

# Unleash the Power of Generative AI Systems in Your Enterprise

# Table of Contents

Introduction / 3

AI: Is This the New Business Reality? / 4

AI Use Cases Across Industries / 5

What Are Large Language Models, and Why Are They Important for AI? / 5

How LLMs Work / 7

The Role of Data in AI Models / 8

The Merging of AI and Enterprise Data / 8

Examples of AI Platforms Used in Business / 9

The Risks and Challenges Involved with Enterprise Data and AI / 10

How MarkLogic and Semaphore Assist with Using Enterprise Data for LLMs / 12

Generative AI and Progress Data Platform / 17

Knowledge Graph Database vs. Vector Database / 19

Use Cases: MarkLogic Capabilities & Generative AI / 20

Fine-Tuning AI Models: Is It Worth it? / 20

Applying Retrieval Augmented Generation (RAG) in the Organization / 21

Navigate the Risks of Generative AI / 21

Conclusion / 22

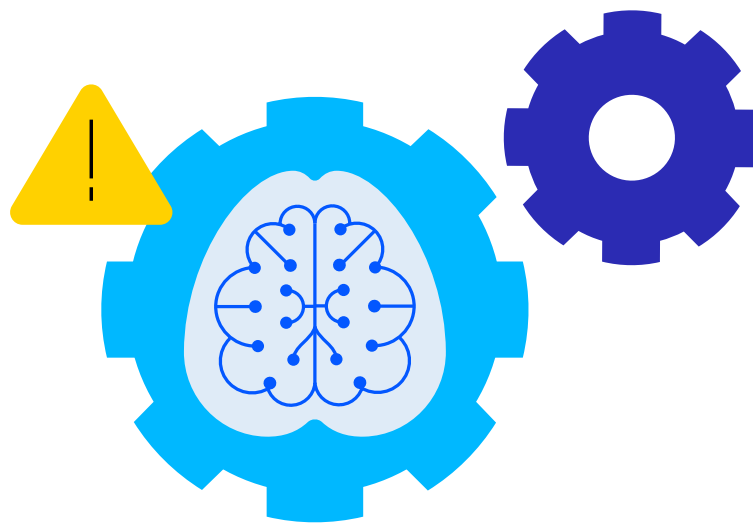
# Introduction

The world has been fundamentally disrupted by the changes generative AI and large language models (LLMs) have brought to life—and businesses are increasingly recognizing the significance of this paradigm shift. Those who harness the possibilities and opportunities of AI and digital technology will be the ones to succeed and flourish in this new digital landscape. AI platforms, like ChatGPT, Bard, Gemini and Claude, can provide answers and insights using suitable LLMs. Combined with private business data, LLMs can further refine and generate more correct output. The convergence of AI and enterprise data is not a new phenomenon. Many enterprises are now experimenting with AI to get on the innovation curve and outperform their competitors.

However, there is an enormous responsibility when using sensitive data or leveraging your proprietary data. With the growing relevancy of AI platforms, businesses need to consider ethical and regulatory requirements surrounding data privacy, fairness, explainability, transparency, robustness and access.

[Progress® MarkLogic®](#) and [Progress® Semaphore™](#) can be used to leverage enterprise data with generative AI and help businesses derive further context, meaning and insight into their data, as well as improve and contextualize the results.

If you are considering combining your enterprise data with a generative AI tool or need more information on the topic, this guide is the perfect place to start.



# AI: Is This the New Business Reality?

## Defining Artificial Intelligence (AI)

AI refers to the science of developing technologies that replicate human cognitive functions, like reasoning, learning, problem-solving and more—enabling them to perform tasks that typically require human intelligence.

Leveraging AI in business provides many benefits, including better decision making, faster time to market, enhanced customer experience, more efficient operations and business growth.



**“AI won’t take your job.  
It’s somebody using AI  
that will take your job.”**

Richard Baldwin,  
Economist and professor at the  
Geneva Graduate Institute in  
Switzerland

The Artificial Intelligence (AI) winter first described in the 1980s is officially over. Its spring has just begun and, with it, products and services democratizing the availability of AI to worldwide business and organizations. It is not just a technological revolution—it is the new business reality that empowers businesses to achieve sustainable competitive advantage. Businesses who choose not to embrace AI and innovate in this space run the suggest with risk of being left behind.

