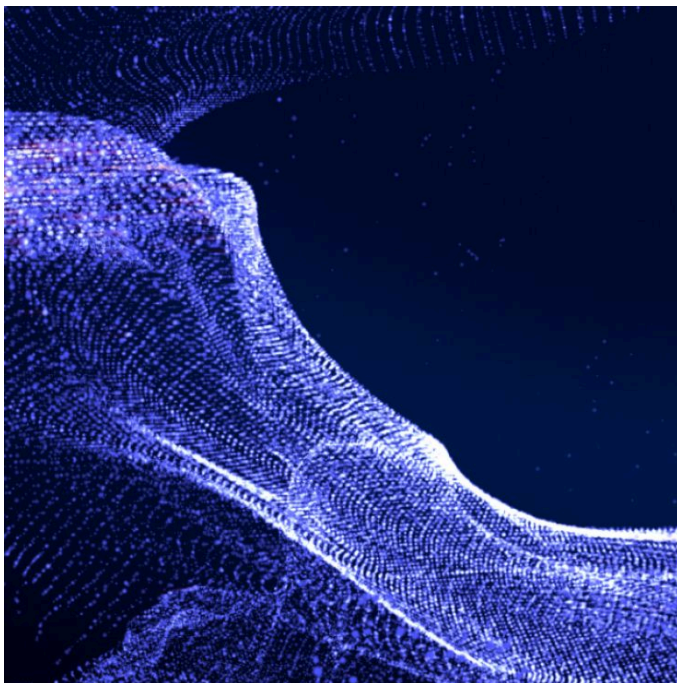


Artificial Intelligence

Compute and servers

IT &amp;

# What is RAG (retrieval augmented generation)?



## About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your [cookie preferences](#) options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

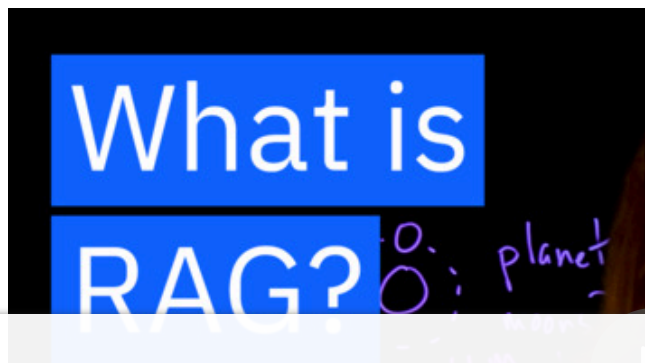
[Accept all](#)[More options](#)

# What is RAG (retrieval generation)?

Retrieval augmented generation (RAG) improves the performance of an [artificial intelligence \(AI\)](#) model by providing it with external knowledge bases. RAG [\(LLMs\)](#) deliver more relevant responses.

[Generative AI](#) (gen AI) models are trained on large datasets to generate outputs. However, training datasets are limited to the data the AI developer can access—public domain data and other publicly accessible data.

RAG allows generative AI models to access a wide range of internal organizational data, scholarly journal articles, and other relevant information into the generation process. [Natural language processing \(NLP\)](#) tools can create more accurate responses through further training.



## About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your

options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

# What are the benefits of RAG?

RAG empowers organizations to avoid high resource costs by applying foundation models to domain-specific use cases. Enterprise RAG leverages a [machine learning](#) model's knowledge base so that it can answer questions about domain-specific data.

The primary benefits of RAG include:

- Cost-efficient AI implementation and [AI savings](#)
- Access to current domain-specific data
- Lower risk of [AI hallucinations](#)
- Increased user trust
- Expanded use cases
- Enhanced developer control and model management
- Greater data security

## Cost-efficient AI implementation

When implementing AI, most organizations find that [machine learning](#) models that serve as the basis for their applications are often generalized. Foundation models typically have generalized knowledge based on

available training data, such as internet content.



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

# Access to current and dynamic data

Generative AI models have a *knowledge cutoff* last updated. As a model ages further past its time, RAG systems connect models with suppliers to incorporate up-to-date information into generated responses.

Enterprises use RAG to equip models with specific customer data, authoritative research and other relevant information.

