



# Disrupting Access to Information in Companies with LLM and Advanced RAG.



How can we quickly and effectively access knowledge dispersed among internal resources and focus on high-value tasks? How can generative AI and Large Language Models (LLMs) assist with analysis, decision-making, and creation in a business context? This white paper explains how to connect Large Language Models (LLMs) to internal data, how Retrieval Augmented Generation (RAG) works, and the potential benefits and uses in a business environment.

# Table of contents

<b>I. THE LLMs MARKET</b>	<b>1</b>
<i>Generative AI and Large Language Models: What's the Difference?</i>	2
<i>The Journey to Generative AI</i>	2
<i>2023: The Proliferation of Gen AI</i>	3
<i>2024 and Tomorrow: the Experimentation</i>	4
<i>Generative AI Market Projection</i>	4
<i>Main Objectives of Generative AI Initiatives</i>	5
<i>Case Study: Examples of Business Use Cases</i>	6
<i>Generative AI: What's the ROI?</i>	7
<b>2. LLM IN BUSINESS: HOW DOES IT WORK AND WHAT ARE ITS LIMITATIONS?</b>	<b>9</b>
<i>How Do Large Language Models Function?</i>	10
<i>Lifecycle of LLMs</i>	11
<i>What Are LLMs Used For?</i>	14
<i>Available Providers and Technologies</i>	15
<i>The Other Types of Actors</i>	16
<i>The Limitations of LLMs</i>	17

<b>3. OPERATING THE LLM WITH INTERNAL DATA THROUGH THE RAG</b>	<b>20</b>
<i>What is the RAG?</i>	21
<i>How Does it Work?</i>	21
<i>RAG Preparation: Indexing the Knowledge Base</i>	22
<i>Using the RAG</i>	23
<i>Limitations of Simple RAG in Business Context</i>	24
<i>Advanced RAG</i>	25
<i>Zoom on RAG Vision</i>	28
<i>A Powerful RAG to Unlock the Doors to Knowledge Intelligence</i>	29
<b>4. BENEFITS AND USE CASES IN BUSINESS</b>	<b>32</b>
<i>Benefits of RAG in Business</i>	33
<i>Industry Use Cases</i>	34
<b>5. SOLUTION USING ADVANCED RAG: ENTERPRISE CHAT "PLAYGROUND" CONNECTED TO INTERNAL DATA</b>	<b>36</b>
<i>What is it for?</i>	37



# Part 1



## The LLMs Market

# The LLMs Market

## Generative AI and Large Language Models: What's the Difference?

LLMs use data to learn and develop an understanding of language. These are models that process natural language and generate natural language. Not all generative AIs are based on LLMs, but all LLMs are a form of generative AI.

Generative AI can produce a variety of content: text, images, videos, music, voices, code, etc. It encompasses a multitude of tools designed to exploit data from LLMs and other AI models using machine learning to create new content.

In contrast, an LLM is a specific type of AI model that uses machine learning based on billions of parameters to understand and generate text.

## The Journey to Generative AI

Source : Gartner.

**2010**

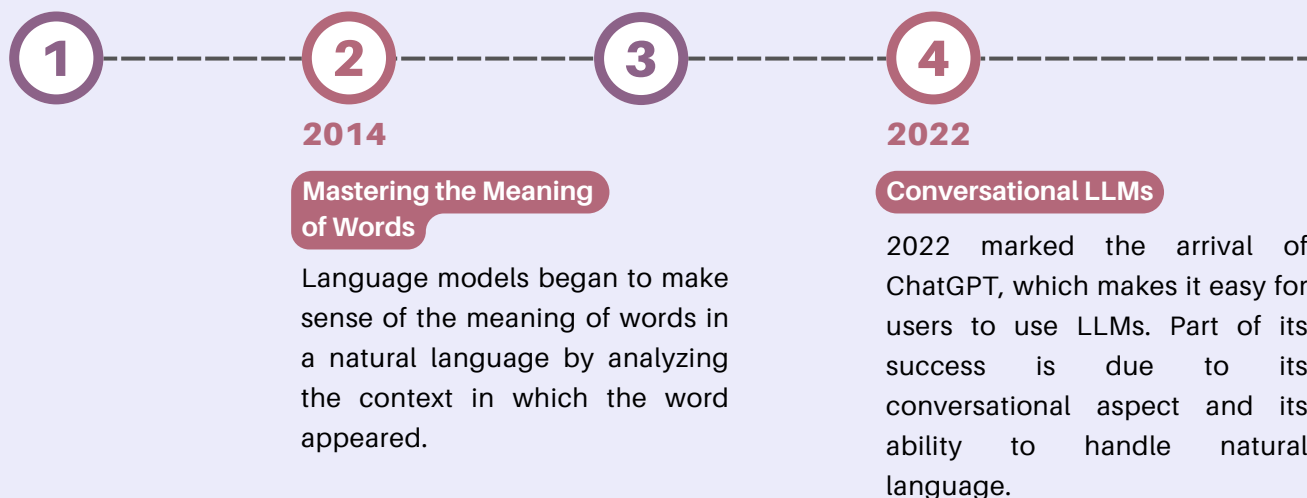
### Near-Perfect Translation of Natural Language

AI researchers working on natural language translation discovered that models exposed to vast amounts of text produced much better results than models that just using top-down grammatical rules.

**2017-2022**

### Large language foundation models

Creating foundation models is cost-prohibitive, but once created, they can be customized using a small amount of additional data to achieve state-of-the-art performance on new tasks.



## 2023: The Proliferation of Gen AI

Generative AI was the major technological revolution of 2023, with many people trying and enjoying ChatGPT or Midjourney.

While the generative AI market was valued at \$23.2 billion in 2022, it reached \$44.9 billion in 2023. It was in this same year that we witnessed the fast proliferation of Generative AI.

This technology emerged as a revolution, and many individuals tested generative AI for personal use before adopting it for more professional purposes.

ChatGPT emerged as the flagship of text generators, but competition was quick to follow. In 2023, ChatGPT held 19.7% of the market share, followed by Jasper Chat (13.4%), YouChat (12.3%), DeepL (12.1%), and Simplifies (9.7%), leaving a third of the market to other players. In image generation, Midjourney (26.9%), DALL-E (24.4%), and NightCafe (23.5%) were the most well-known.

The success of generative AI quickly led to its monetization, with over 200 million monthly users for ChatGPT, 30 million daily users, and 25 to 50 million paying users.

Rarely has a technology been adopted so quickly, as illustrated by a study conducted by O'Reilly (an American publishing house specializing in computer science) among its users:

67% report that their company uses generative AI.

16% of those working with AI use open-source models.

54% of AI users expect the greatest benefit of AI to be increased productivity.

### It's just the beginning

Many AI enthusiasts are still in the early stages: 26% have been working with AI for less than a year.

### One obstacle remains

The difficulty in finding appropriate use cases is the main barrier to adoption, both for users and non-users.