Businesses across different industries are investing in AI. Similar to many other cutting-edge technologies, AI will always raise some questions, and companies that are deploying it will face challenges.

Businesses exploit the opportunities of AI to:

- Quickly and accurately analyze and interpret vast amounts of data
- Automate business workflows and processes
- Reduce human error in data-processing activities
- Gain valuable insights to streamline their decision-making processes
- Accelerate research and development endeavors

# AI Use Cases Across Industries

Businesses from different industries are now implementing AI technology to improve their processes, automate tasks and get insights from data.

<b>Manufacturing</b>	<b>Healthcare</b>
Demand forecasting	Drug discovery and development
Optimization of inventory	Medical imaging
Development of production scenarios	Disease diagnosis
Research and development	Enhanced patient care
—	Medical research
<b>Energy &amp; Utilities</b>	<b>Banking</b>
Chatbots and virtual assistants	Fraud detection
Data analytics	Risk management
Risk mitigation	Predictive analysis
Demand forecasting	Credit evaluation
Energy optimization	Financial analysis

## What Are Large Language Models, and Why Are They Important for AI?

Large language models (LLMs) present a significant shift in the realm of AI, transforming the business environment that we know. LLMs are powerful generative AI models trained on a massive library of information to execute diverse natural language processing functions, encompassing language translation, question answering, summary creation and sentiment gauging. In contrast to the AI systems that perform tasks like data classification, data categorization and data automation, the primary function of LLMs is to generate content. These models are designed to analyze language in a way that mimics human intelligence, allowing them to process, understand and generate human speech. They can perform complex tasks based on skilfully crafted input text that prompts the model to tackle the specific tasks.

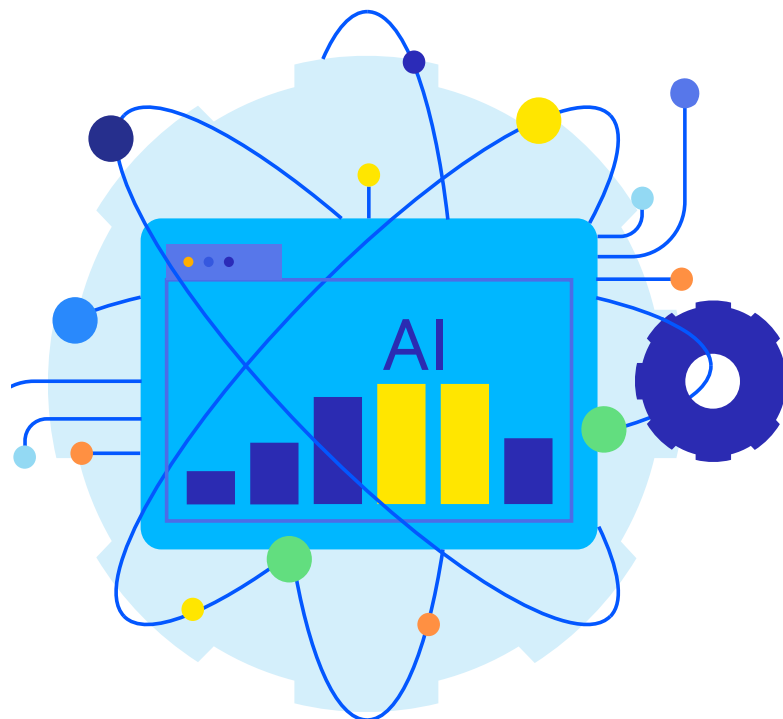
LLMs are also referred to as neural networks, text predictors, agents, generative AI, writing assistants, chatbots, simulators, content generators, etc. What all these systems have in common is their content-creation function.

The models are trained on vast amounts of data, allowing them to detect patterns and make predictions that would be difficult or impossible for a human to do manually. The information ingested, or entered, can be private or proprietary data, or in the case of ChatGPT, data that comes directly from the internet. This has many potential applications in healthcare, finance and customer service.

However, the complexity of these models comes with ethical and technical challenges for businesses, including data bias, copyright infringement and potential libel cases. Essentially genAI systems are hacking the 200,000-year-old human language system.