RAG models can also connect to the internet (APIs) and gain access to real-time social media feeds to gain a better understanding of market sentiment. Meanwhile, RAG engines can lead to more accurate responses by incorporating relevant information into the text-generation process.

## Lower risk of AI hallucinations

Generative AI models such as OpenAI's GPT-4 use patterns from training data to predict the most likely next word in a sequence. However, these models detect patterns that don't exist. A hallucination is when a model presents incorrect or made-up information.

RAG anchors LLMs in specific knowledge back to a source of data. Compared to a generative model operating on its own, RAG models tend to provide more accurate answers within their knowledge base. By anchoring RAG, it can reduce the risk of hallucinations, it can



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

information. Corporate data storage is often a with citations point users directly toward the

## Expanded use cases

Access to more data means that one model c. Enterprises can optimize models and gain mo knowledge bases, in turn expanding the conte results.

By combining generative AI with retrieval sys information from multiple data sources in res

## Enhanced developer con maintenance

Modern organizations constantly process mas What if AI projections to employee turnover and and data storage is paramount for strong RAG

At the same time, developers and [data scient](#) models have access at any time. Repositionin a task of adjusting its external knowledge sou If fine-tuning is needed, developers can prior model's data sources.

## Greater data security



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

# RAG use cases

RAG systems essentially enable users to query the data-powered [question-answering](#) abilities across a range of use cases, including:

- Specialized chatbots and virtual assistants
- Research
- Content generation
- Market analysis and product development
- Knowledge engines
- Recommendation services

## Specialized chatbots and virtual assistants

Enterprises wanting to [automate](#) customer support can leverage the specialized knowledge needed to augment generative AI models into internal data to equip customer service agents with information about a company's products, services and policies.

The same principle applies to AI avatars and digital assistants. By combining an underlying model with the user's personal data, RAG systems can provide a more customized user experience.

## Research



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your

options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

# Market analysis and product development

Business leaders can consult social media trending, breaking news and other online sources to better understand market trends. Product managers can reference customer feedback to inform future development choices.

## Knowledge engines

RAG systems can empower employees with information for onboarding processes, faster HR support and field service. These are just a few ways businesses can use RAG.

## Recommendation services

By analyzing previous user behavior and company data, recommendation systems power more accurate recommendations. A content delivery service can both use RAG to

## How does RAG work?

RAG works by combining information retrieval with generative AI to produce more authoritative content. RAG systems use a retrieval component to search for relevant context to a user prompt before generating a response.



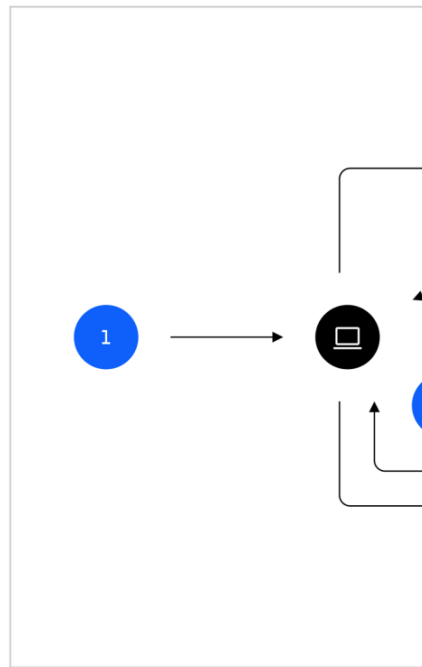
Standard LLMs source information from their training data. RAG adds a retrieval component to the AI workflow, gathering relevant information from external sources to enhance responses.

### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).



1. The user submits a prompt.
2. The information **retrieval** model queries the
3. Relevant information is returned from the
4. The RAG system engineers an **augmented** from the retrieved data.
5. The LLM **generates** an output and returns

This process showcases how RAG gets its narrow knowledge base, *augments* the prompt with a



#### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).



# Components of a RAG

RAG systems contain four primary components:

- The **knowledge base**: The external data repository.
- The **retriever**: An AI model that searches the knowledge base for relevant information.
- The **integration layer**: The portion of the RAG system that integrates the retrieved information with the user's query.
- The **generator**: A generative AI model that generates a response based on the integrated information.

