coveo™

3 Key Steps to Getting Started with **Generative Answering**

White Paper

Contents

Implications for **Next-Generation Digital Experiences**

Thanks to natural language applications such as ChatGPT and BARD, generative AI (GenAI) and large language models (LLMs) are now mainstream. What does this mean for businesses? Enterprises now have access to a technology that can take **unstructured data** and create meaningful answers — taking digital experiences to the next level.

The enthusiasm for this technology makes sense. In the past few years, there has been a subtle shift in people's expectations of digital experiences. They've gone from digging for information to seeking answers. They don't want a list of documents that might contain the answer they're looking for. They want *the* answer — one that satisfies their specific questions or problems so they don't have to pick up the phone.

Can GenAI help in these scenarios? The answer is yes, but not without careful consideration. This white paper will walk you through questions, implications, and concerns around GenAI. In the end, you'll be able to build a solid business case to begin using this innovative technology in your organization.

The current landscape

Many organizations have multiple, often siloed search infrastructures. One for educational curriculum, another for their customer community, one for customer support, another for documentation, knowledge bases on specific domains, one for their corporate website, and so on.

The answer to any question might have bits and pieces in any or all of these repositories – and the user usually has to wade through all the links, read the text, and then figure out how to reconcile all the versions. Needless to say, all of these touchpoints, and corresponding efforts, create a disjointed and unsatisfactory journey.

Generative answering will best work if you can unify all these knowledge repositories to generate the most accurate answer.

What are LLMs and GenAI?

Just to make sure we're all on the same page, here's a primer.

LLMs are language models that are trained on large data sets to recognize parts of speech and structure of text to realistically predict human-like responses.

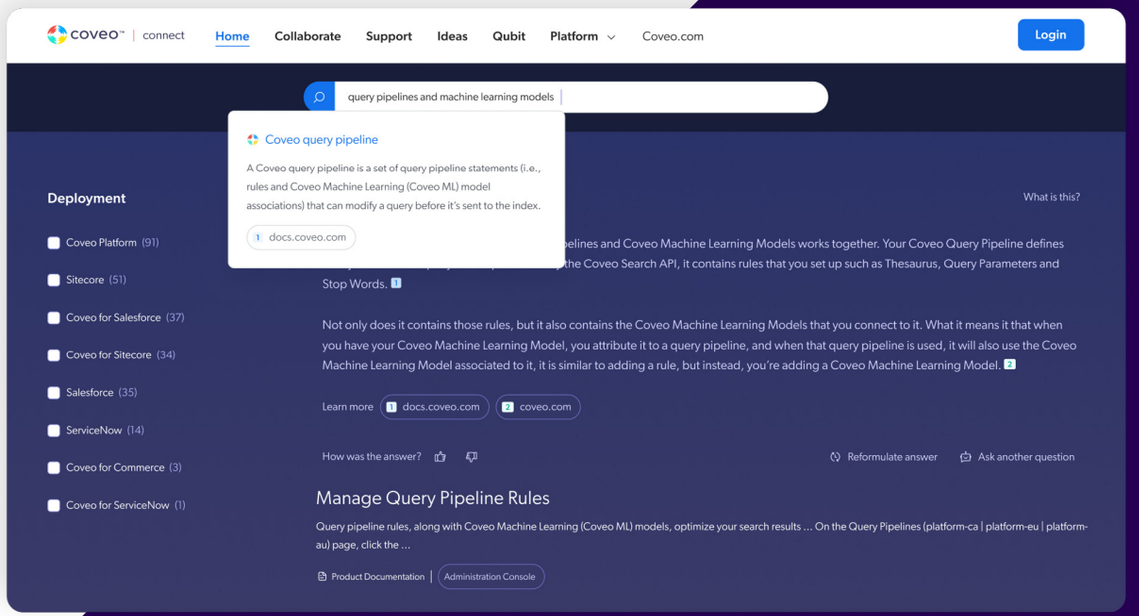
GenAI involves the use of algorithms to generate new, original data or content. In the case of conversational technologies such as BARD and ChatGPT, the algorithm is based on LLMs to generate text and create fluent conversations.

What can GenAI do?

The main benefit of GenAI is generating output that resembles human-created content. This ranges from text to video and much more, but this white paper will focus on **generative answering**.

Generative answering enables classification, the extraction of passages, and the formulation of an answer in response to a question. This will transform how audiences — from customers to employees and far beyond — consume information, powering a dramatic shift in the entire digital experience paradigm.

Visitors can now engage in an advisory exchange digitally. They can ask a question and get a well-constructed answer that spans multiple contexts — instead of typing a query and getting a list of links that they must sift through to find their answers.