Characteristics of generative AI platforms:

- Mimic human conversations and respond to users' queries
- Create content
- Translate languages
- Summarize texts
- Code generation



# How LLMs Work

LLMs are built upon neural network architectures. They mainly use transformative architecture that includes several levels of self-attention mechanisms responsible for determining the importance of words and elements in a sequence of text. The input text is broken down into tokens, which are smaller units representing words, elements and characters. Each token has a numerical representation. For example, in the images below we can see that the word “the” is included in the text three times, and its numerical value of 262 also shows up three times in the corresponding spots.

**Tokens**      **Characters**  
**99**            **465**

```
[17875, 5623, 2671, 1229, 577, 37484, 29362, 4013, 1326, 14358, 262, 1790, 12, 4354, 357, 4761, 2250, 284, 352, 1285, 8, 23584, 779, 286, 12948, 15701, 9904, 4410, 35031, 3306, 329, 6506, 351, 12593, 508, 423, 2035, 286, 262, 1708, 2761, 287, 12755, 2910, 15701, 1241, 11, 555, 39772, 284, 10224, 14588, 10742, 15068, 25, 21209, 34619, 344, 20730, 17261, 1108, 26, 393, 30558, 515, 5328, 34619, 344, 20730, 357, 1203, 621, 2026, 10527, 14, 45582, 8, 290, 8718, 10853, 344, 20730, 357, 18223, 263, 621, 6640, 10527, 14, 45582, 8, 379, 262, 976, 640, 1123, 1110, 13, 220]
```

TEXT      TOKEN IDS

**Tokens**      **Characters**  
**99**            **465**

```
Continuous Glucose Monitoring Devices Acme considers the short-term (72 hours to 1 week) diagnostic use of continuous glucose monitoring devices medically necessary for persons with diabetes who have either of the following problems in controlling blood glucose level, unresponsive to conventional insulin dose adjustment: Hypoglycemia unawareness; or Repeated hypoglycemia (less than 50 mg/dL) and hyperglycemia (greater than 150 mg/dL) at the same time each day.
```

TEXT      TOKEN IDS

A helpful rule of thumb is that one token generally corresponds to ≈4 characters of text for common English text. This translates to roughly 3/4 of a word (so 100 tokens ≈75 words).

Once trained, an LLM’s primary aim is to create predictive models for generative word sequences, using the context to predict the next token. Based on its learned patterns, the model generates text word-by-word and token-by-token. Under the hood, the core operational generative AI code is designed to take in the tokens in the context window (the short-term memory), combine tokens with its model and predict the next most probable token (usually using a GPU). This then repeats until the desired number of tokens is produced or the prediction indicates that it should stop and await further input.

# The Role of Data in AI Models

Data has a crucial role in AI models, ensuring their accuracy and reliability. The development of AI applications requires the adoption of copious amounts of data, both public and private, for training, testing and the evaluation of applications. For this reason, it is essential that data used for training the model is accurate and up to date.

Generative AI is effective when it comes to asking general questions and it delivers good results. But if you ask LLMs to compare your company's performance last year to this year, it will fail terribly. This is one of the biggest challenges of LLMs: they have a good public and general knowledge, but they are unaware of and lack access to proprietary and non-public information. Proprietary knowledge is important for businesses, as having access to it, LLMs will be able to deliver more trustworthy and quality results.

## The Merging of AI and Enterprise Data

In recent years, organizations have been challenged to manage and interpret the large amounts of business data they possess and make adequate actions in a timely manner. With the implementation of AI, companies can interpret and act upon their data by embedding LLMs in their business-critical applications and systems.

Training AI models with enterprise data can lead to greater insights and accurate predictions simply because others don't have access to this same data. Proprietary data is unique and exclusive to companies, and exploiting its capabilities can provide a competitive advantage to businesses.

Companies with access to confidential or sensitive information, such as healthcare companies legally allowed to use medical records, have a competitive advantage when building AI-based solutions.



**“The winners of AI will be those who know how to use AI and have private data.”**

Cathie Wood,  
CEO ARK Investment Management



# Examples of AI Platforms Used in Business



## ChatGPT

[ChatGPT](#), launched in November 2022 by OpenAI, represents a generative AI platform designed to mimic human conversations and generate business information and responses to human queries. ChatGPT was trained on an extensive collection of human knowledge, from which the model is able to produce engaging and persuasive answers, create content and generate code. The current version of ChatGPT supports 4K tokens, while new versions are expected to have bigger token buffers.



