In This Article

Generative artificial intelligence (AI) excels at creating text response
(LLMs) where the AI is trained on a massive number of data points.
text is often easy to read and provides detailed responses that are
asked of the software, often called prompts.

The bad news is that the information used to generate the respons
train the AI, often a generalized LLM. The LLM's data may be weeks
corporate AI chatbot may not include specific information about the
That can lead to incorrect responses that erode confidence in the te
employees.

What Is Retrieval-Augmented Generation (RAG)?

That's where retrieval-augmented generation (RAG) comes in. RAG provides a way to optimize the output of an LLM with targeted information without modifying the underlying model itself; that targeted information can be more up-to-date than the LLM as well as specific to a particular organization and industry. That means the generative AI system can provide more contextually appropriate answers to prompts as well as base those answers on extremely current data.

RAG first came to the attention of generative AI developers after the publication of "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," a 2020 paper published by Patrick Lewis and a team at Facebook AI Research. The RAG concept has been embraced by many academic and industry researchers, who see it as a way to significantly improve the value of generative AI systems.

Retrieval-Augmented Generation Explained

Consider a sports league that wants fans and the media to be able to use chat to access its data and answer questions about players, teams, the sport's history and rules, and current stats and standings. A generalized LLM could answer questions about the history and rules or perhaps describe a particular team's stadium. It wouldn't be able to discuss last night's game or provide current information about a particular athlete's injury because the LLM wouldn't have that information—and given that an LLM takes significant computing horsepower to retrain, it isn't feasible to keep the model current.

In addition to the large, fairly static LLM, the sports league owns or can access many other information sources, including databases, data warehouses, documents containing player bios, and news feeds that discuss each game in depth. RAG lets the generative AI ingest this information. Now, the chat can provide information that's more timely, more contextually appropriate, and more accurate.

Simply put, RAG helps LLMs give better answers.

Key Takeaways

- RAG is a relatively new artificial intelligence technique that can improve the quality of generative AI by allowing large language model (LLMs) to tap additional data resources without retraining.
- RAG models build knowledge repositories based on the organization's own data, and the repositories can be continually updated to help the generative AI provide timely, contextual answers.
- Chatbots and other conversational systems that use natural lang from RAG and generative AI.
- Implementing RAG requires technologies such as vector databas new data, and searches against that data to feed into the LLM.

How Does Retrieval-Augmented Generati

Consider all the information that an organization has—the structured databases, the unstructured PDFs and other documents, the blogs, the news feeds, the chat transcripts from past customer service sessions. In RAG, this vast quantity of dynamic data is translated into a common format and stored in a knowledge library that's accessible to the generative AI system.

The data in that knowledge library is then processed into numerical representations using a special type of algorithm called an embedded language model and stored in a vector database, which can be quickly searched and used to retrieve the correct contextual information.

RAG and Large Language Models (LLMs)

Now, say an end user sends the generative AI system a specific prompt, for example, "Where will tonight's game be played, who are the starting players, and what are reporters saying about the matchup?" The query is transformed into a vector and used to query the vector database, which retrieves information relevant to that question's context. That contextual information plus the original prompt are then fed into the LLM, which generates a text response based on both its somewhat out-of-date generalized knowledge and the extremely timely contextual information.

Interestingly, while the process of training the generalized LLM is time-consuming and costly, updates to the RAG model are just the opposite. New data can be loaded into the embedded language model and translated into vectors on a continuous, incremental basis. In fact, the answers from the entire generative AI system can be fed back into the RAG model, improving its performance and accuracy, because, in effect, it knows how it has already answered a similar question.

An additional benefit of RAG is that by using the vector database, the generative AI can provide the specific source of data cited in its answer—something LLMs can't do. Therefore, if there's an inaccuracy in the generative AI's output, the document that contains that erroneous information can be quickly identified and corrected, and then the corrected information can be fed into the vector database.

In short, RAG provides timeliness, context, and accuracy grounded in evidence to generative AI, going beyond what the LLM itself can provide.

Retrieval-Augmented Generation vs. Semantic Search

RAG isn't the only technique used to improve the accuracy of LLM-based generative AI. Another technique is semantic search, which helps the AI system narrow down the meaning of a query by seeking deep understanding of the specific words and phrases in the prompt.

Traditional search is focused on keywords. For example, a basic query asking about the tree species native
to France might search the AI system's database using "trees" and "
contains both keywords—but the system might not truly comprehe
therefore may retrieve too much information, too little, or even the
search might also miss information because the keyword search is
might be missed, even though they're in France, because that keyw

Semantic search goes beyond keyword search by determining the r
documents and using that meaning to retrieve more accurate resul
RAG.

Using RAG in Chat Applications

When a person wants an instant answer to a question, it's hard to beat the immediacy and usability of a chatbot. Most bots are trained on a finite number of intents—that is, the customer's desired tasks or outcomes—and they respond to those intents. RAG capabilities can make current bots better by allowing the AI system to provide natural language answers to questions that aren't in the intent list.

The "ask a question, get an answer" paradigm makes chatbots a perfect use case for generative AI, for many reasons. Questions often require specific context to generate an accurate answer, and given that chatbot users' expectations about relevance and accuracy are often high, it's clear how RAG techniques apply. In fact, for many organizations, chatbots may indeed be the starting point for RAG and generative AI use.

Questions often require specific context to deliver an accurate answer. Customer queries about a newly introduced product, for example, aren't useful if the data pertains to the previous model and may in fact be misleading. And a hiker who wants to know if a park is open this Sunday expects timely, accurate information about that specific park on that specific date.

