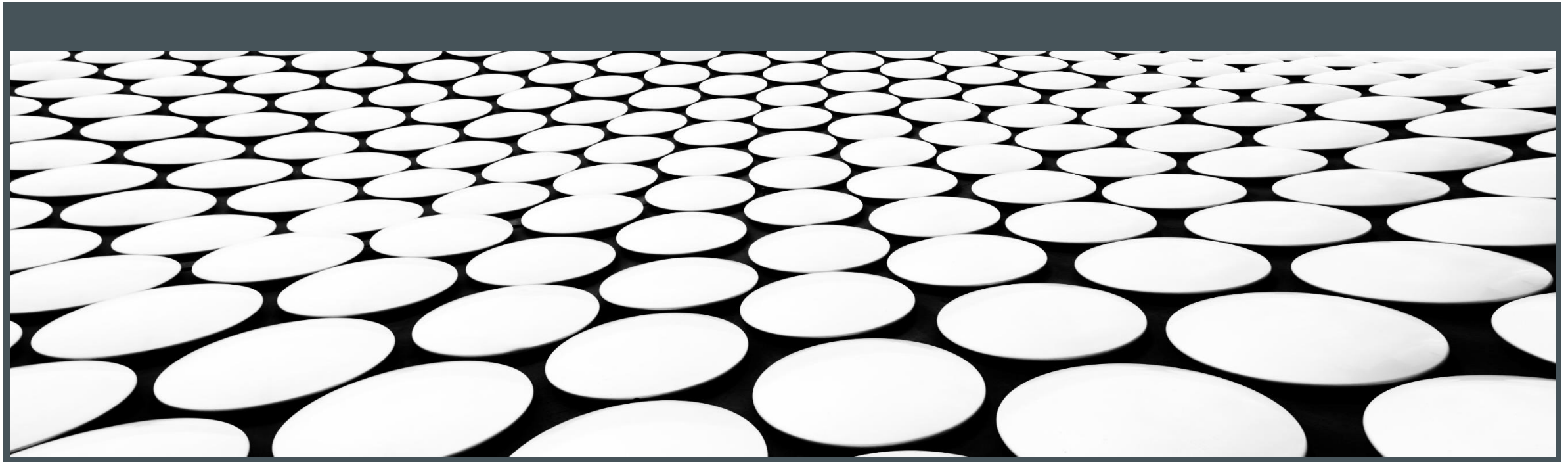
INTRODUCTION TO RETRIEVAL AUGMENTED GENERATION



What you'll learn after completing this module, you should be able to:

1. Identify the key use cases for RAG
2. Explain the benefits of using RAG to expand the knowledge base of a large language model
3. Describe the key steps in a RAG workflow

Topic will be covered in this session…

1. An overview of retrieval-augmented generation
2. Benefits of implementing RAG
3. How does RAG work?
4. Simulation: Enhancing a large language model using RAG

An overview

Of

Retrieval-Augmented Generation

IBM SkillsBuild

edunet foundation

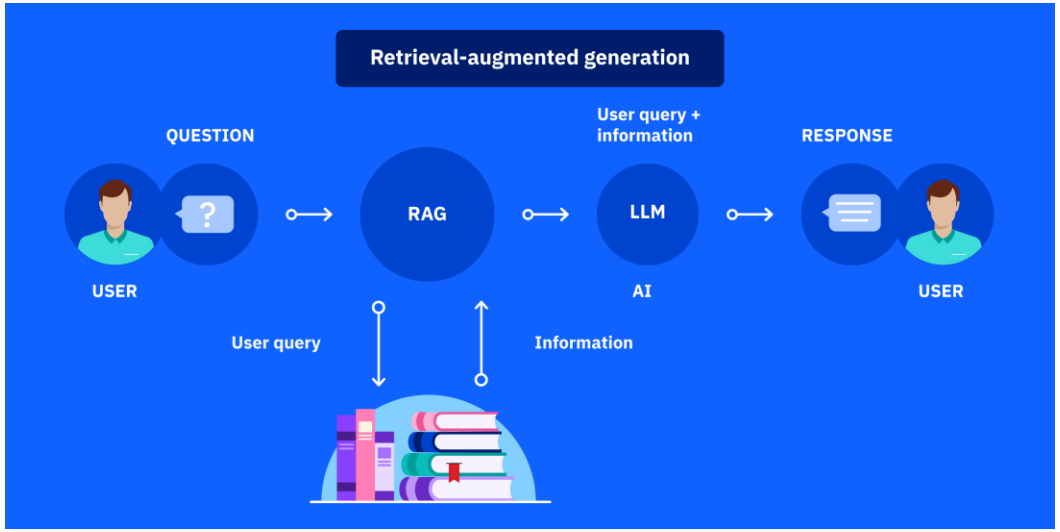
Objective

- What is Retrieval Augmented Generation
- Practical Use cases of RAG
- Activity: Identify key use cases for RAG