## Google Bard

[Google Bard](#), released in March 2023, is a conversational AI chat service. Bard is powered by Google's proprietary LLM, PaLM 2 and Google's Language Model for Dialogue Applications (LaMDA). Similar to other chatbots, Bard is capable of coding and creating and summarizing content. The current version of Google Bard supports 1K tokens, compared to ChatGPT's 4k-large token library.



## Anthropic Claude

[Anthropic Claude](#), released in 2021, is a chatbot with capabilities similar to Bard and ChatGPT. Anthropic has spent significant time improving the trustworthiness of their AI platform by focusing on the dangers of AI and training their AIs to be "helpful, harmless and honest." An important characteristic of Claude is its ability to allow users to delete conversations and support VPN browsing. The chatbot supports 100,000 tokens, which is significantly more than other chatbots.

# The Risks and Challenges Involved with Enterprise Data and AI

One underlying question linked to these LLMs is the level of transparency and the extent to which they provide accurate and unbiased data. Due to the unpredictable nature of LLMs, there are numerous concerns around ethics, compliance and governance surrounding generative AI. While it can effectively enhance and contribute to many real-world use cases, generative AI does come with its fair share of downsides.

- **Hallucinations**

A hallucination, as Britannica states, is “an experience involving the apparent perception of something not present. When it comes to AI, a hallucination happens when AI reports error-filled answers to the user. In human terms, these are like memory errors or lies. Even though these error-filled answers do sound plausible, the information may be incomplete or altogether false. For example, if a user asks a chatbot about the average revenue of a competitor, chances are those numbers will be way off. These kinds of errors are a regular occurrence. The rate of hallucinations that ChatGPT experiences varies between 15% and 20%, which should be considered when querying your AI (Datanami, 2023).

- **Fairness/Biases**

Data bias impacts day-to-day businesses and occurs when the available data fails to represent the entire population of the phenomenon. This diminishes the fairness and equity of the systems as they may produce results that reflect the biases encoded in the training data rather than present objective reality.

A recent [study on Data Bias](#) by Progress highlights the following tendencies:

- 65% of businesses and IT executives currently believe there is data bias in their respective organizations
- 13% of businesses are currently addressing data bias
- 78% believe data will become a bigger concern as AI/ML use increases

For AI systems to deliver unbiased results, they need to be trained on unbiased data. While the potential benefits of these models are immense, users should carefully examine the ethical and practical considerations.

- **Reasoning/Understanding**

Even though LLMs demonstrate exceptional natural language abilities, they are struggling with logical reasoning and understanding of complex concepts. Certain queries that are common sense for human beings may confuse LLMs as they are missing this information in the training data. LLMs can lead to stereotypical or incorrect answers if not carefully monitored and trained.

- **Data Cutoffs**

Data cutoffs present a major challenge for LLMs, impacting the model's ability to accurately understand and respond to users' queries. Considering the extensive time required to train the model, its memory can quickly become outdated. When LLMs have limited access to information, they may produce answers that do not incorporate recent trends and developments, affecting its accuracy and relevance.

- **Explainability**

How LLMs generate responses can be questionable. They should be trained or prompted to show their reasoning and reference to the data they used to construct a user's response. A traceable AI-generated response will enhance users' trust in the model and provide accountability.

- **Robustness**

LLMs, like any other technology, should focus on maintaining their performance even when they are challenged with unexpected or conflicting inputs or situations. If we can successfully address these issues, the trustworthiness of our solutions will increase along with user satisfaction, ultimately leading to the solution's success.

The ground-breaking generative AI applications like ChatGPT are quickly entering the business world. The effect of these technologies on businesses is profound. The ability of generative AI to process vast amounts of data allows businesses to “know” everything about their organizations—and all the information a business stores can be harnessed to optimize business operations. That said, how can someone receive more accurate answers when using ChatGPT? Can a user influence a language model with private data to obtain correct answers? The short answer is yes.