## 2024 and Tomorrow: the Experimentation

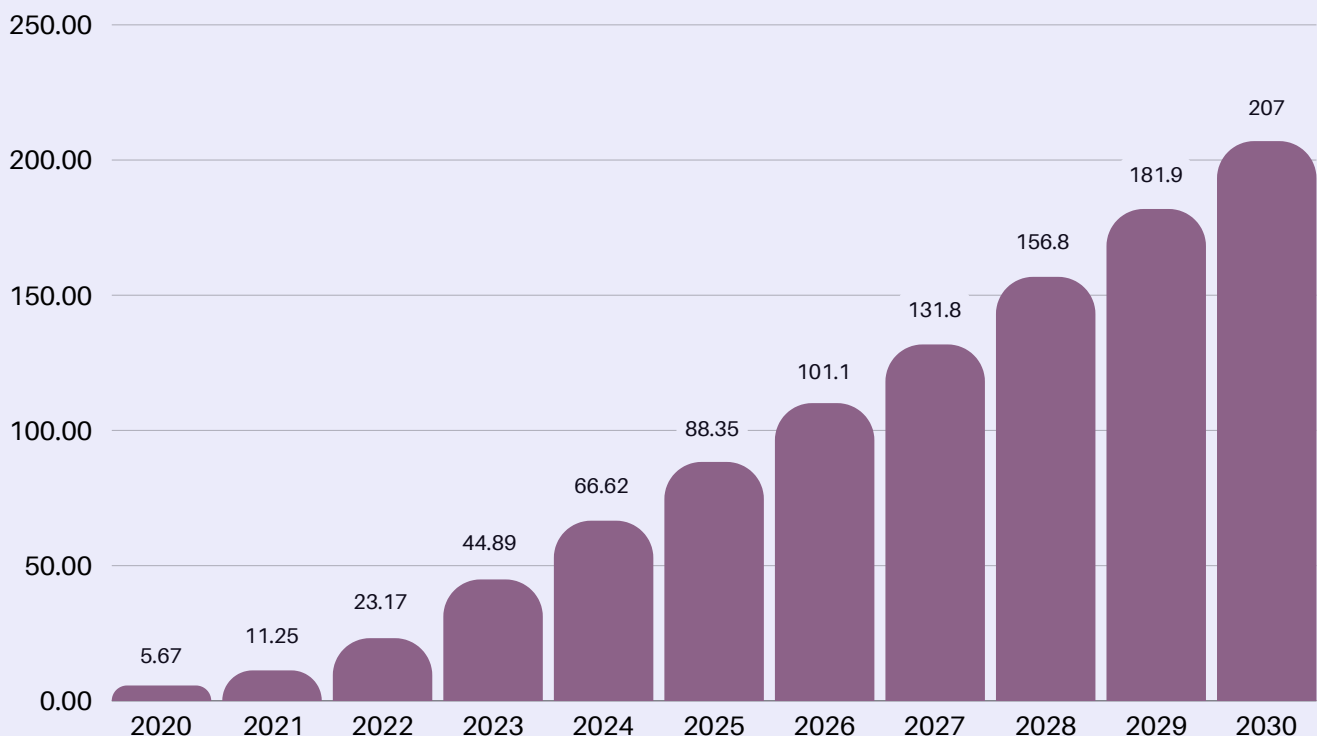
According to research conducted by Sopra Steria Next, the generative AI market is expected to increase tenfold by 2028, reaching approximately \$100 billion, with an annual growth rate of 65%.

**By then, the time spent on generative AI applications will double**, from 30 minutes to 1 hour per day, thereby increasing the average revenue per user from \$30 to \$40 per month.

## Generative AI Market Projection

Global market size in billions of dollars, estimated in August 2023.

Source : Statista.

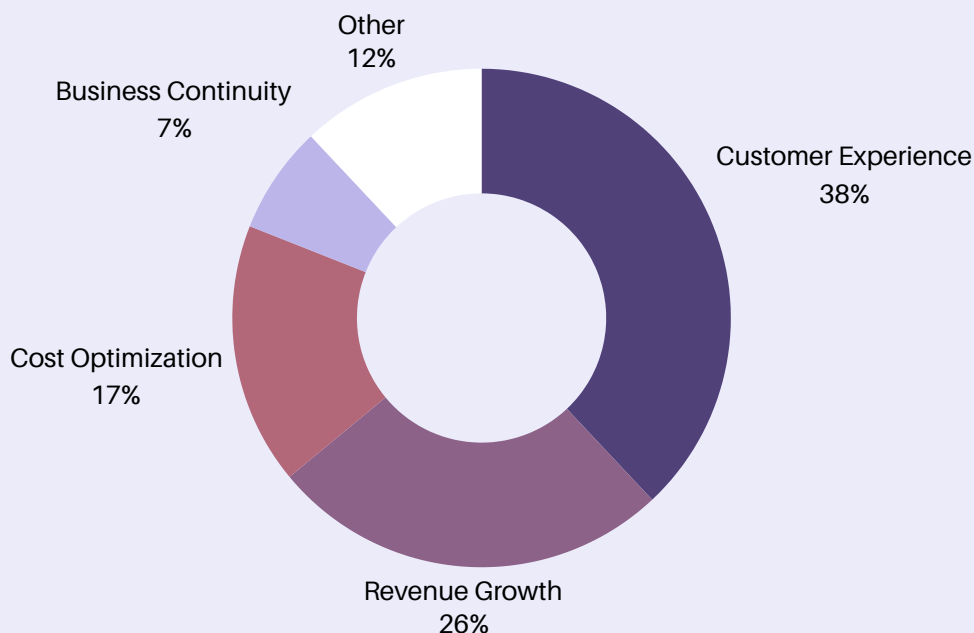


In a three-year horizon, the potential productivity gain in businesses could reach 10% if generative AI is deployed across four priority application areas: customer service, digital marketing, software engineering, and knowledge management.

In a recent Gartner survey of over 2,500 executives, 38% indicated that improving customer experience is the primary goal of their investments in generative AI. This is followed by revenue growth (26%), cost optimization (17%), and business continuity (7%).

## Main Objectives of Generative AI Initiatives

Source: Gartner.



The impact of generative AI will be significant across a multitude of sectors, including medicine, manufacturing, media, architecture, interior design, engineering, automotive, aerospace, defense, healthcare, electronics, and energy.

### Médical

By 2025, it is projected that over 30% of new drugs and materials will be systematically discovered through generative AI techniques.

### Marketing

By 2025, 30% of marketing messages from major organizations will be generated by AI, a sharp increase from less than 2% recorded in 2022.

And more broadly, generative AI is promised a bright future within the enterprise:

- By 2024, 40% of enterprise applications will integrate conversational AI, compared to less than 5% in 2020.
- By 2025, 30% of companies will have implemented an AI-optimized development and testing strategy, compared to 5% in 2021.

- By 2026, generative AI will automate 60% of the design efforts for new websites and mobile applications.
- By 2027, nearly 15% of applications will be generated automatically by AI without human intervention. In comparison, this does not happen at all today.

## Case Study: Examples of Business Use Cases

### L'ORÉAL

Take the example of L'Oréal, which recently deployed its internal generative AI, not only for the group's developers but also for certain professions.

Launched at the end of 2023, L'OréalGPT already had 19,000 distinct users after 3 months. The AI generated 330,000 messages and 46,000 images. In the marketing department, this AI allows for idea validation. But L'Oréal does not intend to stop there and wants to multiply possible use cases.

### Morgan Stanley

Morgan Stanley, an investment bank and wealth management giant, announced in September 2023 that it was working on an assistant based on GPT-4. Named AI @ Morgan Stanley Assistant, this tool enables financial advisors to quickly access a database of approximately 100,000 research reports and documents.