Other components might include a *ranker*, which ranks the retrieved information, and an *output handler*, which formats the generated response.

## The knowledge base

The first stage in constructing a RAG system is identifying the external data repository. This repository can contain data from a variety of sources, including books, guides, websites, audio files and more. Much of this data is unlabeled, which means that it hasn't yet been categorized.

RAG systems use a process called embedding, which converts text into multidimensional mathematical representations called vectors. The embedding process arranges these vectors in a space where similar items are judged to be closer in relevance to each other.



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

Chunk size is an important hyperparameter for large, the data points can become too general for user queries. But if chunks are too small, the

## The retriever

Vectorizing the data prepares the knowledge that identifies points in the database that are [machine learning algorithms](#) can query massive information, reducing latency as compared to

The information retrieval model transforms the searches the knowledge base for similar embeddings in the knowledge base.

## The integration layer

The integration layer is the center of the RAG system, passing data around the network. With the adapter system creates a new prompt for the LLM combining the user query plus the enhanced context returned

RAG systems employ various [prompt engineering](#) techniques for prompt creation and help the LLM return the best results. Orchestration frameworks such as the open source [watsonx Orchestrate™](#) govern the overall function

## The generator



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

# RAG vs. Fine-Tuning Which AI Strategy Should You Use?

RAG vs. Fine Tuning (8:57 min)

The [difference between RAG and fine-tuning](#) is that RAG lets an LLM query an external data source while fine-tuning trains an LLM on domain-specific data. Both have the same general goal: to make an LLM perform better in a specified domain.

RAG and fine-tuning are often contrasted but can be used in tandem. Fine-tuning increases a model's familiarity with the intended domain and output requirements, while RAG assists the model in generating relevant, high-quality outputs.



## About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

Ebook

## Generative AI + ML for the enterprise

Learn how to successfully implement AI technologies for significant competitive advantage.



## Related solutions

I B M	I B M ®	A r
<div data-bbox="136 1528 479 1560"><b>About cookies on this site</b></div> <div data-bbox="136 1566 406 1955"><p>Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.</p></div> <div data-bbox="451 1566 771 1955"><p>For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's <a href="#">privacy statement</a>.</p></div> <div data-bbox="764 1566 980 1881"><p>To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed <a href="#">here</a>.</p></div> <div data-bbox="1502 1472 1534 1507">X</div>		

i  
TM

L  
e  
a  
r  
n  
m  
o  
r  
e  
a  
b  
o  
u  
t  
a  
n  
e  
x  
t  
-  
g  
e  
n  
e  
r

a  
t  
a  
TM

P  
u  
t  
y  
o  
u  
r  
d  
a  
t  
a  
t  
o  
w  
o  
r  
k  
,  
w  
h  
e  
r  
e

e

ll  
i  
g  
e  
n  
c  
e  
(  
A  
I  
)  
c  
o  
n  
s  
u  
l  
t  
i  
n  
g  
s  
e  
r  
v

a

About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

ices

https://www.ibm.com/think/topics/retrieval-augmented-generation

13/16

fine how you work with AI for business.

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).

I m o d el s.  Ex pl or e w at so nx .ai →	ti c s.  E x pl or e w at s o n x. d at a →	o n s ul ti n g s er vi c e s →
---	--	---

## Resources



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).



To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).





The recipe for RAG: How cloud services enable generative AI outcomes across industries

Article

[Learn more →](#)



Generative AI + ML for the enterprise

Guide

[Register and download →](#)

## Take the next step

Train, validate, tune and deploy generative AI, foundation models and machine learning capabilities with IBM watsonx.ai, a next-generation enterprise studio for AI builders. Build AI applications in a fraction of the time with a fraction of the data.

[Explore watsonx.ai](#)



[Book a live demo](#)



### About cookies on this site

Our websites require some cookies to function properly (required). In addition, other cookies may be used with your consent to analyze site usage, improve the user experience and for advertising.

For more information, please review your options. By visiting our website, you agree to our processing of information as described in IBM's [privacy statement](#).

To provide a smooth navigation, your cookie preferences will be shared across the IBM web domains listed [here](#).