INVISIBLE

A New AI Capability is Reshaping Knowledge Management

Here's What Leaders Should Know

INVISIBLE

A New AI Capability is Reshaping Knowledge Management—Here's What Leaders Should Know

> Introducing Retrieval-Augmented Generation

Imagine being able to converse with all of the knowledge in your company and the world's best library at the same time—with sources cited, weighted, and up-to-date. That's the kind of thing that's possible with RAG, a technique for grounding an LLM's responses in a trusted set of sources. Resulting responses are accurate, transparent, and—unlike legacy knowledge solutions—easier to keep up to date. Source data can be changed in real-time with no need to retrain the model.

AI21labs | cohere | Microsoft | OpenAI

With RAG, Conversational LLMs Deliver Reliable Outcomes

Traditional Tech is Reliable

Traditional technology, such as software or workflow automation platforms, are predictable: the same input will consistently generate the same output. Other traditional technologies, such as narrow AI algorithms, can also be predictive, seeing microscopic patterns in data that human analysts miss. But, however powerful and reliable, traditional technologies are limited to executing workflows that have already been defined.

Conversational LLMs are Versatile

By contrast, users can leverage conversational LLMs to achieve nearly any goal or task across myriad contexts. Despite this power, companies struggle to capture value from LLMs because they have an Achilles heel. LLMs are prone to hallucinate. While they lack factual grounding, their applications are limited.

With RAG, Conversational LLMs engage reliably and unlock the power of traditional tech

When RAG ensures LLMs rely on context-relevant evidence sets, it increases their accuracy and reliability dramatically. Not only can RAG make LLMs more trustworthy, RAG agents can also help LLMs use technology on a user's behalf.

The combined effect increases the ways LLMs can be deployed exponentially, and creates opportunities to deploy existing technologies more broadly.

> Use Cases. New Possibilities for GenAI Business Applications

RAG is already being combined with LLMs to accelerate investment decisions, to give chatbots "memories" that help them engage like human assistants, and to animate question-answer bots for companies such as the Mayo Clinic. Whether improving the economics of decision-making workflows or enabling novel, personalized consumer experiences, RAG is poised to reshape value chains throughout the knowledge economy itself. For example:

Accessing Investment Briefs

Financial services do deep investigation on all their prospective investments. What if they could query past write-ups to speed up current diligence?

Invisible helped one firm make all past research accessible and comparable, reducing time to make decisions and giving them a competitive advantage versus other prospective investors.

Giving Bots "Memory"

While chatbots are helpful assistants at given points in time, they can't remember the people, places, and things that matter well enough to be effective over time.

Invisible helped train a RAG-augmented chatbot to recognize and store the context that matters but to forget sensitive details, like login information and PII. As a result, the chatbot came one step closer to handling context like a human assistant.

Other Uses We're Seeing

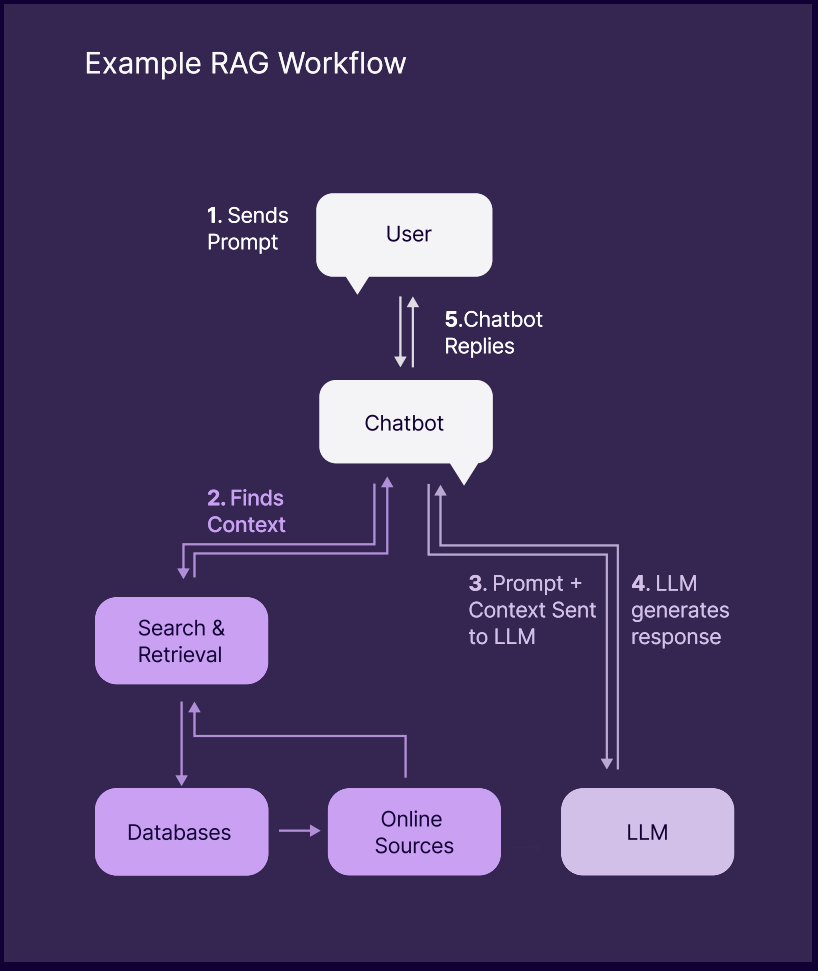
- Quantitative trading in finance
- Clinical question-answering systems in healthcare
- Streamlining decisions in the power and energy sectors
- Next-generation customer support
- Connected vehicle service repair
- Personalized document queries

