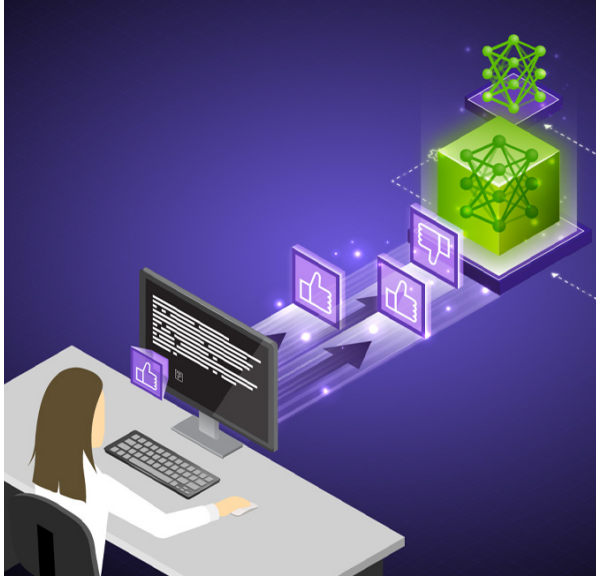
Industry Brief



Activate Your Data With Custom Generative AI

Retrieval-augmented generation (RAG) is making generative AI more precise, more affordable, and more effective for organization-specific use cases.

Generative AI applications powered by large language models (LLMs) can automate tasks, understand and generate natural language in text or speech, and swiftly analyze large bodies of data to provide instant responses to queries. With **generative AI capabilities, enterprises are transforming business** functions such as customer service, content creation, software development, and IT operations. Large multimodal models (LMMs) can process various data types, including text, images, videos, audio, graphs, and more–unlocking new use cases across many industries such as healthcare, finance, and energy.

While large-scale LLMs can bring about impressive efficiency improvements for time-consuming tasks such as generating scripted responses, translating text, and summarizing documents, they have limited enterprise utility because they lack knowledge of organization-specific policies, best practices, and operations. Fine-tuning LLMs on enterprise data requires robust computing power and AI expertise, putting customized AI out of reach for many organizations.

Industry leaders are turning to a new AI technique known as retrieval-augmented generation (RAG) to make enterprise AI more cost-effective and customized to their use cases. By connecting AI models to hand-picked data sources, RAG breaks down barriers to fast, effective, and case-specific generative AI deployment.

Key Challenges for Generative AI Workloads

- **Time-intensive:** Testing generative AI deployments is a laborious, time-consuming process.
- **Accuracy:** Generative AI tools that rely on LLMs are only as accurate as the data they were trained on, which may quickly become outdated.
- **Cost:** Fine-tuning and retraining LLMs is compute-intensive, increasing data center costs.
- **Explainability:** It may be difficult or impossible to determine how LLMs arrived at their generated output, reducing model trust.
- **Scaling:** Going from proof of concept to enterprise deployment requires robust computing infrastructure optimized for generative AI workloads.

Superior Generative AI With RAG

Retrieval-augmented generation is an AI framework that connects foundation or general-purpose LLMs to proprietary knowledge sources such as inventory management systems, sales records, and customer services protocols. This allows AI assistants, conversational chatbots, and other solutions to respond to organization-specific inquiries with the most up-to-date, accurate, and specific information.

Leveraging RAG techniques brings numerous benefits to existing generative AI tools.

Greater Model Output Control

RAG lets enterprises easily control what their model does and doesn't know by simply altering the data in the connected knowledge source, without the time-consuming process of retraining the entire model on new data.

Improved Accuracy

RAG-equipped models can accurately respond to queries even without prior training on the relevant data, minimizing the chance that the model will hallucinate or return false information.

Advanced Contextual Understanding

By connecting generative AI to proprietary data, RAG improves the model's ability to understand domain-specific context, boosting the quality of output delivered to users. This results in more useful generated answers to queries.

Improved Model Explainability

With retrieval from designated knowledge bases, the AI model can offer explanations and justifications for its output by referencing specific information sources. This boosts user confidence by providing clear reasoning behind generated responses.

Reduced Data Center Costs

RAG bypasses the process of model retraining and reduces model size at inference, the stage of AI where models run computations to make predictions or generate outputs in near real time. This saves computational costs and opens up resources for innovation in other areas.

RAG enables enterprises to make use of their own data.	
Information retrieval and analysis	> Document summarization > Information retrieval > Semantic searches > Classification > Sentiment analysis > Recommendations
Knowledge base enhancement	> Conversion of technical and policy manuals, videos, or logs into knowledge bases > Customer and field support > Employee training > Developer productivity
Trust and verification	> Citation of sources for model claims to build trust with users
Specific applications	> Multilingual customer service chatbots > AI writing assistants for marketing teams > Virtual assistants for employees > Professional aids connected to legal or medical databases

Benefits of RAG

- Improves explainability with controlled retrieval from vetted sources.
- Domain-specific generation informed by proprietary data.
- Smaller model size reduces training and inference costs.
- Updates AI tools without retraining models.