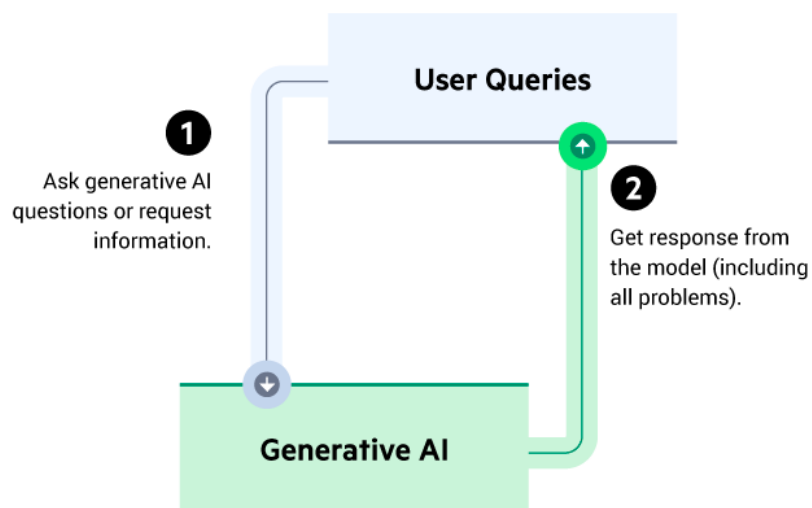
**This is possible through the connection of generative AI models with a business's proprietary data.**

# How MarkLogic and Semaphore Assist with Using Enterprise Data for LLMs

Progress MarkLogic is capable of storing and querying structured and unstructured data. Additionally, Progress Semaphore can capture subject matter expert (SME) content via its intuitive GUI. The resulting knowledge graphs can extract facts found within the data and can tag the enterprise data with semantic knowledge. Semaphore can also use this semantic knowledge to start tagging user questions/inputs and specific LLMs, answers with this knowledge. Users can then use MarkLogic and Semaphore to fetch semantically relevant enterprise data for the LLM.

The best way to compare how MarkLogic and Semaphore work in conjunction with LLMs is to think about closed-book and open-book exams.

## 1. Closed-Book Exam Model

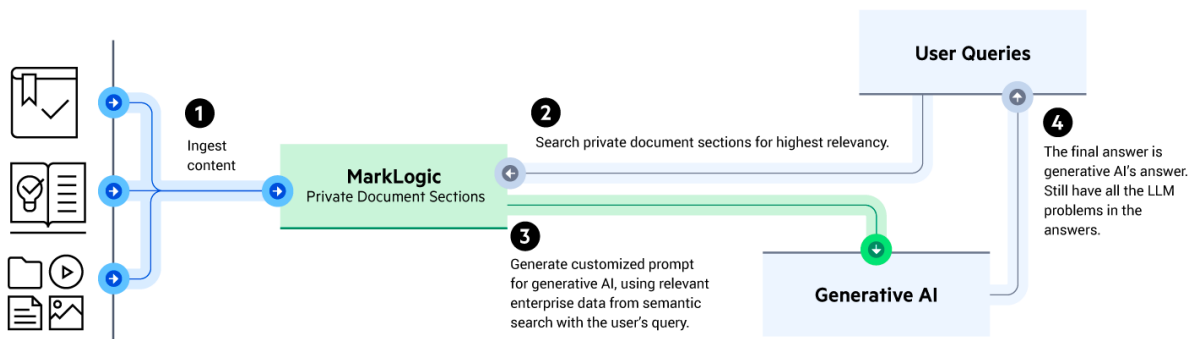


The closed-book exam model can be described in two steps:

1. The user asks the LLM model the question or requests information.
2. The model provides answers to these inquiries based on the knowledge embedded within the language model that is currently being used.

Without semantic and tagged data, users may not get entirely accurate information and may not even be aware of a response's inaccuracies.