"Financial advisors will always be the center of Morgan Stanley wealth management's universe," said Andy Saperstein, co-president of Morgan Stanley. He also said that Gen AI will "revolutionize client interactions, bring new efficiencies to advisor practices, and ultimately help free up time to do what you do best: serve your clients. »

### Carrefour

The Carrefour group is using generative AI to create product pages for over 2000 references, saving time on production and online posting.

Ultimately, Carrefour plans to use it for all its products. Internally, the company employs generative AI for purchasing processes to assist with daily tasks, such as drafting calls for tenders or analyzing quotes.

### McKinsey & Company

McKinsey studied how generative AI could replicate the capabilities of their employees and subsequently launched Lilli, their own generative AI solution for employees. It's a platform that provides streamlined and unbiased research and synthesis of the firm's vast knowledge reservoirs to deliver the best information quickly and efficiently to clients.

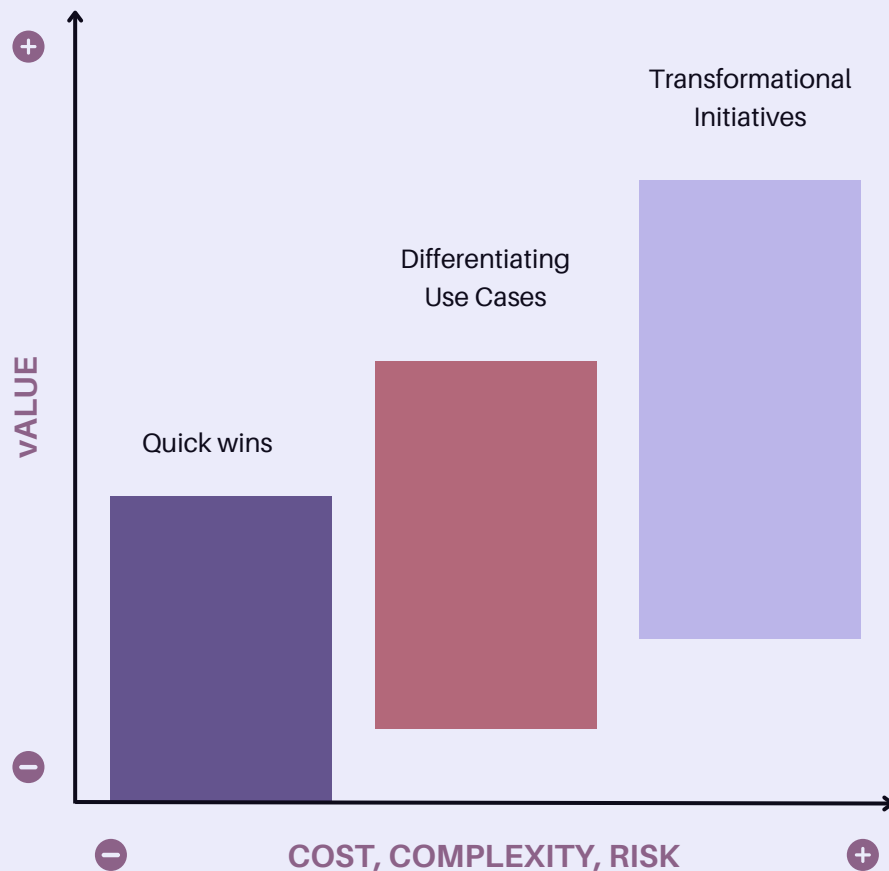
"Lilli aggregates our knowledge and capabilities in one place for the first time and will allow us to spend more time with clients activating those insights and recommendations and maximizing the value we can create," says Erik Roth, McKinsey senior partner leading Lilli's development.

# Generative AI: What's the ROI?

Integrating generative AI into businesses is both a technical and financial challenge. The consulting firm Gartner distinguishes three types of AI projects in enterprises: "quick wins," differentiating use cases, and transformational initiatives.

## Generative AI Use Case Categories

Source : Gartner.



Quick wins, or rapid gains, are projects that can be implemented quickly because they do not pose significant technical challenges. The valuation period is short (less than a year), and productivity improvements provide a competitive advantage that diminishes over time.

Differential use cases represent those that leverage generative AI within a specific industry with customized applications, enabling the utilization of internal company data. Costs and risks are higher than for quick wins, but the competitive advantage is more significant.

Costs may then be offset by potential revenue generation, but the valuation period is slightly longer: between one and two years. Transformational initiatives have a strong impact on markets and revenue models. They come with high costs, risks, and great complexity. These innovations may take much longer to be valued (more than two years), but they offer a significant competitive advantage. Decision criteria for investing in such projects should prioritize strategic advantages that may be difficult to quantify rather than immediately identifiable financial benefits.

### Elqano's perspective

The year 2023 saw the rise of generative AI in the enterprise for quick wins use cases (summary of calls, email sorting, etc.). The year 2024 will be characterized by more complex, higher value-added use cases (analysis work, decision support, drafting complex documents, etc.).

**60 to 70%**

of employee time is spent on tasks that could be automated by today's generative AI.

*Source : MIT.*

**75%**

of the value that generative AI could bring can be broken down into 4 areas: customer service, marketing/sales, software engineering and R&D.

*Source : McKinsey.*

**0.1 to 0.6%**

is the potential growth in workplace productivity per year up to 2040, thanks to generative AI.

*Source : McKinsey.*



**98%**

of executives worldwide say AI models will play an important role in their organizational strategies over the next 3-5 years.

*Source : Accenture.*



# Part 2



## LLM in Business: How Does It Work and What Are Its Limitations?

# LLM in Business: How Does It Work and What Are Its Limitations?

## How Do Large Language Models Function?

Large Language Models operate through a complex architecture based on artificial neural networks, specifically Transformers.

Before the emergence of Transformers, model size was restricted due to computational constraints and technical challenges related to learning from vast amounts of data. However, Transformers changed the game by introducing attention mechanisms that efficiently capture long-distance relationships within sequences.

This ability to model complex dependencies paved the way for creating large-scale models without sacrificing performance.

Transformer architectures are inherently parallelizable, enabling efficient training on large datasets using modern hardware such as GPUs and TPUs.

Consequently, researchers have been able to develop gigantic language models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) with tens or hundreds of billions of parameters.

