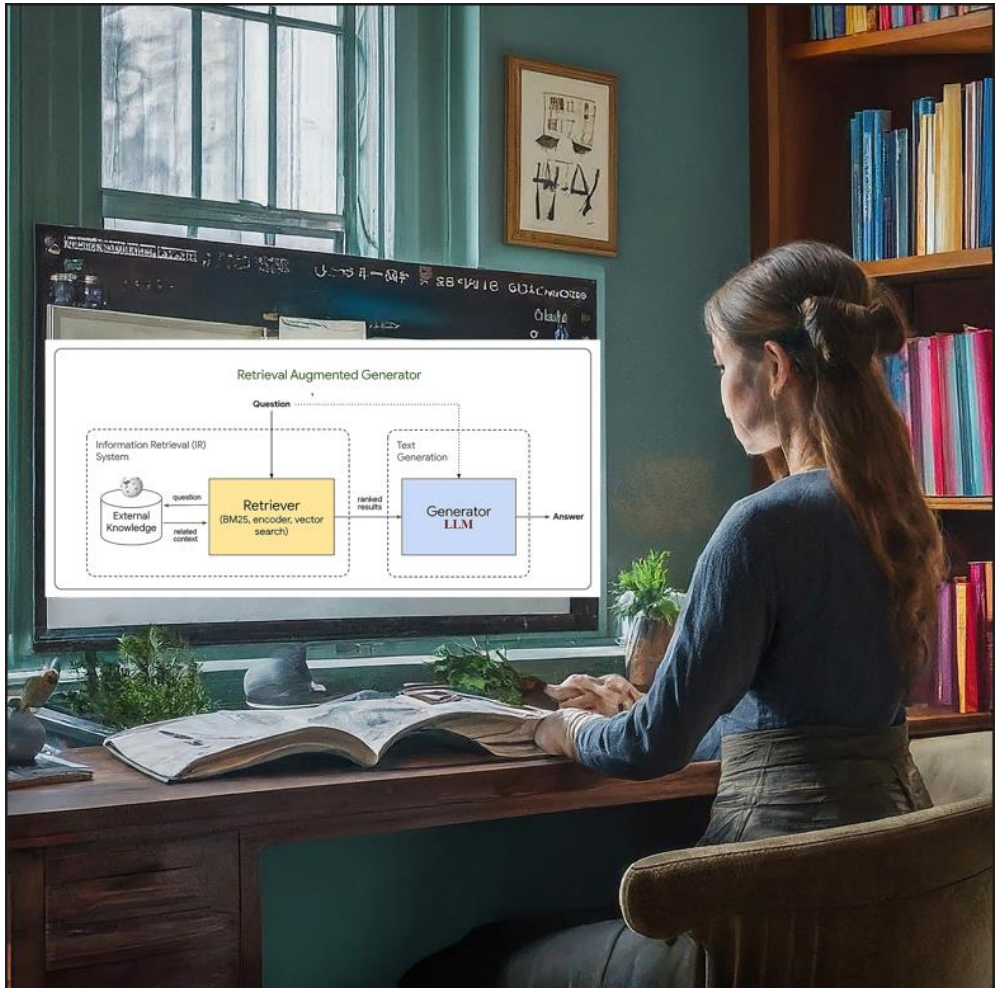
RAG: PROs

- **PRO: Enhanced Accuracy**: Supports responses with factual evidence
  - → Improved factual accuracy
- **PRO: Up-to-date Information**: RAG can access information beyond the LLM's training data.
  - → Adaptability to update knowledge
- **PRO: Transparency**: The retrieval step provides some level of explainability
  - → Interpretability and traceability
- **PRO: No need** for LLM retraining

27: Generative AI: LLM: Retrieval-Augmented Generation (RAG)

RAG: CONs

- **CON: Computational Cost**: The retrieval process adds overhead, additional layer of complexity and cost.
- **CON: Reliance on Knowledge Base**: RAG's quality depends heavily on the quality and relevance of the knowledge base.
- **CON: Dependency on Retrieval** – Retrieval Errors:
  - Mistakenly retrieved information can negatively impact the response.
  - If the retrieval component fails to identify relevant documents, the generated answer may still be incorrect.
- **CON: Potential for Misalignment**: Mismatches in writing style or factuality between the knowledge base and the language model can lead to inconsistencies in the output.



***Source**: Thuan L Nguyen – AI Generated images using Google-DeepMind Imagen 2.0*