What is Retrieval Augmented Generation

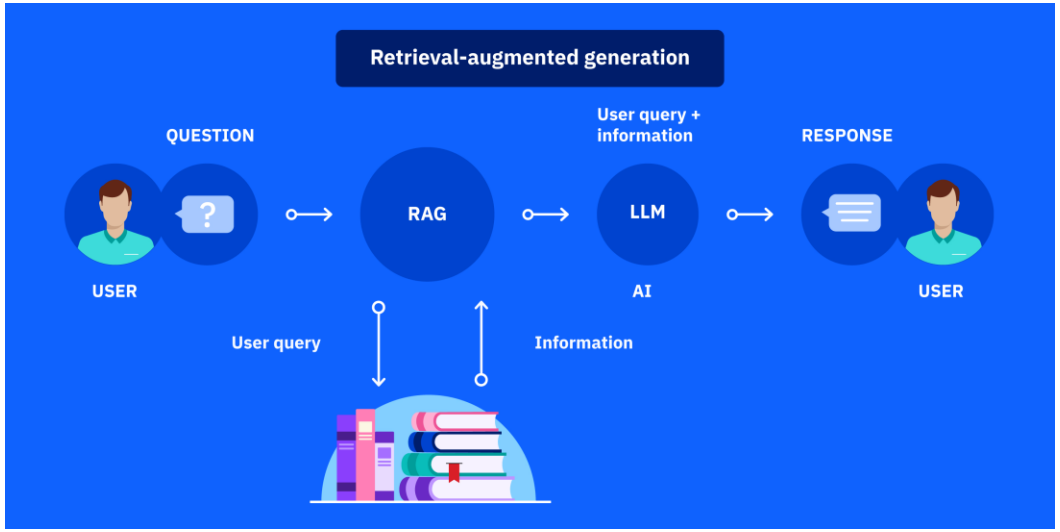
- **Retrieval augmented generation (RAG)** is an architecture for optimizing the performance of an **artificial intelligence (AI)** model by connecting it with external knowledge bases.

- RAG helps **large language models (LLMs)** deliver more relevant responses at a higher quality.

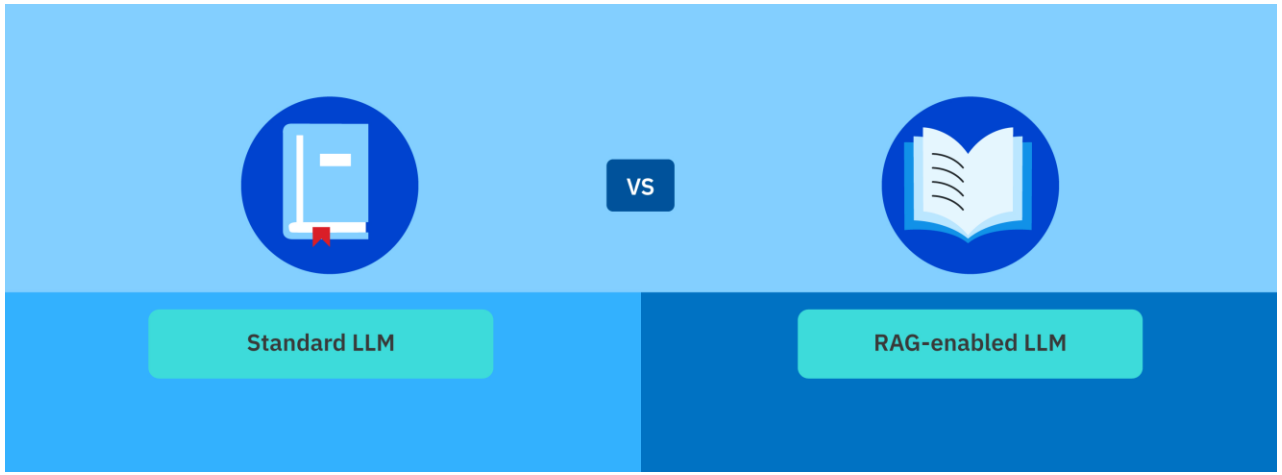


What is Retrieval Augmented Generation

- RAG is **an architecture for optimizing** the performance of **LLMs** by connecting them to external knowledge bases. RAG helps **LLMs deliver** more relevant responses at a higher quality



What is Retrieval Augmented Generation – Standard LLM VS RAG LLM



A **standard LLM** is like a student taking an exam without access to textbooks. While answering the exam questions, these students must rely only on what they have learned. Similarly, to answer user queries, a **standard LLM** can rely only on **its training data.** If it doesn't know specific details, it might guess or provide incomplete answers.

A **RAG-enabled LLM** or a **RAG system** is like a student taking an open-book exam. These students can refer to textbooks during the exam to find answers they don't remember. Similarly, a **RAG system needn't rely only on its training data**. It can look up **external sources** before responding, generating more accurate and up-to-date responses.

Objective

- What is Retrieval Augmented Generation
- Practical Use cases of RAG
- Activity: Identify key use cases for RAG

Practical use cases for RAG - Question answering

- AI-supported question answering generates responses based on a **fixed dataset** it was trained on. While helpful, this approach can sometimes lead to inadequate answers.

Practical use cases for RAG - Question answering

RAG systems can retrieve the latest and most relevant information from sources, such as **public databases**, **websites**, or **company documents**, before generating an answer.

Keeps the **AI responses** are current and reliable

Enables the **AI model to adapt** dynamically to new information and handle complex queries

An **e-commerce portal** uses a **RAG-powered chatbot** to search up-to-date website and product information, providing accurate responses to customer queries in real time

Practical use cases for RAG - Research augmentation

- AI models augment the **research process**, they might not be completely reliable because they extract responses from their **training data**, which might not reflect recent developments in the field.

Practical use cases for RAG - Research augmentation

RAG systems retrieve the most **current and relevant information** from credible external sources, including **journals, market reports, and databases**, before generating insights. This maintains research accuracy with the latest evidence.