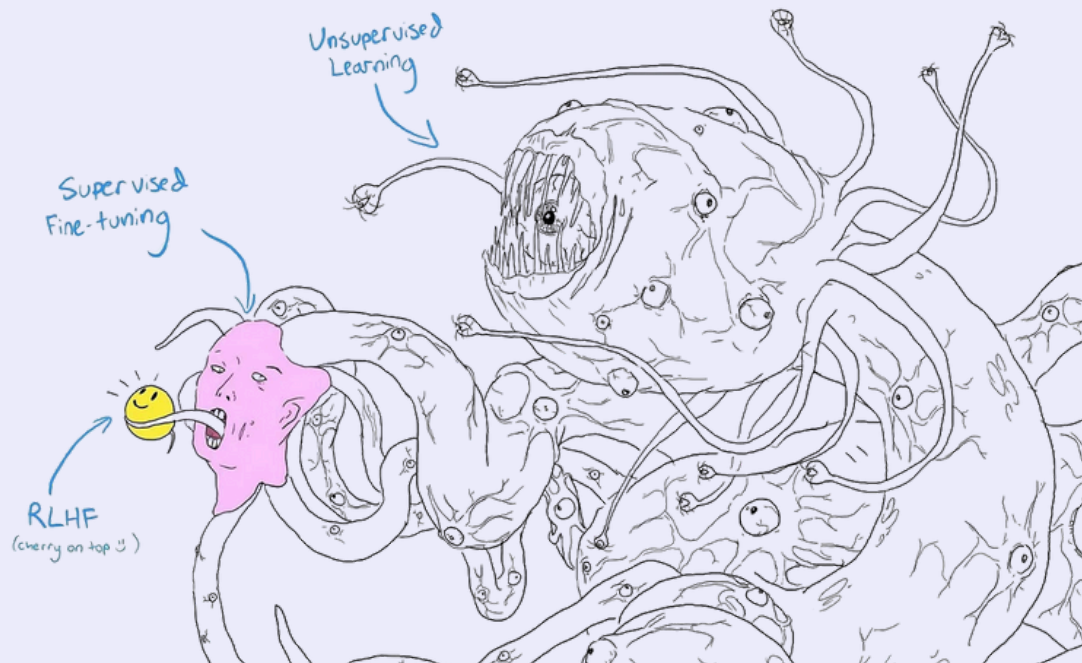
In simplified terms, parameters can be numerical weights in the connections between neurons of a neural network, determining the importance of the connection between two neurons.

With billions of parameters, Large Language Models (LLMs) can grasp complex structures and nuances in textual data: the machine can represent the meaning of a text within the parameters.

# Lifecycle of LLMs:

## 1 Training

LLMs are pre-trained on vast sets of textual data to learn basic linguistic patterns, such as syntax, semantics, and sentence structure. During this phase, the model learns to predict the next word in a given sequence of words.



Conducting this pre-training in the enterprise is a hefty task: it's very expensive (costing several million euros in GPU computing power), requires a vast amount of data, and significant expertise.


Then comes fine-tuning (SFT, RLHF): the pre-trained model is designed to predict the next word. Most market-relevant models are those capable of answering questions, potentially with instructions. To adapt the model, fine-tuning methods (Supervised fine-tuning, and Reinforcement Learning with Human Feedback) are used, which train it on "question answering" and "instruction prompt" tasks, specific tasks using smaller, specialized datasets.

Through these steps, the model is now capable of correctly answering questions.

However, to learn new concepts and knowledge, it needs to be retrained in the lower layers to "imprint" knowledge into the model's weights. This type of method has not yet been proven and generally leads models to hallucinate as they learn to speak on less familiar topics.

Here's what the GPT4all tuning dataset looks like in the Hugging Face dataset viewer. It's quite simple: it consists of a series of prompts (first column) and responses/expected responses (second column) to refine the model's responses.

Prompt	Réponse
<code>&lt;p&gt;Good morning&lt;/p&gt; &lt;p&gt;I have a Wpf datagrid that is displaying an observable collection of a...</code>	One possible solution is to use a fixed width for the GroupItem header and align the header and...
<code>&lt;h2&gt;Hi, How can I generate a pdf with the screen visual data, or generate a pdf of the data being...</code>	To generate a PDF with the screen visual data, you can use a library such as pdf. Here's an...
<code>&lt;pre&gt;&lt;code&gt;package com.kovair.omnibus.adapter.platform; import...</code>	The issue might be related to class loading and garbage collection. When a class loader loads a...

 [See full dataset](#)

## 2 Inference - Model Usage

Once the model is trained, it can be deployed on a computer, often in the cloud as models can be very large: 7 billion parameters for "small" ones and up to hundreds of billions of parameters.

During inference, a Large Language Model (LLM) utilizes the knowledge it acquired during training to perform various language-related tasks.

When a query is submitted to the model, it traverses the input word sequence, assigning varying attention to each word based on its context and meaning.

Using this attention, the model evaluates the probability of the next words in the sequence, thus generating a prediction or response coherent with the given query.

## Lexical Point

### Token

A token represents a basic unit of language, such as a word, sub-word, or even a character. When text is processed by a LLM (Large Language Model), it is first divided into tokens, allowing the model to understand and manipulate language in a more granular way.

### Embedding

An embedding is the numerical representation of a token. Tokens are input into the model and then converted to be processed mathematically. Each token is associated with a numerical vector that captures its semantic and syntactic properties, enabling the model to understand the relationships between words in the language.

### Sentence Embedding

The numerical representation of a sentence or text in a continuous vector space. This allows for capturing the semantic and syntactic features of the sentence, facilitating the comparison and analysis of similarity between different sentences.

Sentence embeddings are typically obtained from pre-trained language models, such as Sentence Transformers, which convert individual words of the sentence into numerical vectors and then combine them to form a representative global embedding of the entire sentence.

### Keep in mind

LLMs are AI models that can predict the most probable responses, thanks to Transformer technologies, enabling the machine to translate the semantic meaning of a word and a sentence into something it handles well: numbers (vectors).

## What Are LLMs Used For?

Generally, LLMs are useful in all natural language tasks. Here are some of the main uses of LLMs:



### Text Generation

LLMs can be used to automatically generate text, whether it's for content creation, auto-email drafting, report generation, or code generation.



### Personalization

They can be used to personalize online experiences, such as product recommendations, content adaptation, or customization of user interactions.



### Automatic Summarization

They can summarize large volumes of text concisely and accurately, helping extract relevant information from a document.



### Automatic Translation

LLMs can be employed to enhance automatic translation systems by providing more accurate and fluent translations in different languages.



### Sentiment Analysis

LLMs can analyze large amounts of text to determine sentiments, opinions, or trends from data such as social media, product reviews, or online comments.



### Decision Making

LLMs can provide analyses and insights based on natural language to aid decision-making in various fields such as finance, medicine, law, or business management.



### Information Retrieval

They can assist in information retrieval by answering questions asked in natural language, extracting relevant information from large databases or texts.

### Keep in mind

LLMs are experts in natural language; they can understand instructions, make connections between words and their meanings, and have learned a significant portion of human knowledge (public) through training on very large datasets.

## Available Providers and Technologies



The most well-known and leading market leader, produces the entire GPT suite (GPT-3.5, GPT-4, and soon GPT-5). These models are used for various applications such as text generation, translation, and question answering.



From the early days of generative AI, Microsoft positioned itself as a major player in the ecosystem, initially with strategic partnerships with OpenAI.

Microsoft then deployed these AI models (GPT-3.5, GPT-4) on its cloud service, and quickly integrated generative AI into its Office suite with Copilot.



Offers several large language models, including its latest model Gemini 1.5 Pro, a multimodal model (for selected companies: up to 1M tokens, 10M intended for research).



Amazon also positions itself as a major player in cloud services for AI usage with tools that facilitate developers' work in implementing AI.

Amazon has heavily invested in Anthropic, which is working on an LLM as powerful as OpenAI's: Claude (1,2,3).



The French startup that has become a flagship of open source, offers models such as 8x22b, 8x7b, 7b, large.

The latest release: 8x22b is "a sparse mixture-of-experts (SMoE) model that uses only 39 billion active parameters out of 141 billion, offering unmatched efficiency for its size."



Meta offers, among others, the Llama 3 model, the latest version of its Llama model family, also open source.