## 2. Open-Book Exam Model with the Ability to Get to the Right Page

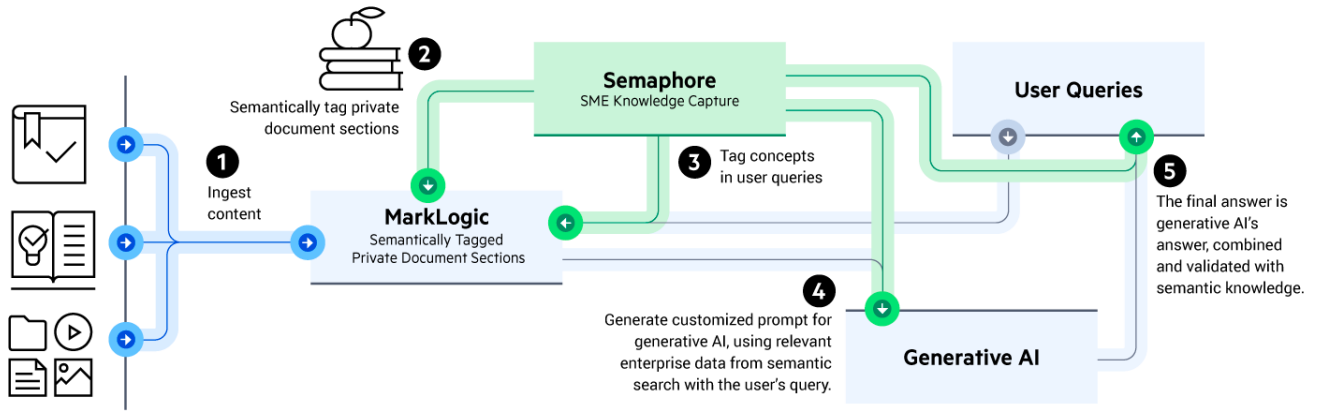


The open-book exam model, using MarkLogic's abilities alongside an LLM model, can be described in four steps:

1. Content is ingested into MarkLogic and segmented into smaller sections of the documents so that search results can fit into the model's prompt window.
2. MarkLogic searches for the most relevant private document sections based on the user's question.
3. The middle tier or MarkLogic can then generate a customized prompt for the model, using relevant private data from search with the user query.
4. The model provides the final answer.

With MarkLogic, users can increase accuracy and efficiency while introducing governance. Despite the increase in accuracy and efficiency, LLMs—and their users—may still need help with answers, especially given the data could lack context and meaning.

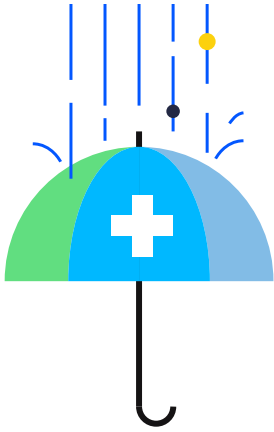
# 3. Open-Book Exam Model with Expert Help



The open-book exam model with expert help can be described in five steps:

1. Content is ingested into MarkLogic and segmented into smaller sections of the documents so that search results can fit into the model's prompt window.
2. Semaphore semantically tags, categorizes and fact-extracts private document sections.
3. Semaphore tags key concepts in user queries.
4. The middle tier of MarkLogic can then generate a customized prompt for the model, using only semantically relevant private data from semantic search with the user query.
5. The final answer is a combination of the model's answer and semantic knowledge from Semaphore to validate it.

The combined results will allow LLMs and users to easily access and fact-check the results against the source content and the captured SME knowledge.



# Example Use Case with ChatGPT & Claude

The case: A healthcare insurance company has released a plan, accompanied by documents, outlining specific benefits coverage. Members, providers and customer support agents can query the plan documents for elements like coverages, conditions treated, etc.

The company compares the results of the LLM models and their performance in conjunction with MarkLogic and Semaphore.

Example Question: Does Acme Insurance pay for CGM?

## Chat GPT Results

Closed-Book Exam Model	Open-Book Exam Model with Enterprise Data	Open-Book Exam Model with Semantic Enterprise Data
<p>I cannot provide specific insurance coverage information for Acme Insurance, as coverage options vary by plan and depend on a variety of factors. You should contact your insurance provider or healthcare provider for more information about your specific coverage options.&lt; im_end &gt;</p>	<p>Acme Insurance covers CGM devices for people with diabetes who meet certain criteria. To be eligible for coverage, you must meet Acme's coverage criteria, which may include having a diagnosis of diabetes, requiring insulin therapy, and having documented episodes of hypoglycemia.</p>	<p>Yes, Acme may provide coverage for continuous glucose monitoring (CGM) devices for individuals with diabetes who meet certain criteria. Some criteria may include frequent episodes of hypoglycemia (low blood sugar) or hypoglycemia unawareness, or the inability to control blood glucose with traditional methods of monitoring. However, specific coverage policies can vary depending on the individual patient and the specific device being used. Patients should check with their healthcare provider or insurance provider to determine whether a particular CGM device or service is covered under their specific plan, and to learn about any costs, restrictions, or documentation requirements that may apply.&lt; im_end &gt;</p>