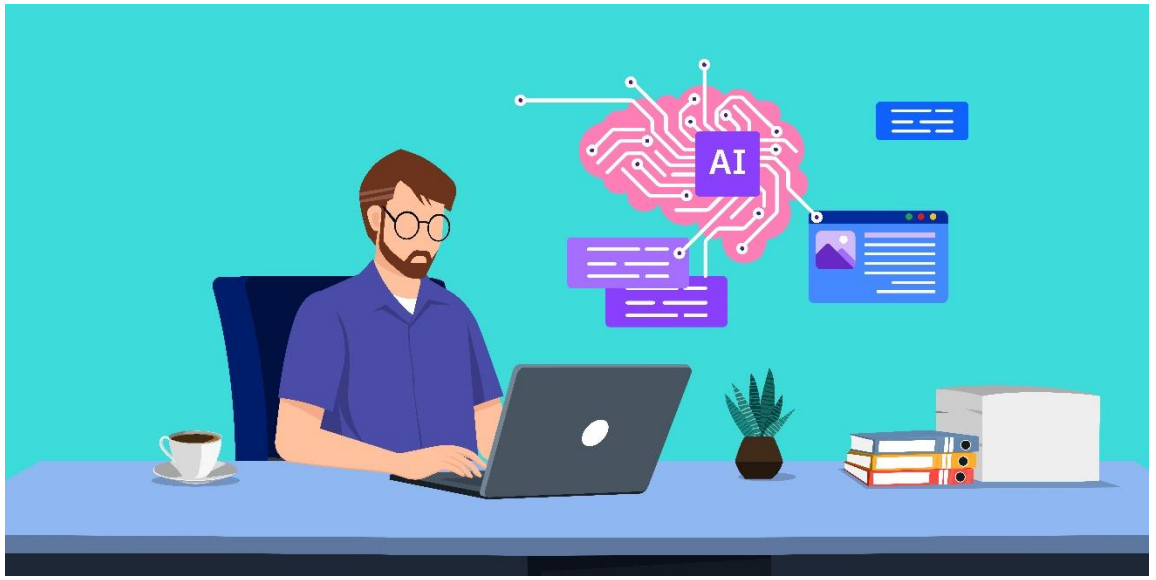
Provides access to **up-to-date** information sources

Combines **historical data** with **recent trends** to inform research findings

Financial professionals at a stock trading firm use a **RAG-enabled virtual assistant** to research the **latest stock reports, economic forecasts, and market trends**. This enables them to make timely and relevant investment decisions.

Practical use cases for RAG - Content generation

- **RAG systems** enhance content generation by retrieving real-time information from **external sources**, such as **news feeds**, **websites, or databases, before generating content**. In addition, they can include citations for users to verify the credibility of sources.



Practical use cases for RAG - Content generation

Improves the **factual accuracy** and timeliness of generated content

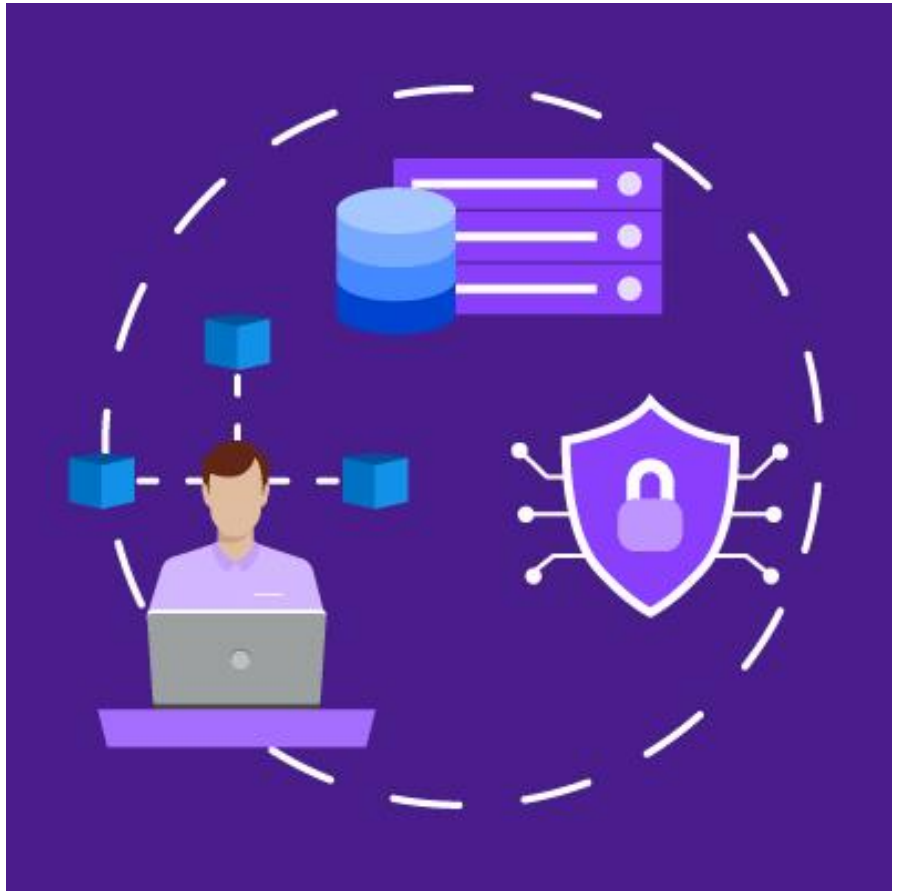
Enhances the **overall credibility** of content by including credible citations

A leading media channel uses a **RAG system** to **generate news articles** that include **real-time information**, updated statistics, and direct citations from credible sources.

This improves the factual accuracy of the channel's content and boosts its overall reliability and trustworthiness among its subscribers.

