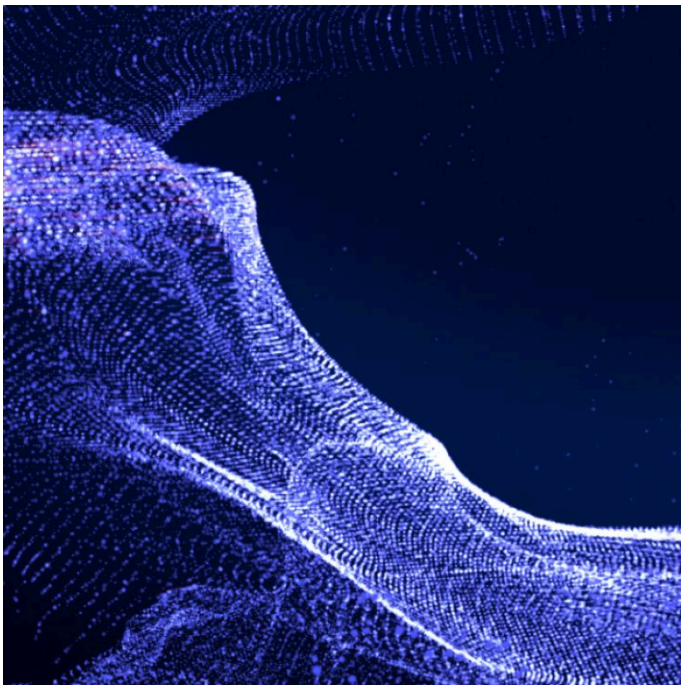
Artificial Intelligence Compute and servers IT a

What is RAG (retrieva generation)?



What is RAG (retriev generation)?

Retrieval augmented generation (RA the performance of an artificial intell with external knowledge bases. RAG (LLMs) deliver more relevant respon

Generative AI (gen AI) models are trained on to generate outputs. However, training datase the AI developer can access—public domain v and other publicly accessible data.

RAG allows generative AI models to access a internal organizational data, scholarly journal relevant information into the generation proc processing (NLP) tools can create more accur further training.

What are the benefi

RAG empowers organizations to avoid high re
models to domain-specific use cases. Enterpı
machine learning model's knowledge base so

The primary benefits of RAG include:

- Cost-efficient AI implementation and AI s
- Access to current domain-specific data
- Lower risk of AI hallucinations
- Increased user trust
- Expanded use cases
- Enhanced developer control and model m
- Greater data security

Cost-efficient AI implem

When implementing AI, most organizations fi
learning models that serve as the basis for th
Foundation models typically have generalizec
available training data, such as internet conte

Access to current and dc

Generative AI models have a *knowledge cutof* last updated. As a model ages further past its time. RAG systems connect models with supp incorporate up-to-date information into gene

Enterprises use RAG to equip models with sp customer data, authoritative research and oth

RAG models can also connect to the internet (APIs) and gain access to real-time social me understanding of market sentiment. Meanwhi engines can lead to more accurate responses information into the text-generation process.

Lower risk of AI hallucina

Generative AI models such as OpenAI's GPT using those patterns to predict the most likely models detect patterns that don't exist. A hal models present incorrect or made-up informa

RAG anchors LLMs in specific knowledge bacl data. Compared to a generative model operat tend to provide more accurate answers withir RAG can reduce the risk of hallucinations, it c

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 syst
information from multiple data sources in res

Enhanced developer con maintenance

Modern organizations constantly process mas
make RAG 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

RAG use cases

RAG systems essentially enable users to quer
The data-powered question-answering abiliti
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

Enterprises wanting to automate customer su
the specialized knowledge needed to adequa
models into internal data to equip customer s
about a company's products, services and po

The same principle applies to AI avatars and
underlying model with the user's personal da
provides a more customized user experience.

Research

Market analysis and proc

Business leaders can consult social media tre
breaking news and other online sources to be
product managers can reference customer fe
future development choices.

Knowledge engines

RAG systems can empower employees with i
onboarding processes, faster HR support and
field are just a few ways businesses can use F

Recommendation servic

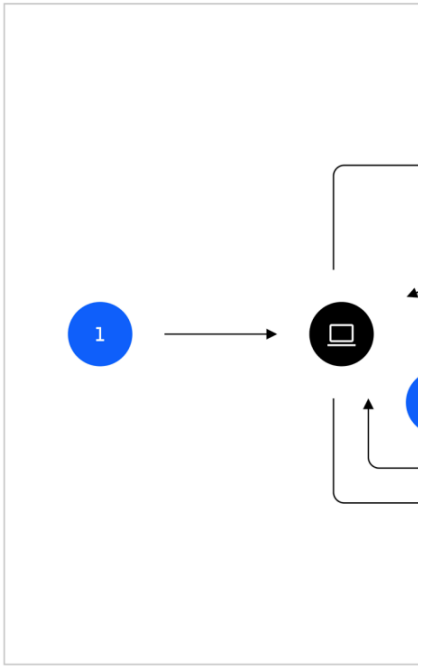
By analyzing previous user behavior and com
systems power more accurate recommendati
content delivery service can both use RAG to

How does RAG work

RAG works by combining information retrieva
produce more authoritative content. RAG syst
context to a user prompt before generating a

Standard LLMs source information from their
retrieval component to the AI workflow, gath
the generative AI model to enhance respo

RAG systems follow a five-stage process:



1. The user submits a prompt.
2. The information **retrieval** model queries tl
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 nar knowledge base, *augments* the prompt with a

Components of a RA

RAG systems contain four primary componen

- The **knowledge base**: The external data re
- The **retriever**: An AI model that searches
- The **integration layer**: The portion of the F functioning.
- The **generator**: A generative AI model tha and retrieved data.