Concerns and limitations of generative answering

Generative answering puts enterprises in the difficult position of figuring out the best ways to take advantage of the rapidly developing tech while mitigating the risks. Through interviews with more than 50 CIOs, we've identified the **biggest headaches to overcome**:

1. Security

- **Permissions**, privacy cybersecurity
- **Proprietary** content vs public LLMs retention governance and IP

2. Accuracy

- **Currency**, of generative content
- **Factuality**, veracity of answers
- **Content lineage, traceability** to **sources of truth**

3. Content

- **Multiple sources, volume** and **variety** of **content** increases the **value of GenAI** exponentially
- Flexibility to evolve data landscape
- Ethical use of first party data only

4. Costs

- GenAI experiences can be 100x more expensive if not engineered right
- **Business case & ROI**
- Looking-in with unique GenAI provider

5. Experience

- **Relevancy** for users
- **Unified "intent" and engagement** experience combining search, answering and disambiguation

These challenges can all be addressed with a robust unified AI search platform that offers Relevance Augmented Generation. Here are three steps you can take to put your initiative on the path to success.

3 Key Steps Businesses Can Take to Get Started With **Generative Answering**

1. Invest in a robust unified AI search platform

Generative answering excels at generating language. But generating language alone won't be satisfactory if the model doesn't have enough content — nor the right content — to draw on. This means investing in a search platform that can index, categorize, and surface content in a way that respects security and access rights.

What's more, your index should be able to interface with an LLM intelligently. Adopting a search platform that employs Retrieval Augmented Generation (RAG) means your LLM generates an answer only with contextual information relevant to the user and prompt at that moment in time.

Investing in search means investing in content creation and in feeding the LLM with information relevant to your domain based on a user's needs. It means you're committing to creating a new way of interacting with customers through new enabling technologies.

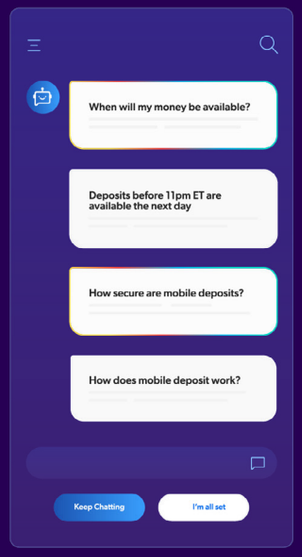
2. Select a starter use case for generative answering

As with any technology, there have to be clear reasons for adoption. If an organization starts deploying generative answering without a clear purpose in mind, it is unlikely to see results.

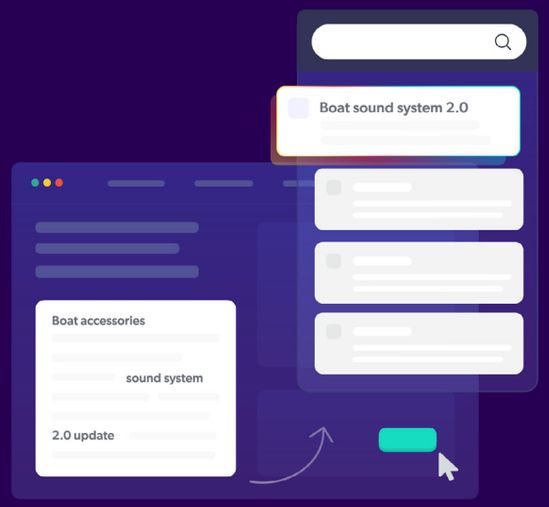
First, decide which audience you want to target with generative answering. Customers, employees, site visitors? It's helpful to choose a use case using your audience's perspective: improving customer self-service might mean unifying access to content from multiple repositories, and making the generative answering interface accessible from the customer service portal.

Or, if you're hoping to improve employee productivity, you need to assess those data resources used most often, those that might be overlooked; analyze information that might be lacking or stale — and provide an interface in a commonly accessed area, such as a company intranet.

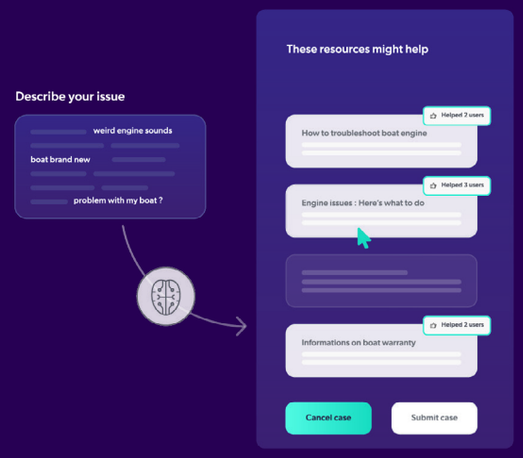
From there, think about what outcomes you're looking to achieve. List the risks and barriers to entry. Identifying the main issue to be solved and establishing metrics to quantify its success in advance will help ensure the most effective use of this new technology — while providing a north star to track your return on investment.