Practical use cases for RAG - Domain-specific and proprietary data integration

- RAG systems retrieve information from proprietary sources, such as internal knowledge bases, company documents, or industry-specific repositories, without embedding it with training data.



Practical use cases for RAG - Domain-specific and proprietary data integration

Delivers tailored and nuanced responses based on **up-to-date** internal and industry-specific data

Eliminates the need for proprietary information to be shared in public domain by dynamically incorporating it

Professionals at a law firm use a RAG system to **find recent court rulings** and **regulations, ensuring** they advise clients based on the latest available data. This approach enables the **dynamic integration** of the latest **legal information, strengthening client trust.**

By integrating real-time extraction of **data with LLMs'** text generation capability, **RAG enables LLMs** to deliver timely, accurate, and context-rich responses

Objective

- What is Retrieval Augmented Generation
- Practical Use cases of RAG
- Activity: Identify key use cases for RAG

Activity: Identify key use cases for RAG by using below color blocks

An internet provider implements a RAG-powered system to provide customers with real-time details on the latest service plans and promotions.

Question answering

Research augmentation

An environmental consultancy uses RAG to aggregate the latest climate studies and regulatory updates.

Research augmentation

Domain-specific and proprietary data integration

A marketing agency uses a RAG-enabled tool to generate materials to incorporate current market trends and verified statistics in its blog posts.

Content generation

Question answering

A pharmaceutical firm uses a RAG system to merge internal production data with up-to-date industry benchmarks.

Domain-specific and proprietary data integration

Content generation

Topic will be covered in this session…

1. An overview of retrieval-augmented generation
2. Benefits of implementing RAG
3. How does RAG work?
4. Simulation: Enhancing a large language model using RAG

Benefits of implementing RAG

IBM SkillsBuild

edunet foundation

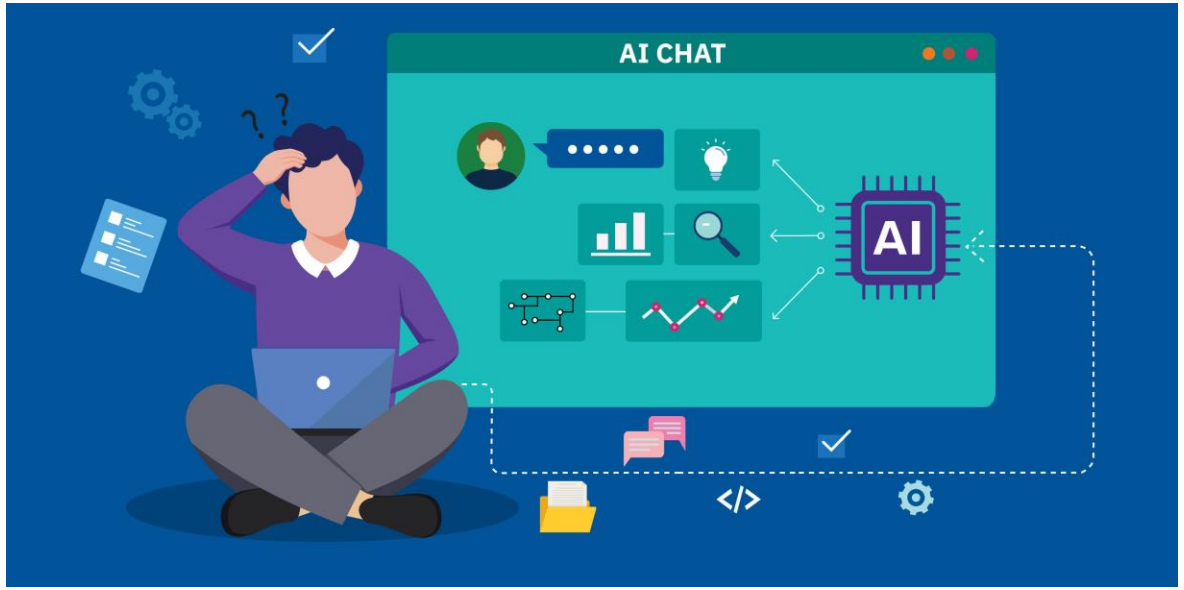
Objective

- Why do organizations use RAG?
- Benefits of implementing RAG
- Example RAG in action
- Activity: Explain the benefits of using RAG

Why do organizations use RAG?

- LLMs are trained on **extensive data sets**, which is one of their key strengths.