Other components might include a *ranker*, wh and an *output handler*, which formats the gen

The knowledge base

The first stage in constructing a RAG system i external data repository can contain data fror guides, websites, audio files and more. Much means that it hasn't yet been labeled.

RAG systems use a process called embeddin representations called vectors. The embeddi multidimensional mathematical space, arran judged to be closer in relevance to each othe

Chunk size is an important hyperparameter fc
large, the data points can become too genera
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 massi
information, reducing latency as compared to

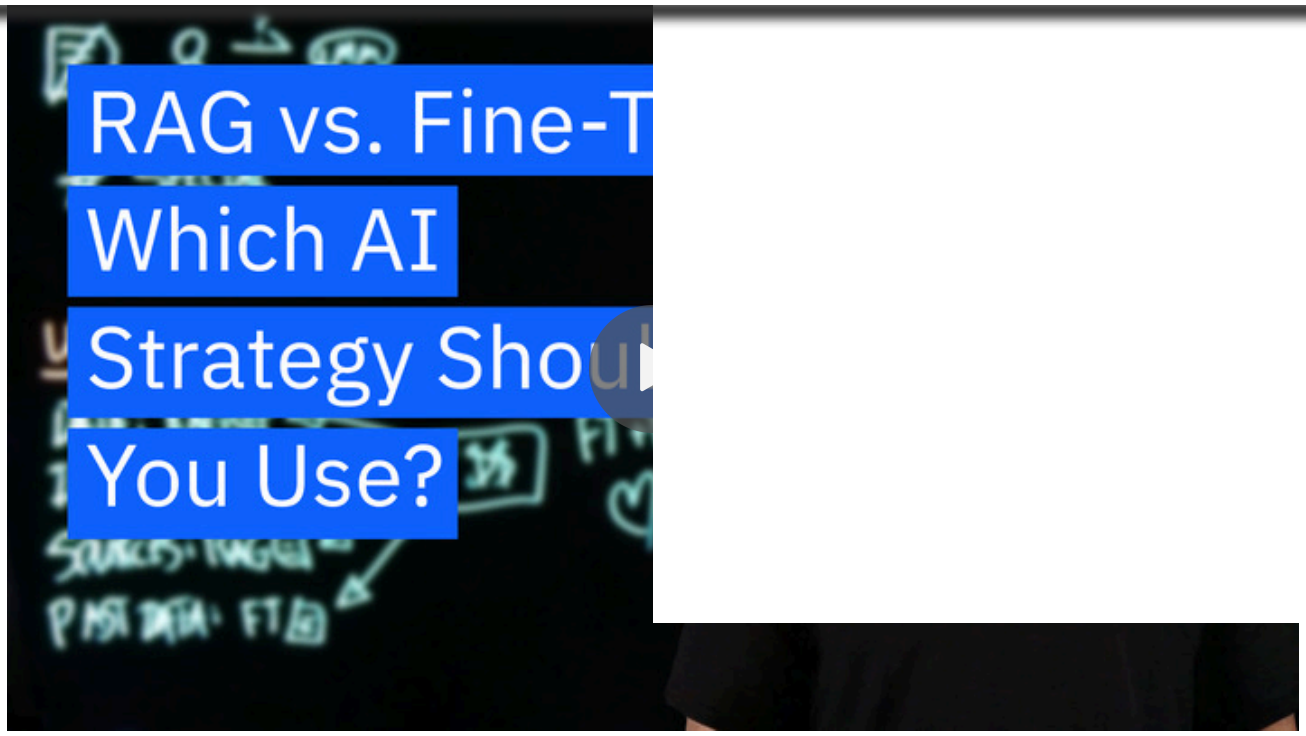
The information retrieval model transforms th
searches the knowledge base for similar emb
the knowledge base.

The integration layer

The integration layer is the center of the RAG
passing data around the network. With the ad
system creates a new prompt for the LLM con
user query plus the enhanced context returne

RAG systems employ various prompt enginee
prompt creation and help the LLM return the l
orchestration frameworks such as the open s
watsonx Orchestrate™ govern the overall func

The generator



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.

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 ® w a s o n A

A r t i f i c i a l i n t

i™

Learn more about the next-gener

ata™

Put your data to work, wher

elligence (AI) consulting servi

s
e
st
u
di
o
f
o
r
A
I
b
ui
ld
e
rs
t
o
tr
ai
n,
v
al
id
a
t
e,
t
u

it
h
t
h
e
o
p
e
n
,
h
y
b
ri
d
d
a
t
a
l
a
k
e
h
o
u
s
e
f

fi
n
e
h
o
w
y
o
u
w
o
r
k
w
it
h
A
I
f
o
r
b
u
si
n
e
s
s.

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

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 →