Two pretrained and fine-tuned models with parameters of 8B and 70B have been released with the ability to support a wide range of use cases.

---

## The Other Types of Actors



A powerful open-source framework for developing language model-powered applications. It connects to the AI models you want to use and links them to external sources.

LangSmith provides visibility into what is happening with your LLM application, whether it is built with LangChain or not, so you know how to act and improve quality.

Finally, LangServe provides turnkey API delivery for your LangChain application.



A data framework for LLM applications that integrates, structures, and accesses private or domain-specific data.



An all-in-one development platform for every stage of the LLM-based application lifecycle, whether you build it with LangChain or not.



A framework that helps evaluate Retrieval Augmented Generation (RAG) pipelines.

### Keep in mind

The excitement surrounding the power of LLMs induces an ecosystem that evolves very quickly. Every month the cards are shuffled, and leaders are dethroned. However, very few companies are capable of developing their own LLM from scratch due to the complexity and cost of training.

## The Limitations of LLMs

Despite their power, LLMs also have limitations. Here are some important limitations to be aware of:

### LLMs do not know a company's private data

The information known by the AI is what it learned during its training; its knowledge is static. LLMs are trained on vast sets of textual data from the internet (among other sources), but they are not natively integrated with internal systems and the specific data of each company.

Many companies are seeking solutions to make these AIs work with their knowledge bases. We will later see the difficulties and possible solutions to achieve this.

Such a model can therefore formulate outdated or generic responses when you expect a specific and current response, if it has not integrated the most recent data.

For example, trends, industry standards, consumer preferences, and business practices can evolve rapidly, making the data on which models were trained less relevant for current tasks.

This can be particularly concerning in fields where the accuracy and relevance of information are crucial, such as business decision-making, research and development, or customer service personalization.

When using generative AI in business, it is crucial that the model provides reliable and up-to-date responses. This is where the Retrieval Augmented Generation (RAG) model comes into play.

### Elqano's perspective

In the vast majority of cases, fine-tuning techniques or large contextual windows are not the most recommended for achieving quality results, notably due to their cost, complexity, and performance.

## Building a good prompt isn't always easy

Crafting an effective prompt isn't a skill everyone possesses, and LLMs are particularly sensitive to it. The same question, phrased differently, can elicit entirely different responses. That's why many companies hire Prompt Engineers. Here are some common limitations associated with this task:

**Complexity of professional language:** Companies often operate in specialized domains with specific jargon and linguistic conventions. Formulating prompts that accurately reflect these nuances can be challenging, especially if language models are not pre-trained on specific sectoral data.

**Need for expertise:** Constructing effective prompts requires a deep understanding of the company's domain and the specific task requirements. This task may require input from domain experts to formulate relevant queries.

**Sensitivity to biases:** Poorly formulated prompts can introduce biases into the results generated by LLMs. For example, ambiguous or tendentious formulations can lead to inappropriate or biased responses, compromising the quality and objectivity of the results obtained.

**Performance optimization:** Striking the right balance between query specificity and response diversity can be a challenge. Overly restrictive formulations can limit the model's ability to generate diverse responses, while overly vague formulations can yield imprecise or irrelevant results.

In general, writing a good prompt involves adhering to the basic codes of communication: who am I talking to, what do they know about the subject... Here's a (non-exhaustive) list of useful elements for writing a prompt: persona, context, tasks, expected format, audience, additional information, tone.

## The size of the contextual window

The size of the contextual window is a crucial concept in the field of natural language processing. It refers to the scope of information (range of words) considered by a language model to generate a prediction or response...

Models with large contextual windows have the ability to consider more context around each word when generating text or making decisions.

As the amount of text to consider increases, the computational complexity of the task also increases.

GPT-4 has extended its window to 128K tokens for gpt-4-turbo/gpt4-vision, and although larger than most competitors, this limit remains restrictive for more complex tasks such as reviewing numerous documents.

The size of the contextual window is crucial in leveraging LLMs as it enables providing all the necessary information to answer a question regarding unknown information to the LLM (e.g., company data or recent information).

The size of the context window is a subject of debate in the research world. On one side, Google presents LLMs with window sizes of 1M+ tokens, while on the other side, researchers show that the relevance of responses decreases after a certain threshold.

"Through our study, we discovered that while LLMs show promising performance on inputs up to 20K tokens, their ability to process and understand longer sequences decreases significantly."

Source : [Arxiv](#).

It's important to note that the primary mode of remuneration for LLM providers is billing per token (per word) sent to the LLM and the words generated by the LLM.

### Keep in mind

For enterprise use, LLMs have limitations. And even though some methods allow circumvention, they are very costly and complex to implement. Nonetheless, there are promising solutions: by combining text generation with information retrieval systems, we can improve the accuracy, relevance, and efficiency of LLM responses. This technique, called RAG, paves the way for more powerful applications tailored to specific business needs.



# Part 3



## Operating the LLM with Internal Data through the RAG

# Operating the LLM with Internal Data through the RAG

## What is the RAG?

The RAG expands the capabilities of LLMs by incorporating specific information from domains or an internal knowledge base. It involves providing additional information (via a prompt) to assist the LLM in answering a question. Today, the RAG is the most effective method for enriching LLMs with new information they haven't been trained on.

---

## How does it work?

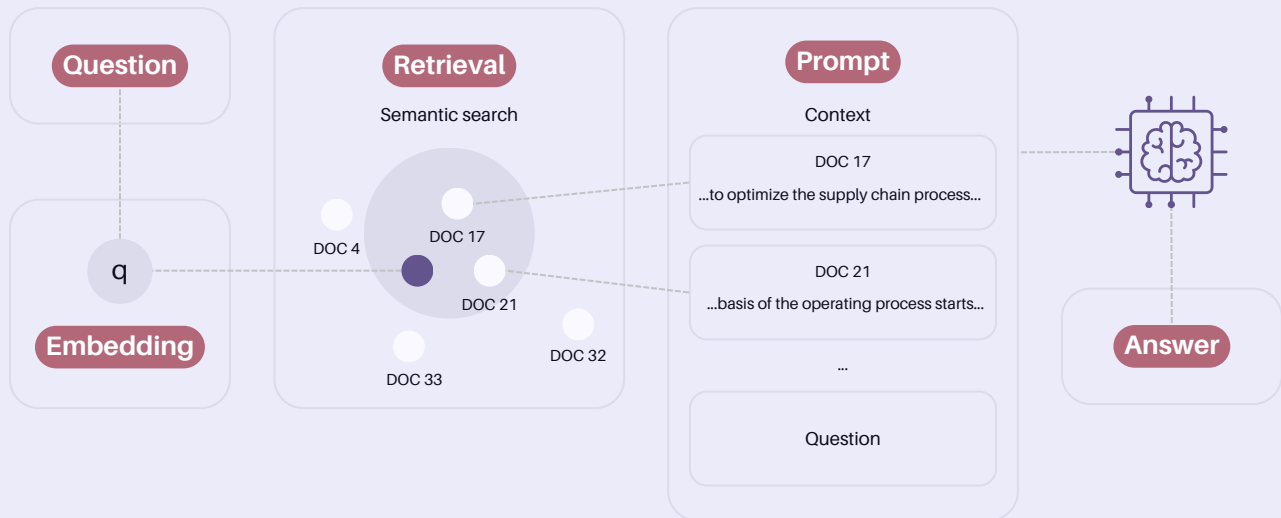
RAGs combine user prompts with databases or documents to enrich context and follow two main steps.