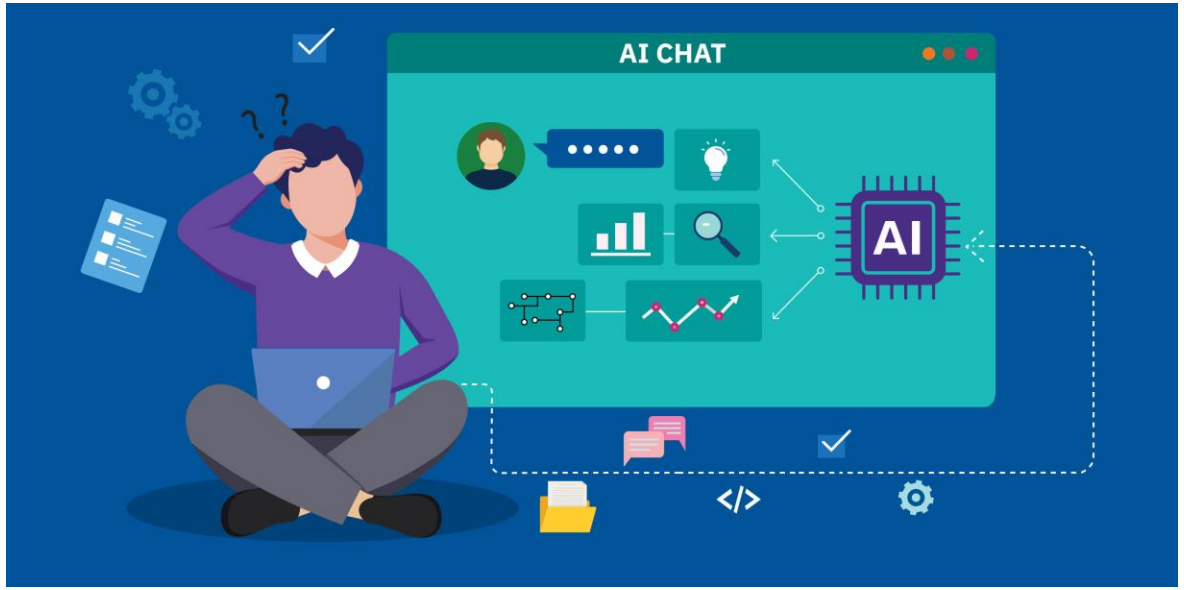
- The huge volume of training data enables LLMs to generate **human-like responses**, and this capability transforms how individuals and organizations interact with and use AI.



Why do organizations use RAG? – (Continued..)

- With standard LLMs, this is a challenge because their training data is static, meaning it is updated only until a certain point in time.

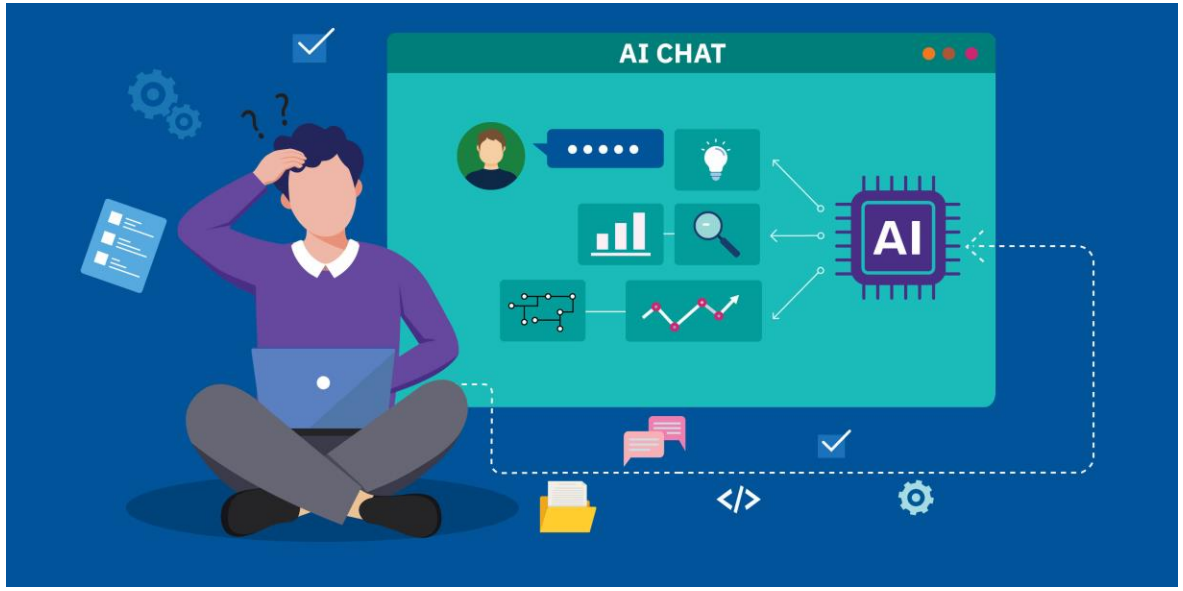
- **AI hallucination** is a phenomenon in which an **LLM** detects patterns or objects that do not exist or are not visible to humans, leading to inaccurate or meaningless outputs



Why do organizations use RAG? – (Continued..)

- **RAG is an alternative solution** that connects an **LLM to external**, authoritative data sources, expanding the model's knowledge base without the need for constant retraining.

- This integration allows the model to **retrieve real-time**, **relevant information**, enhancing the **accuracy** and relevance of its responses while minimizing the **risk of hallucination**.



What are AI hallucinations?

- **AI hallucination** is a phenomenon where, in a **large language model (LLM)** often **a generative AI chatbot** or **computer vision tool**, perceives patterns or objects that are nonexistent or imperceptible **to human observers**, creating outputs that are **nonsensical or altogether** inaccurate.

Objective

- Why do organizations use RAG?
- Benefits of implementing RAG
- Example RAG in action

Benefits of implementing RAG

- Cost-efficient AI implementation and AI scaling
- Real-time access to current data
- Enhanced accuracy and relevance
- Increased user trust
- Enhanced data security

Benefits of implementing RAG - Cost-efficient AI implementation and AI scaling

When implementing AI, most organizations start with **LLMs** trained on vast amounts of publicly **available data**.

RAG enhances **LLM**s by enabling them to retrieve **updated information** from internal, authoritative sources without retraining

For example, an insurance company uses **RAG** so that its virtual assistant can extract correct and **up-to-date** information about the company's products from **internal data** sources without **costly retraining**.