Chatbot



IPX



Case Form

3. Amplify your knowledge strategy

While generative answering will bring significant changes to the structure of today's organizations, language models, no matter how good, will never be up-to-date enough to know all the answers. It's important to remember that for all its power, generative answering is another tool that allows us to organize knowledge — and knowledge comes from people.

This introduces the concept of 'human in the loop', or HITL. While generative answering will help people interact with one another (such as revealing new insights or providing a faster summarization), subject matter experts are still needed to manage your knowledge strategy and identify that answers generated by your AI are accurate for your industry.

As stated previously, the power of an LLM is only realized with the amount of information it has access to. This means developing a robust knowledge-sharing culture that brings institutional information out of people's heads and onto the digital page. Likely, you will want a team that:

- Understands the current state of your knowledge landscape
- Prioritizes quality content
- Establishes knowledge management practices for generative answering

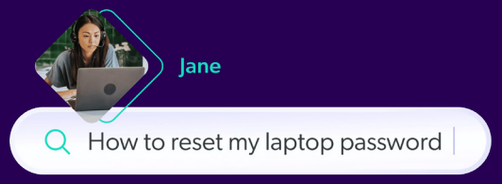
Your search platform should also provide usage analytics to help identify the above. It can reveal content gaps of what people search for and aren't finding. After all, your LLM can't generate an answer to a question it doesn't have the content for.

Coveo Relevance Generative Answering

Out-of-the-box AI search platforms like Coveo let you quickly use an LLM. It does this by providing a much-needed administrative wrapper around an LLM, making it enterprise-ready and accessible to organizations of all types and sizes.

1. Coveo creates an index of all your data (PDFs, HTML files, docs, you name it)

2. Your user searches this index for an answer



3. Coveo GenAI uses mature LLMs to identify relevant snippets in your documents

4. An accurate, tailored answer is generated. Factual, up-to-date — complete with citations.

Coveo's history with LLMs

Coveo has been using LLMs in production for large enterprises since 2020, particularly for Case Classification and Smart Snippets.

Smart Snippets

Smart Snippets models provide answers to customers' queries directly on the results page by displaying a snippet of the most relevant result. Users can quickly find answers without having to open links.

To surface the most relevant content snippet, the system leverages an ML algorithm backed by LLMs to compute a similarity score between a user query and item headings in a results list. Based on that, the system finds an item that is the most relevant to the query, then pushes it to the top of the results list in the main snippet section.

Coveo ML Smart Snippets also powers the "Related Content" (People also ask) section that users can click to find further information related to their query without leaving the results page.

Case Classification

Coveo uses LLMs to classify a case based on previous cases, thereby removing some customer effort as they submit a support case.

Coveo ML Case Classification models learn from support cases that have been correctly classified to provide classification suggestions for cases that haven't yet been classified.

To do that, Coveo uses the Case Similarity approach for support cases that contain little text, while using the Context Recognition approach when dealing with larger support cases.

For the Context Recognition approach, the system leverages NLP techniques and a deep learning algorithm to process text appearing in support cases. It takes into account common vocabulary, contextual nuances, and key concepts used in case descriptions to suggest classifications related to the case in question.

Mitigating security concerns

Coveo is committed to democratizing AI and allowing enterprises to modernize their digital offering while ensuring accuracy, confidentiality, and security for your organization. Because enterprise knowledge systems often contain secure and proprietary information, we take special care to mitigate security concerns.

Learn more about our generative answering security measures.

While using LLMs to assemble an answer, Coveo reduces security risks **by leveraging the permissions and security of your existing content.** We rely on the corporate enterprise content in an organization's knowledge base, product documentation, and help content.

Built for Enterprise **Safety**—and Performance

Secure Content Retrieval

Secure access to +100 cloud and on-premise content sources.

Structured and unstructured data with document-level security in the index.

Grounding Context

The LLM prompt context is grounded with passages from secure, personalized search results.

Helps to ensure that the output is relevant, accurate, consistent and secure.

Auditable Prompts and Responses

All prompts and responses generated by the model are tracked and recorded for auditing purposes.

Ensures the model is being used to generate answers in a responsible and ethical manner.

Zero Retention

Data (queries) is not stored by the LLM provider—nor used for training the LLM.

This helps to maintain the privacy and security of an organization's information.

Ensuring: Mitigation of Financial Security Risks + Best-in-Class Practices in Governance

Ready To Take Your Digital Experience To The Next Level?

Generative answering is cutting-edge technology gone mainstream. While the more sophisticated demand-specific applications will take time to develop, businesses need to start investing in this technology now.

The investment you make now in this technology is a long-term investment that will grant you unprecedented opportunities in the future.

coveo™

The Future is **Business-to-Person**,
powered by **AI Search** and **Generative Experiences**

Learn more about Coveo

Coveo, a leading provider of enterprise AI platforms that enable individualized, connected, and trusted digital experiences at scale with semantic search, AI recommendations, and GenAI answering.

Contact us