The first step, the retrieval phase, involves exploring data to gather relevant information for the user's question.

This information, along with the question, is then integrated into the prompt sent to the LLM. The second step, generation, allows the LLM to naturally respond to the question, akin to what ChatGPT does.

Although the RAG may seem simple, its impact in business is significant. Unlike traditional language models restricted to sometimes outdated data from months or years ago, the RAG keeps the model up to date.

The operation of RAG can be simplified as follows:



## RAG Preparation: Indexing the Knowledge Base

The goal is to transform knowledge bases of various formats (PDF, Word, PPTX, etc.) into a knowledge base easily and quickly exploitable by information retrieval tools.

This generally involves extracting text from selected documents (PDFs, Word, etc.) and storing it in a database specifically designed for information retrieval.

With the advent of transformers, information retrieval technologies have also improved; today, we talk about semantic search and vector-based databases.

To create a vector database, we simply convert our text into embeddings using "sentence transformer" models. It is advisable to divide the text into paragraphs for model embedding performance (and limitation) reasons.

## Using the RAG

### Step 1 - Retrieval: Retrieving relevant information

To find the most relevant text excerpts to answer the user query, we perform a semantic comparison between the query and a piece of text from our index. For this, we need to:

- Convert the user query into a vector (embedding) using the same sentence transformer model.

- Compare the vector derived from the user query with vectors in our knowledge base and select those that are closest (vector distance).

The output is the text excerpts most semantically close to the query. They will serve as context provided to the LLM.

### Step 2: Final prompt

In this final step of the RAG process, the system takes into account the relevant excerpts from the previous step to formulate the final prompt.

This is the final request or description given to the language model to generate the desired content. It is the last step before the model begins generating the final text in response to this request.

This prompt may include keywords, summaries, relevant excerpts, specific questions, or any other elements that effectively guide the model towards producing the desired content.

The simple RAG process as presented above is highly effective in the ideal case: when the information is similar to that of a dictionary, the questions are factual, and the answer is directly in the corpus.

The RAG is no longer suitable for unstructured or visual data such as presentation slides or documents, for example.


This is where advanced RAG comes in, to amplify the benefits of the RAG by making it more efficient and more suitable for business needs.


Here is the difference in response between a generative AI without RAG and one that uses the RAG model:

Gen AI without RAG

12:55PM

Have we ever worked on generative AI?




 LLM 12:56PM


I'm sorry, but I don't have the necessary information to answer this question.



Gen AI with RAG

12:55PM

Have we ever worked on generative AI?



 LLM 12:56PM

There are in fact two ongoing projects on generative AI:  
 [Generative\\_AI\\_Project.docx](#)  
 [GenAI\\_Consulting.docx](#)

## Limitations of Simple RAG in Business Context

The text parsing approach of documents, segmenting them into paragraphs, and inserting these segments into a vector base for "Retriever-Answer Generation" (RAG) is a common method in natural language processing to facilitate information retrieval and answer generation.

However, this method has certain limitations in terms of preserving and fully representing information, primarily for the following reasons:

- Loss of global context:

When documents are segmented into individual paragraphs, each paragraph is often treated as an independent unit of information.

This segmentation can lead to the loss of the global context or narrative continuity that exists in the entire document. Paragraphs taken in isolation may not contain all the necessary information to understand the subject in depth, or may be misinterpreted without the context provided by adjacent sections.

- Loss of structural and multimodal information:

Business documents may include not only text but also graphics, tables, and other visual or structured elements that are difficult to efficiently encode in a purely textual vector base.

Segmenting into paragraphs and standard textual encoding ignore these elements, which may be essential for understanding the complete information.

- Granularity issues:

The choice of granularity of text segments (paragraphs, sentences, or entire documents) has a significant impact on the performance of the RAG system.

Paragraphs may not be of good granularity for certain types of information or questions, such as when an adequate response requires integration of information from multiple paragraphs or even different documents.

- Challenges related to coherence and continuity:

In cases where responses or information need to be generated or retrieved from multiple vector inputs (paragraphs), maintaining coherence and continuity in the generated responses can be challenging, especially if the segments come from different parts of the document or different documents.

---

## Advanced RAG

There are numerous possible strategies for optimizing the RAG.

Advanced RAG is an evolution of the simple RAG concept, introduced to overcome the limitations of the RAG discussed earlier.

The objective is both to be able to retain 100% of the information in the search tool during indexing, to transform it in a way that an LLM (Language Learning Model) can better understand the context of the information, and to be able to answer a wider range of questions.

Advanced RAG also allows for more precise information retrieval within the knowledge base.

New language processing technologies enable fully leveraging the knowledge base by manipulating and transforming it to extract maximum information.

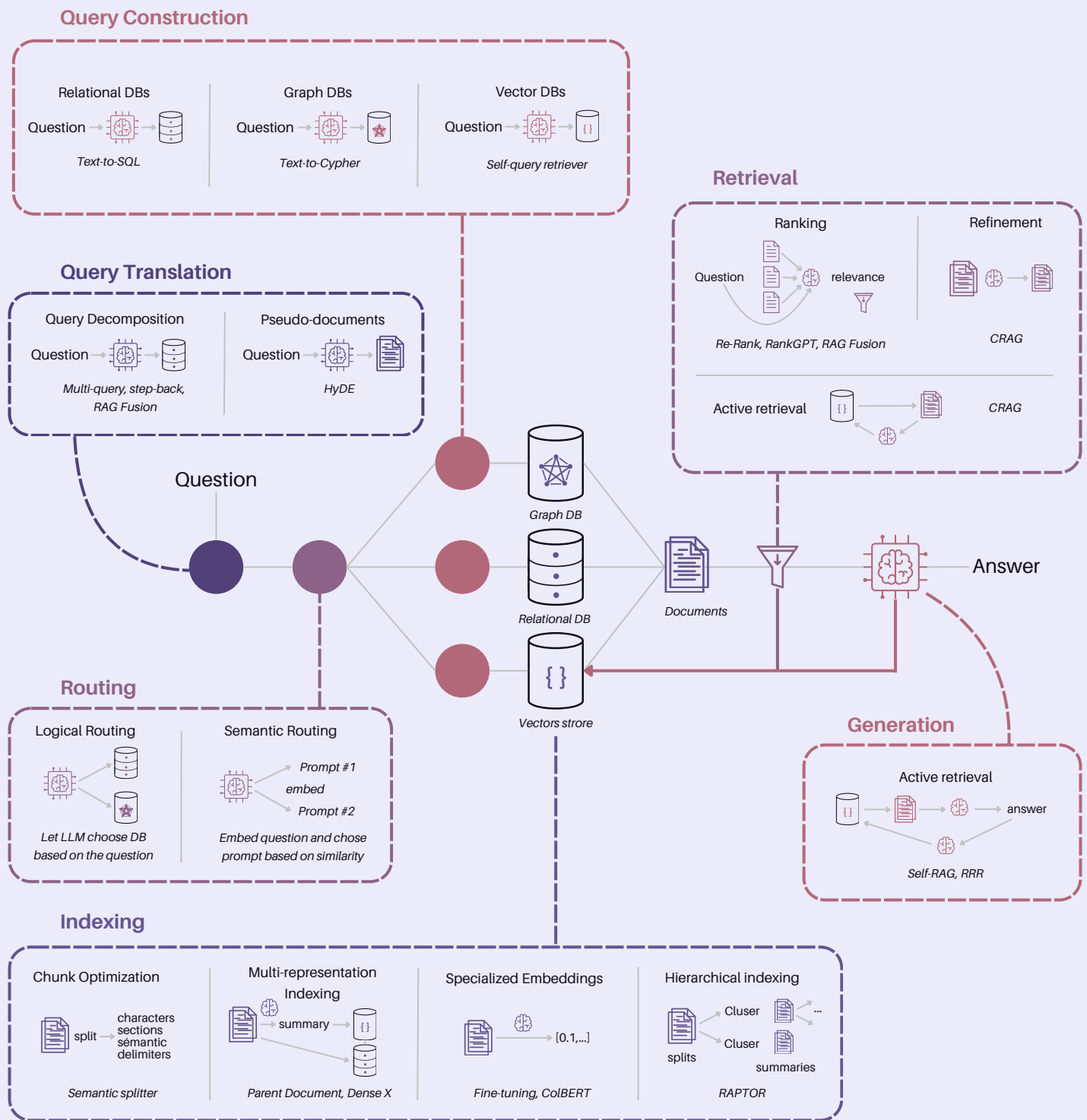
The main properties of Advanced RAG are:

1 - transforming a heterogeneous database into a structured knowledge base with a multi-level vector index.

2 - using multi-step reasoning to efficiently traverse the database and answer complex questions.

Advanced RAG strategies can be grouped into 2 major categories: indexing strategy and information retrieval strategy.

The Advanced RAG can be summarized with this diagram:



The most important stage is the fourth stage: indexing, represented at the bottom in the previous diagram. Advanced RAG then enriches it as follows:

- Multi-level indexing:

This stage summarizes and contextualizes documents. Based on the user query, the most relevant chunk (portion of a document) is extracted and passed to the LLM along with the document it belongs to. This allows for a deeper understanding of the context, more precise extraction of data characteristics, and better adaptability to various tasks and domains.

- RAG vision:

We'll discuss this in more detail in the following section, but this stage enables the textual description of slides or visual documents. It's simply the integration of image processing capabilities into the RAG model. This extension enables the model to understand and utilize visual information in addition to textual data to improve its performance.

- Metadata:

Automatic assignment of metadata by AI to enrich content understanding and improve text generation. Metadata are structured pieces of information that describe, identify, and help organize data. In the context of RAG, they're used to enrich understanding and information retrieval in different data sources (texts, PDFs, visuals...).

- Sequential indexing:

Creation of page summaries from a document. To retain the context of previous pages, we perform sequential indexing, meaning that in addition to the document page, we provide information on the previous pages.

## Information retrieval strategy

To search for the right information, Advanced RAG first modifies the user's initial query. It's enriched and/or translated into a semantically rich query. Metadata is extracted, and the query is broken down into multiple queries.

Multi-index/multi-level search: Based on the user query, the search is then performed in different indexes.

Active retrieval: During information retrieval, an information verification step is added. If the information is not relevant, the algorithm retries by modifying the search parameters.

Here's an example query and the steps Advanced RAG will go through:

Query: "What is the industry sector of our clients?"

Advanced RAG will detect the user's intent from a macro perspective based on its knowledge base, search in the document summaries index. It retrieves a large amount of document summaries related to the company's clients. Finally, it synthesizes the information into a final response categorizing clients by industry type.

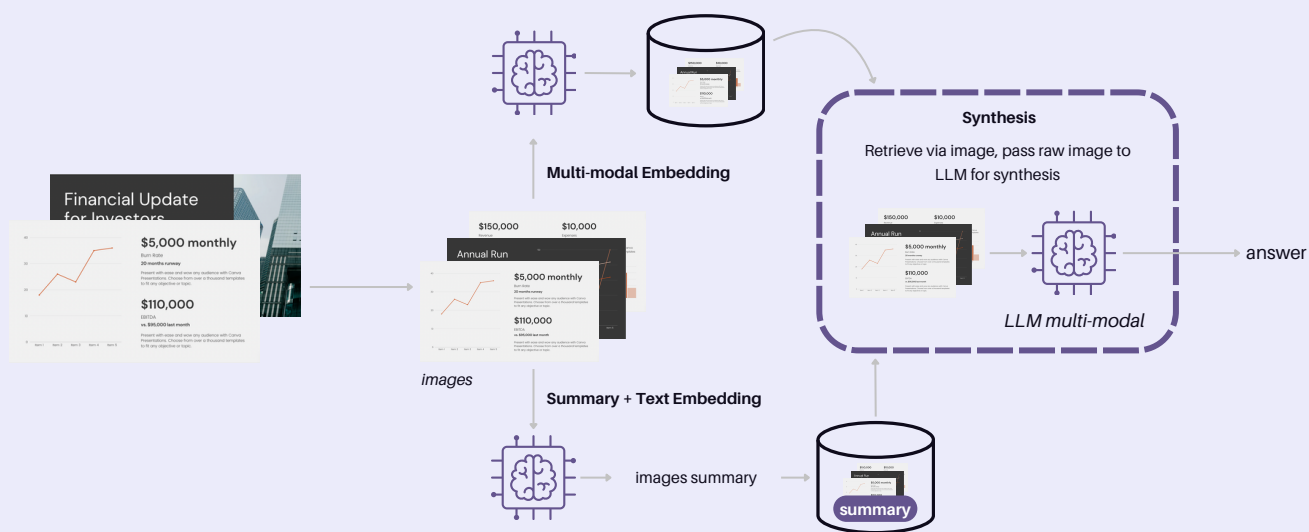
## Zoom on RAG Vision

Within a company, numerous data and insights are housed in presentations in the form of slides or other visual content.

The RAG vision methodology allows for the incorporation of such data into the knowledge base.

This process extracts slides as images, employs an LLM to summarize each image, integrates image summaries with a link to the original images, retrieves the relevant image based on the similarity between the image summary and the user's query, and finally passes everything to the LLM for response synthesis. However, there are two approaches:


### Approach 1 : vectorstore w/ multi-modal embedding



### Approach 2 : vectorstore image summaries + text embeddings

Between these two options, the second, which allows for summarizing the image, is much more efficient. However, it is more complex and costly. Here are the different precision scores resulting from a Langchain study in December 2023:

Approach	Score (CoT precision)
High k RAG (text only)	20%
Approach 1 : vectorstore w/ multi-modal embedding	60%
Approach 2 : vectorstore image summaries + text embeddings	90%

 [See complete study.](#)

### Elqano's perspective

For efficient advanced RAG, the most critical step is enriching and transforming the database using AI. This step is absolutely necessary to effectively address complex questions and enable multi-step reasoning.

## A Powerful RAG to Unlock the Doors to Knowledge Intelligence

The RAG represents a major advancement in how businesses can access and utilize their data to generate actionable insights. It goes beyond simple data extraction and provides relevant answers and insights, thus transforming how businesses leverage their knowledge capital.

The RAG fosters collaboration and knowledge sharing within businesses. By allowing users to ask natural questions and get instant answers, it promotes information exchange and collective intelligence.