Benefits of implementing RAG - Real-time access to current data

The training data of **LLMs** is current up to a certain point in time, usually when the **model was trained.**

RAG allows **LLMs to integrate** with **online application programming interfaces (APIs), social media feeds**, and **search engines.**

For example, a **travel planning assistant** that uses **RAG** can pull real-time flight information from **airline websites**. This allows it to provide users with the **latest updates** on flight availability and pricing.

Benefits of implementing RAG - Enhanced accuracy and relevance

LLMs generate responses by identifying patterns in their **training data**. In the absence of relevant patterns in the underlying data, they **hallucinate** or provide inaccurate or fabricated information..

RAG reduces **hallucinations** by anchoring LLMs in factual, authoritative, and current data. Because **LLMs** using **RAG** retrieve information from reliable sources

For example, a **hardware manufacturer** uses **RAG** so that its **technical documentation** assistant retrieves the latest product specifications, technical bulletins, and service instructions from internal databases.

Benefits of implementing RAG - Increased user trust

For **LLM**-powered applications, such as **chatbots**, to be effective, users need to trust the responses they generate.

RAG improves user trust by allowing **LLMs** to include source citations in their responses. Using these verifiable sources, users can cross **check AI-generated** outputs to verify their accuracy and reliability.

For example, consider the **knowledge management system** of a **global corporation**. When a Europe-based employee asks about a recent update in **company policies**, the **RAG-powered system** provides the answer and includes citations with links to the original policy documents. This transparency allows the **employee** to verify the information easily, thereby **increasing trust in** the system.

Benefits of implementing RAG - Enhanced data security

As **cyber threats** escalate, ensuring data security and protecting **proprietary data** is a top priority for most organizations.

RAG improves data security by allowing **LLMs to access internal knowledge** sources without embedding them into the LLMs' training data.

For example, a **bank implementing** an AI assistant for customer inquiries uses **RAG to retrieve** loan policy details from secure internal databases instead of training the model on **confidential customer data**. This maintains compliance with data privacy regulations while providing accurate responses to users.

Objective

- Why do organizations use RAG?
- Benefits of implementing RAG
- Example RAG in action
- Activity: Explain the benefits of using RAG

Example: RAG in action

Eliza is the **HR manager** at a large multinational. She oversees the company's **AI-powered HR assistant**, which helps employees with questions about vacation policies, compensation, and benefits

However, as **company policies change**, employees start receiving outdated or incorrect responses, causing **confusion and frustration**. As a result, Eliza and her team are **inundated with employee queries** to resolve AI-generated **misinformation** and find it difficult to manage their workload.

Realizing that the **HR assistant** needs a better solution to address employee queries and ease the HR team's workload, the **company integrates RAG into its LLM**. Now, instead of relying only on static training data, the **assistant retrieves** the **latest policies** from internal HR databases before generating responses.

Example: RAG in action - explore the benefits Eliza's company derives by deploying RAG.

Cost-efficient AI scaling

By allowing **the HR assistant** to retrieve policy updates in **real time, RAG eliminates** the need for frequent retraining. As a result, **AI responses** remain relevant without requiring **extensive model adjustments** or resource-intensive **fine-tuning, reducing operational costs.**

Enhanced accuracy and relevance

By dynamically referencing current **HR documents** before generating responses, the **AI assistant** provides employees with **precise, fact-based, up-to-date answers** aligned with the **latest policies, reducing confusion** and **improving decision making**. This reduces the need to manually verify **AI-generated information**.

Improved data security

By retrieving external knowledge rather than **embedding sensitive company policies** directly into the AI model, **RAG keeps confidential HR information** protected. The AI assistant can access **relevant data** when needed without storing **sensitive content**, ensuring **compliance and security.**

Topic will be covered in this session…

1. An overview of retrieval-augmented generation
2. Benefits of implementing RAG
3. How does RAG work?
4. Simulation: Enhancing a large language model using RAG

How does RAG work?