NVIDIA AI Enterprise includes software for Retrieval-Augmented Generation

- **NVIDIA NIM** is a set of easy-to-use microservices designed to accelerate the deployment of generative AI models across the cloud, data center, and workstations.
- **NVIDIA NeMo Retriever**, part of the NeMo platform, is a collection of generative AI microservices that enable enterprises to enhance generative AI applications with RAG. AI models can be connected to diverse business data wherever it resides.
- **NVIDIA Riva**, provides speech and translation capabilities for RAG applications.

Industry Outlook: Building Business-Specific Applications With RAG

With guardrails designed to ensure that AI models stay on topic and return accurate results, generative AI solutions using RAG bring greater precision and efficiency to operations across industries.

Retailers can use RAG-equipped solutions to help monitor equipment in the supply chain. For example, one large retailer is using RAG to query telemetry data from cold storage equipment to monitor performance, predict failure, and stage parts before failure occurs. This helps minimize downtime to ensure perishable items are always kept at the optimal temperature.

To enhance the consumer experience, RAG-powered shopping advisors can reference customer histories to deliver hyper personalized fashion recommendations, product recommendations, timely promotions, and more.

In the **public sector**, government agencies can use RAG applications to enhance emergency and disaster response. By retrieving information about previous incidents, response protocols, and best practices, RAG applications can generate risk assessments and communication templates to aid in decision-making, resource allocation, and public communication during emergencies. Agencies can also improve citizen services by building multilingual chatbots to provide information during routine interactions like filing taxes and applying for a passport.

Financial service institutions can use RAG to connect analytics tools to vast bodies of public and proprietary data, allowing traders to query and analyze market data quickly, insurance companies to process claims faster, and home buyers to accelerate real estate transactions.

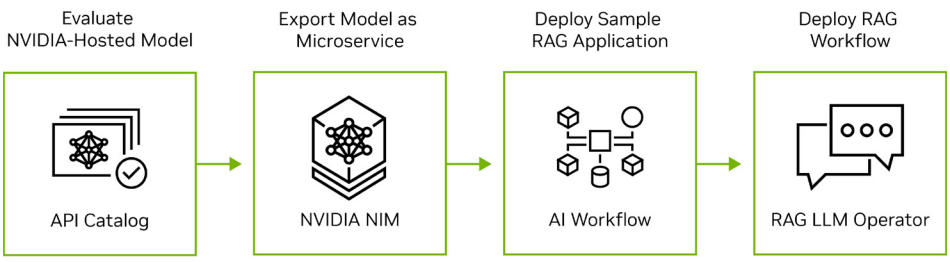
In **healthcare**, medical professionals can use an intelligent clinician's assistant tool to quickly obtain accurate information about a patient's medical history, test results, and treatment plans, improving the efficiency of healthcare delivery and the quality of patient care.

Telecommunications companies can use RAG to support contact center agents resolving customer calls, collect and decipher complex network data in real time to easily identify outages and failures, and help field technicians quickly restore service by generating troubleshooting recommendations.

In areas such as **manufacturing**, RAG chatbots can support worker safety and efficiency by instantly answering questions on safety protocols and providing access to repair manuals and technical documentation.

Getting Started With Retrieval-Augmented Generation

Most enterprises have initiated multiple pilot projects aligned to generative AI use cases. However, it's estimated that **90 percent of those won't move beyond the evaluation** phase in the near future.



> "Retrieval-augmented generation is changing the game for enterprise chatbots. RAG makes it easy to chat with the data buried in your enterprise knowledge systems, unlocking new business insights, and improving employee productivity."
>
> **Rama Akkiraju**,
> Vice President of IT, NVIDIA

Successfully bringing RAG pipelines from pilot production is a complex process that includes multiple steps. NVIDIA helps businesses manage that process by providing a reference architecture for cloud-native, end-to-end RAG applications. With the NVIDIA API catalog, developers and business leaders have a one-stop shop to try models and microservices for their specific use cases that can be downloaded and deployed anywhere.

NVIDIA NIM, part of **NVIDIA AI Enterprise**, provides a streamlined path for developing AI-powered enterprise applications and deploying AI models in production.

NIM is a set of optimized cloud-native microservices designed to shorten time-to-market and simplify deployment of generative AI models, across cloud, data center, and GPU-accelerated workstations. It expands the developer pool by abstracting away the complexities of AI model development and packaging for production using industry-standard APIs.

To help businesses reduce the time and resources needed to build RAG applications, NVIDIA has developed **AI workflow examples**. These examples provide starting points for building and deploying enterprise solutions that accurately generate responses for a variety of RAG use cases.

With RAG, enterprises can build customized generative AI tools that tap into proprietary data and business requirements. Such organization-specific generative solutions promise to usher in a new era of operational efficiency and transform experiences for both employees and customers, while protecting operational security.

Ready to Get Started?

Contact us to speak with an AI expert about RAG:

nvidia.com/generative-ai-chatbots

Contact Sales at: **nvidia.com/contact**

NVIDIA