Benefits of Retrieval-Augmented Generation

RAG techniques can be used to improve the quality of a generative AI system's responses to prompts, beyond what an LLM alone can deliver. Benefits include the following:

- The RAG has access to information that may be fresher than the data used to train the LLM.
- Data in the RAG's knowledge repository can be continually updated without incurring significant costs.
- The RAG's knowledge repository can contain data that's more contextual than the data in a generalized LLM.
- The source of the information in the RAG's vector database can be identified. And because the data sources are known, incorrect information in the RAG can be corrected or deleted.

Challenges of Retrieval-Augmented Generation

Because RAG is a relatively new technology, first proposed in 2020, best implement its information retrieval mechanisms in generative

- Improving organizational knowledge and understanding of RAG
- Increasing costs; while generative AI with RAG will be more expe own, this route is less costly than frequently retraining the LLM i
- Determining how to best model the structured and unstructured and vector database
- Developing requirements for a process to incrementally feed data into the RAG system

Putting processes in place to handle reports of inaccuracies and to correct or delete those information sources in the RAG system

Examples of Retrieval-Augmented Generation

There are many possible examples of generative AI augmented by RAG.

Cohere, a leader in the field of generative AI and RAG, has written about a chatbot that can provide contextual information about a vacation rental in the Canary Islands, including fact-based answers about beach accessibility, lifeguards on nearby beaches, and the availability of volleyball courts within walking distance.

Oracle has described other use cases for RAG, such as analyzing financial reports, assisting with gas and oil discovery, reviewing transcripts from call center customer exchanges, and searching medical databases for relevant research papers.

Future of Retrieval-Augmented Generation

Today, in the early phases of RAG, the technology is being used to provide timely, accurate, and contextual responses to queries. These use cases are appropriate to chatbots, email, text messaging, and other conversational applications.

In the future, possible directions for RAG technology would be to help generative AI take an appropriate action based on contextual information and user prompts. For example, a RAG-augmented AI system might identify the highest-rated beach vacation rental on the Canary Islands and then initiate booking a two-bedroom cabin within walking distance of the beach during a volleyball tournament.

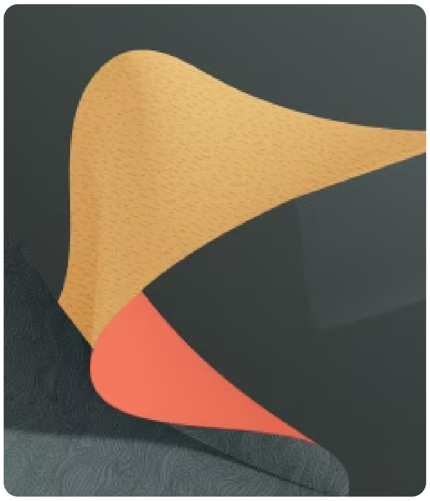
RAG might also be able to assist with more sophisticated lines of questioning. Today, generative AI might be able to tell an employee about the company's tuition reimbursement policy; RAG could add more contextual data to tell the employee which nearby schools have courses that fit into that policy and perhaps recommend programs that are suited to the employee's jobs and previous training—maybe even help apply for those programs and initiate a reimbursement request.

Generative AI With Oracle

Oracle offers a variety of advanced cloud-based AI services, includi
on Oracle Cloud Infrastructure (OCI). Oracle's offerings include robu
unique data and industry knowledge. Customer data is not shared
customers, and custom models trained on customer data can only

In addition, Oracle is integrating generative AI across its wide range
capabilities are available to developers who use OCI and across its c
AI services offer predictable performance and pricing using single-

The power and capabilities of LLMs and generative AI are widely known and understood—they've been the subject of breathless news headlines for the past year. Retrieval-augmented generation builds on the benefits of LLMs by making them more timely, more accurate, and more contextual. For business applications of generative AI, RAG is an important technology to watch, study, and pilot.



What makes Oracle best suited for generative AI?

Oracle offers a modern data platform and low-cost, high-performance AI infrastructure. Additional factors, such as powerful, high-performing models, unrivaled data security, and embedded AI services demonstrate why Oracle's AI offering is truly built for enterprises.

Learn more about Oracle's generative AI strategy

Retrieval-Augmented Generation FAQs

Is RAG the same as generative AI?

No. Retrieval-augmented generation is a technique that can provide more accurate results to queries than a generative large language model on its own because RAG uses knowledge external to data already contained in the LLM.

What type of information is used in RAG?

RAG can incorporate data from many sources, such as relational databases, unstructured document repositories, internet data streams, media newsfeeds, audio transcr

How does generative AI use RAG?

Data from enterprise data sources is embedded into a knowledge r
which are stored in a vector database. When an end user makes a q
relevant contextual information. This contextual information, along
language model, which uses the context to create a more timely, ac

Can a RAG cite references for the data it retrieves?

Yes. The vector databases and knowledge repositories used by RAG contain specific information about the sources of information. This means that sources can be cited, and if there's an error in one of those sources it can be quickly corrected or deleted so that subsequent queries won't return that incorrect information.

Resources for

- Careers
- Developers
- Investors
- Partners
- Startups
- Students and Educators

Why Oracle

- Analyst Reports
- Cloud Economics
- with Microsoft Azure
- vs. AWS
- vs. Google Cloud
- vs. MongoDB

Learn

- What is AI?
- What is Cloud Computing?
- What is Cloud Storage?
- What is HPC?
- What is IaaS?
- What is PaaS?

What's new

- Oracle Supports Ukraine
- Oracle Cloud Free Tier
- Cloud Architecture Center
- Cloud Lift
- Oracle Support Rewards
- Oracle Red Bull Racing

Contact us

- 🇻🇳 VN Sales: 842439447521
- US Sales: +1.800.633.0738
- How can we help?
- Subscribe to emails
- Events
- News
- OCI Blog