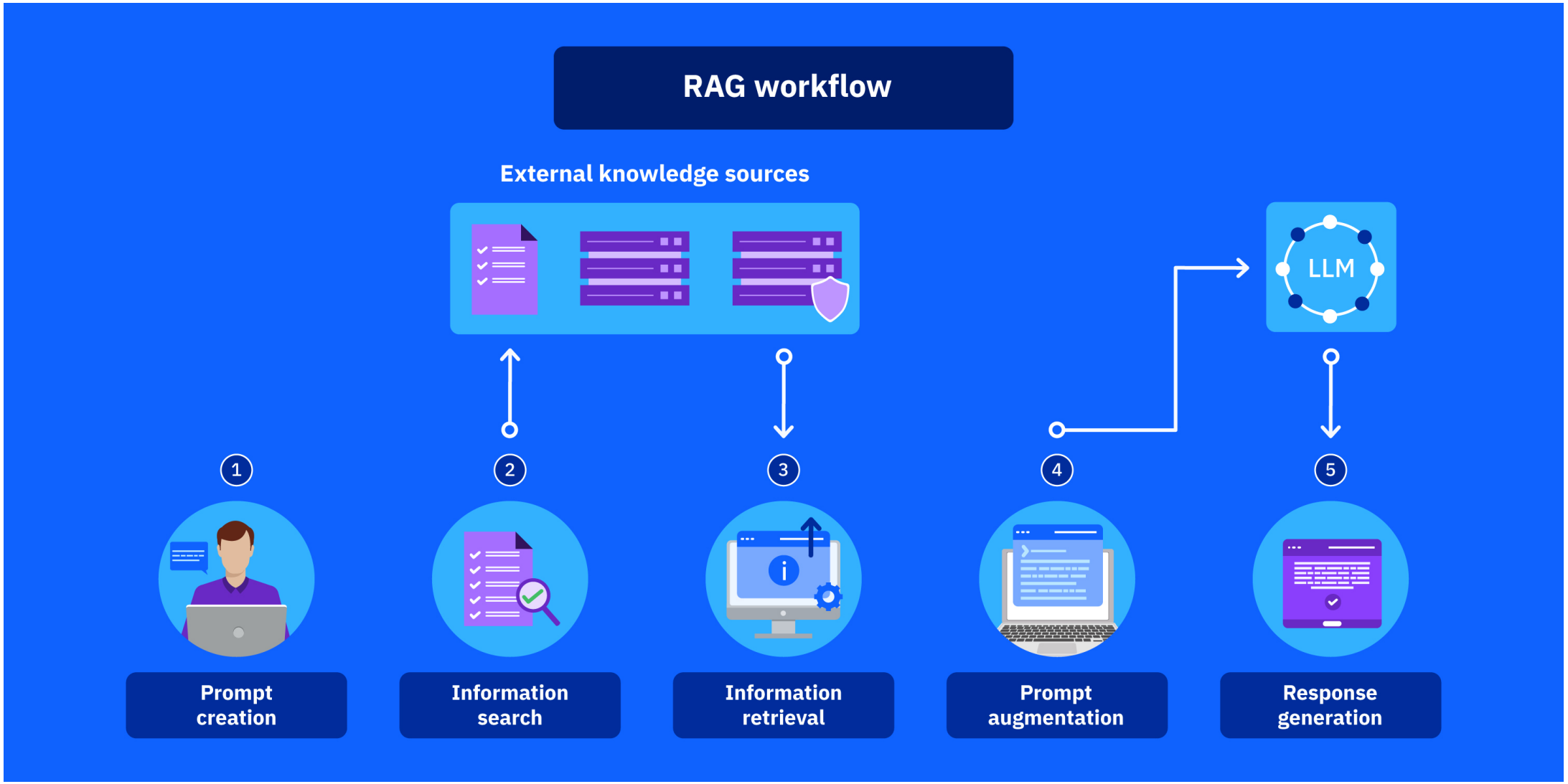
Objective

- The RAG workflow.
- Example: the RAG workflow.
- Activity: Describe the key steps in RAG workflow.

The RAG workflow.

- The RAG workflow consists of steps that combine information extracted from knowledge sources to produce more informed, precise, and relevant responses.

The RAG workflow - Steps



The RAG workflow - Steps

Step 1: Prompt creation

In the first step, the user submits a query, request, or instruction that requires a response. This input acts as a trigger that prompts RAG to capture the user input, determine the intent and scope of the query, and retrieve relevant information.

Step 2: Information search

In the second step, RAG searches external knowledge sources such as research papers, books, company databases, APIs, and web articles. This search allows the response to the user's query to go beyond the LLM's training dataset. Instead, it is based on the most up-to-date and relevant information from external sources.

Step 3: Information retrieval

In the third step, RAG filters and extracts the most relevant and credible information, ensuring that only high-quality data aligned with the user's query is selected.

The RAG workflow - Steps

Step 4: Prompt augmentation

In the fourth step, RAG enhances the original query by integrating the retrieved information. This provides the AI model with additional context to generate a more accurate and detailed answer aligned with the user's intent**.**

.

Step 5: Response generation

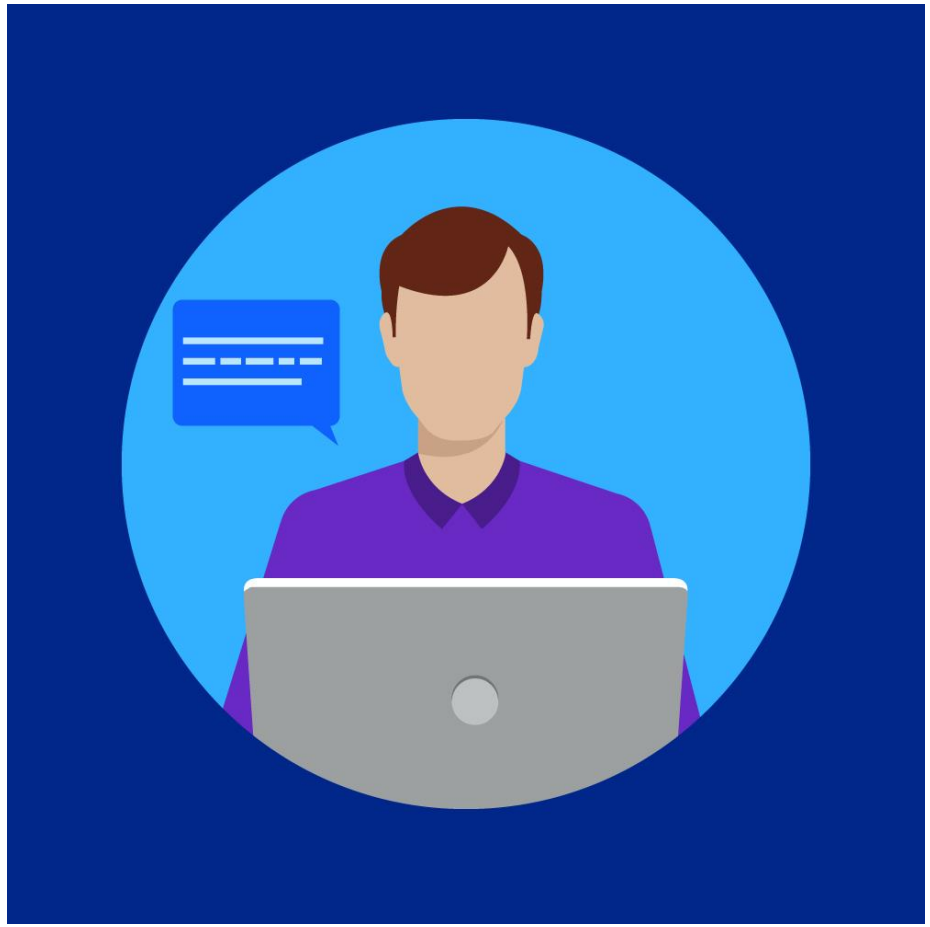
In the final step, RAG uses the enhanced information to generate a clear, accurate, and actionable response, which is then delivered to the user.