> RAG 101. How RAG Actually Works

At its simplest, RAG is like giving an LLM a library and teaching it how to use it.

1. First, clients define and curate the information they want an LLM to access and ensure the LLM can efficiently "read" it, usually through an embedding and a vector database. That information could range from a knowledge base to a repeated internet search, or any combination one can imagine.
2. The system merges a user's prompt with those reference materials, and LLMs respond to the context-enriched prompt with a reply.
3. Usually within a matter of seconds, the LLM's response is passed to the user. Results are highly accurate and can be made transparent. For instance, citations are a commonly added feature.

RAG can be applied to more sophisticated workflows as well, operating less like libraries and more like research assistants. They can perform tasks like web browsing, calculations, and data analysis and integrate all of that into their "thinking." Such applications empower people to more efficiently access, make sense of, and act on knowledge sourced from the reliable sources of truth they most prefer.



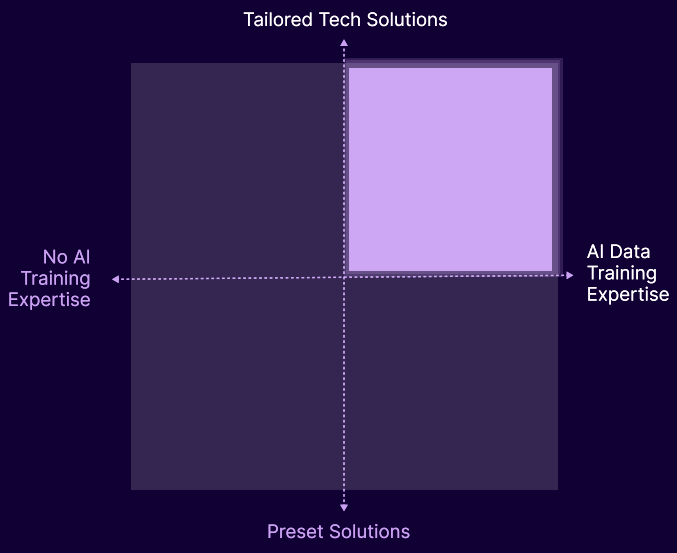
> RAG 102. How RAG Can Work for You

Implementation plans are shaped by a number of variables, including security protocols, data structures, user needs, and projected volumes. For instance: is the data multimodal? Will users input long prompts? Whether an implementation takes 30 days or closer to 300, leaders can expect to progress through the following milestones.

Stage:	Scoping	Setup	Development	Deployment
Example Decisions:	Business and technical stakeholders align on the desired customer impact, product roadmap, high level data pipeline, and technical requirements.	Data pipelines are defined and key choices on architecture, governance, and microservices—such as embedding models, vector databases, and base LLMs—are made.	Technical leads scope, develop, and test each component of the solution. This includes tuning LLMs to ensure they work effectively with new applications.	The solution is beta tested for UX insight and safety prior to launch. Continuity plans are made to ensure the business is positioned to meet user needs.

> Tailored RAG Solutions. The Tailored Solutions and AI Expertise You Need

As an AI training partner for top foundation model developers, Invisible is uniquely suited to build tailored RAG solutions that work seamlessly with the best LLM(s) for your use case. In addition to training LLMs, Invisible creates tailored solutions through its process orchestration platform, integrating LLMs, 3,000+ on-demand subject matter experts, and hundreds of apps to transform how companies are built and run.



Tailored Tech Solutions

Invisible develops RAG solutions based on companies' individual qualities, using pre-built and custom integrations and customizing model training by using domain-specific data and IP.

AI Data Training Expertise

A strategic partner to top model developers, Invisible is uniquely able to transform and generate data sets at industrial scale, and to train AI to use those data sets and technologies.

> Get Started. Strategic Questions with which to Begin

RAG's value is primarily limited by imagination itself. In order to maximize RAG's potential value, start by considering your current knowledge assets and the processes they can be incorporated into for superior efficiency.

1. What internal decision-making workflows could be accelerated to create a business advantage?
2. Do we have access to data sets which could be enriched to create outsized advantages for users?