Content generation features also facilitate the creation of internal knowledge bases, fueled by contributions from all employees, thus reinforcing a culture of sharing and continuous learning within the company.

It can thus be said that the RAG paves the way for more advanced "knowledge intelligence." By combining search, analysis, and generation capabilities, it enables businesses to make informed decisions, anticipate market trends, and innovate more rapidly.

By providing easy and intuitive access to a wealth of data and knowledge, the RAG becomes a strategic asset for companies looking to remain competitive and innovative.

		Simple RAG	Advanced RAG
Indexing	Documents Chunk	✓	✓
	Documents Summary	✗	✓
	Image Description	✗	✓
Retriever/search	On Chunks	✓	✓
	On summary and descriptions	✗	✓
Generation	From text	✓	✓
	From images	✓	✓

		Simple RAG	Advanced RAG
Simple query	Low number of documents (less than 20)	✓	✓
	Large number of documents (more than 20)	✓	✓
	Very large number of documents (over 10,000)	✓	✓
Complex query	Low number of documents (less than 20)	✓	✓
	Large number of documents (more than 20)	✗	✓
	Very large number of documents (over 10,000)	✗	✗

### Keep in mind

Leveraging the power of LLMs on enterprise data is possible. However, the hostile environment of enterprise knowledge bases and the needs of users by profession require the implementation of robust and sophisticated techniques to ensure smooth operation. The proper connection between enterprise data and LLMs opens the doors to Knowledge Intelligence.



# Part 4



## Benefits and Use Cases in Business

# Benefits and Use Cases in Business

## Benefits of RAG in Business

When properly applied, the RAG model can be a significant ally in business. Here are the main advantages of improving generative AI through RAG technology:

### Technical Benefits

- Up-to-date and accurate responses: RAG ensures that responses are based on current information.
- Transparent responses: RAGs are capable of citing sources.
- Resource efficiency: Models do not need to be retrained with new data.

### Strategic Benefits

- Finding internal information: RAG allows for the quick extraction of relevant information from large internal company databases.
- Time savings: Instead of spending hours manually searching for data, employees can obtain accurate and relevant answers in record time.
- Improved quality of outputs: Employees can create higher-quality documents, reports, or presentations more quickly.
- Better decision-making: RAG provides decision-makers with key information needed to inform their decisions.

#### Keep in mind

In addition to responding to simple queries, advanced RAG can be used for problem-solving, decision-making, and answering more complex questions.

# Industry Use Cases

How can the RAG model serve you in business? Here are some specific use cases for different sectors.

## Consulting Firm

- Market analysis and trend research:

By using RAG to explore external databases, sector reports, and relevant articles, a consulting firm can obtain valuable information on market trends, consumer behavior, and sectoral innovations.

This information can then be used to formulate strategic recommendations for clients, saving consultants significant time.

- Report and document development:

RAG can be used to generate summaries, analyses, and reports from large amounts of textual data.

A consulting firm could use RAG to automatically generate customized reports by aggregating and synthesizing information from various sources.

Example Query: "Summarize all recommendations we have made for client X and decipher the results to identify the best recommendations."

## Advertising/Digital Agency

- Creative content generation:

The agency can use RAG to generate advertising campaign ideas by integrating information from external databases, such as market trends, consumer behavior, and past company successes.

It can also use RAG to generate advertising scripts based on examples of effective ads in the same sector or using data on the preferences and interests of the target audience.

- Identification of best practices and past successes:

In response to a tender, RAG can help the agency extract examples of best practices and past successes in similar projects, using data from its own project history or retrieving information on achievements from other agencies in the same field.

Example Query: "Analyze our responses to successful tenders in the FMCG sector and formulate the best possible response to a new tender in the sector."

## Investment Fund

- Portfolio selection:

By integrating data from various sources, such as annual reports, press releases, and financial analyses, RAG could help identify companies or sectors with the best potential return for the fund's portfolio.

It could also be used to assess the risks associated with each potential investment.

- Due diligence:

During due diligence, the fund could use RAG to retrieve and analyze key information, such as financial history, past performance, stakeholder relations, etc.

This would provide a more comprehensive overview and identify potential risks or opportunities.

- Investment monitoring:

Once investments have been made, RAG could be used to continuously monitor the performance of portfolio companies, market trends, and events likely to impact investments. This would enable informed and real-time decision-making.

Example Query: "Summarize the last 10 boards of startup X" or "Analyze these two startups in the same sector and provide an in-depth comparison of their performances."

## Legal Firm

- Enhanced Legal Research:

One of the most obvious use cases for RAG in a legal firm is to enhance legal research. RAG could be used to retrieve relevant legal information from vast databases of case law, statutes, and legal precedents.

This information could then be used to generate summaries, analyses, or responses to specific client inquiries.

- Contract and Legal Document Analysis:

RAG could be utilized to analyze and interpret contracts and other legal documents. It could help identify potential risks, spot ambiguous clauses, and provide guidance on drafting or negotiating contracts more effectively.

- Assistance in Drafting Legal Documents:

RAG could be used to provide assistance in drafting legal documents such as motions, pleadings, or legal opinions.

By using generative language models, RAG could help lawyers draft documents more quickly by providing formulation suggestions, relevant examples, and legal references.



# Part 5



## **Solution Using Advanced RAG: Enterprise Chat "Playground" Connected to Internal Data**

# Solution Using Advanced RAG: Enterprise Chat "Playground" Connected to Internal Data

## What is it for?

Playground is our latest innovation at Elqano, integrating Large Language Models (LLMs) with advanced internal data search to enhance work efficiency. Through a chat-GPT interface, Playground enables real-time content generation, discussion with data, task automation, and quick access to internal information.

Playground utilizes the concept of advanced RAG by combining multi-level vision sequential indexing (refer to the advanced RAG section) and strategies also presented in the section for managing complex user queries.

Playground's Intelligent Search function allows quick access to internal information. Designed to be efficient, it assists employees in swiftly finding the data they need.

*Some examples of Playground solution usage:*

	HR	Finance	IT
SIMPLE QUERY	Can you show me the results of the latest satisfaction survey?	What is the status of the marketing department's budget this month?	What are the latest security protocols for remote access?
	Sales	Legal	Marketing
SIMPLE QUERY	Where can I find the list of leads from the last trade show?	What is the status of the company's latest patent filings?	What are the returns on the last email campaign?

	HR	Finance	IT
COMPLEX QUERY	Can you analyze trends in staff turnover rates over the last five years and identify departments at risk?	Can you compare current spending with the allocated budget and identify significant variances for each department?	What impact have recent security updates had on the company's network performance?
	Sales	Legal	Marketing
COMPLEX QUERY	Can you perform a profitability analysis of different customer segments to identify the most profitable and those at risk?	Can you analyze legal case processing times over the past year to identify processes that require optimization?	What were the returns on investment (ROI) for each marketing channel used over the past year?

### And now?

Would you like to find out more about these topics from our advanced RAG experts? [Contact us now!](#)

## To Go Further

[Advanced RAG series: Indexing](#)

[Multi-modal RAG on slide decks](#)

[Answer the Top Generative AI Questions](#)

[Assess ROI for Generative AI](#)

[AI in Consulting](#)

[Gen AI RAG use cases that can deliver quick impact](#)