Objective

- The RAG workflow.
- Example: the RAG workflow.
- Activity: Describe the key steps in RAG workflow.

Example: The RAG workflow - Prompt creation

- You ask your virtual travel assistant, **“What are the best budget-friendly hotels** in Dubai for this weekend?” **RAG** captures your query and identifies three key parameters that define the intent and scope of your query: “**budget-friendly”, “Dubai”,** and **“weekend dates”.**



Example: The RAG workflow - Information search

- **RAG** searches external knowledge sources such as **travel portals, hotel booking sites, review aggregators, and travel blogs**. The search enables the response to include the most up-to-date information, beyond the static **training data of the LLM.**

Example: The RAG workflow - Information retrieval

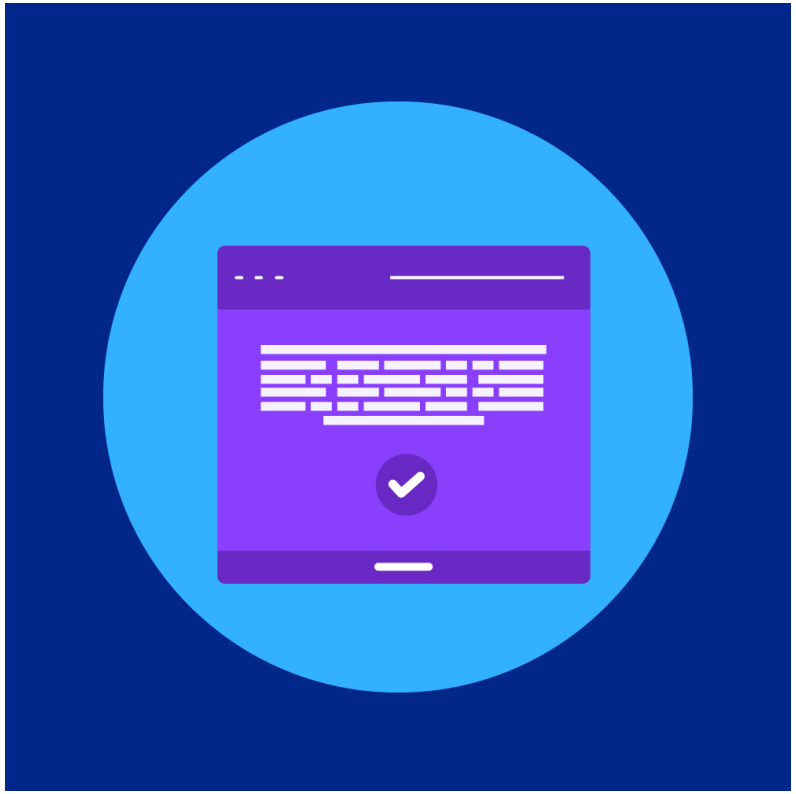
- From the vast pool of available data, **RAG filters** and extracts the most relevant and **credible information**, such as current hotel prices, availability on your **travel dates, ratings, and guest reviews,** ensuring it aligns with your query.

Example: The RAG workflow - Prompt augmentation

- The **retrieved information** is then integrated with your . RAG enhances your prompt with specifics such as "**Hotel X offers a 4-star experience at $110** per night with excellent reviews, while Hotel Y is a more **budget-friendly** option at **$90 per** night but with slightly lower customer ratings." The enriched prompt provides additional context to the **LLM.**

Example: The RAG workflow - Response generation

Finally, the LLM uses the enhanced **prompt to generate** a clear, accurate, and actionable output. It provides a concise recommendation with **budget-friendly hotel options** and all essential details to help you make an informed decision.



Objective

- The RAG workflow.
- Example: the RAG workflow.
- Activity: Describe the key steps in RAG workflow.

Activity: Describe the key steps in RAG workflow by Color blocks

You submit the query, “What free concerts and community events are happening in my city this weekend?”

Prompt creation

Information search

The RAG system checks local event listings, social media feeds, and community bulletin boards to gather the most recent information on events happening in the city.

Information search

Prompt augmentation

Response generation

The RAG system filters and extracts the most relevant events, including their names, dates, venues, and descriptions.

Information retrieval

Prompt creation

The RAG system enriches the query by including specifics to provide additional context for a more accurate response.

Prompt augmentation

The app provides a concise list of free concerts and community events, with all the essential details to help you plan your weekend.

Response generation

Information retrieval

Summary

- RAG systems like RAG help AI get the latest and most accurate information by connecting to outside sources.

- RAG lets AI learn new things without needing constant retraining, which saves money, keeps user data safe, and builds trust by giving better and more reliable results.

- The RAG process involves creating a prompt, searching for info, adding that info to the prompt, and then giving an answer.


Q&A

IBM **SkillsBuild**

edunet foundation