## Claude Results

Closed-Book Exam Model	Open-Book Exam Model with Enterprise Data	Open-Book Exam Model with Semantic Enterprise Data
<p>Unfortunately, I do not have specific coverage information for Acme Insurance and CGMs. Health insurance plans can vary greatly in their policies for covering diabetic supplies like continuous glucose monitors. Some key factors in CGM coverage include:</p> <ul style="list-style-type: none"> <li>• Type of plan (HMO, PPO, etc) and whether CGMs fall under the plan's durable medical equipment benefit.</li> <li>• Requirements for prior authorization or documentation of medical necessity. Approved CGM brands and models covered by the plan.</li> <li>• Cost-sharing amounts (deductible, coinsurance, copays) applicable to CGMs.</li> <li>• Limits on the frequency of CGM replacement. Whether Acme requires the use of an in-network durable medical equipment provider.</li> </ul> <p>The best way to find out about Acme's current coverage for CGMs would be to contact their customer service department directly or check their online member portal. You can inquire about their prior authorization process, covered devices, and any out-of-pocket costs. Many details may ... also be available in the plan's coverage documents. I recommend connecting with Acme directly for specifics on CGM benefits. Please let me know if you need any other assistance!</p>	<p>Based on the reference documentation provided earlier, Acme Insurance considers both short-term diagnostic and long-term therapeutic use of continuous glucose monitors (CGMs) medically necessary under certain circumstances for members with diabetes.</p> <p>Key points about Acme's CGM coverage: For short-term diagnostic use, Acme covers 72 hours to 1 week of CGM monitoring up to 2 times per year for members having trouble controlling blood glucose levels.</p> <p>For long-term use, Acme covers CGMs for adult members with type 1 or type 2 diabetes on intensive insulin therapy who meet criteria related to hypoglycemia, glycemic control, and adherence. Acme also covers long-term CGM use for children/adolescents with type 1 or 2 diabetes. Implantable CGM sensors and closed-loop pump/CGM systems are considered medically necessary alternatives.</p> <p>CGMs are still considered investigational by Acme for patients not on intensive insulin regimens. So in summary, yes Acme Insurance does provide coverage for medically necessary CGM services for eligible members based on their diabetes treatment regimen and utilization criteria. Let me know if you need any clarification!</p>	<p>The reference document indicates that Acme Insurance considers CGMs medically necessary and provides coverage for certain situations, including:</p> <ul style="list-style-type: none"> <li>• Short-term diagnostic use of CGMs for 72 hours to 1 week in cases of hypoglycemia unawareness, recurrent hypoglycemia, or diagnosing conditions like congenital hypoglycemia in infants.</li> <li>• Long-term therapeutic use of CGMs for adults with type 1 or type 2 diabetes on intensive insulin therapy who are not meeting glycemic targets or experiencing hypoglycemia.</li> <li>• Long-term use for children and adolescents under age 18 with type 1 or type 2 diabetes.</li> <li>• Use of CGMs integrated with insulin pumps that have a low glucose suspend feature.</li> <li>• Acme Insurance does consider long-term CGM use experimental and investigational for individuals with type 2 diabetes not on intensive insulin regimens.</li> </ul> <p>So in summary, yes Acme Insurance does cover CGMs in many situations following the medically necessary criteria outlined in the reference document. But there are some limitations on long-term use for type 2 diabetes patients not on intensive insulin therapy.</p>

MarkLogic and Semaphore enhance the overall user experience with ChatGPT and Claude, improving AI trustworthiness.

**Note: Given their popularity, ChatGPT and Claude are used only as an example of an LLM. This particular LLM may or may not fit your specific enterprise use cases or security needs.**



# Generative AI and Progress Data Platform

Different mechanisms can be embedded to improve the capabilities and enhance the short-term memory of LLMs, allowing them to store information from an external source and use it when generating responses. To achieve this, LLMs can be integrated with a third-party database like MarkLogic and a semantic knowledge management tool like Semaphore in order to achieve the following benefits:



## Holds Specific Private or Proprietary Data

Generative AI's capacity to retain specific private or proprietary semantic data enables enterprises to gain competitive, organization-specific insights. In essence, semantically tagged data acts like an associative memory for generative AI, enabling natural language questions against the most relevant private enterprise data. By ingesting and processing private or proprietary data, the generative AI model is enhanced with an understanding of the company's products, services, customers and internal processes. The data is updated and retrieved in real time, solving the generative AI training data cut-off problem, while older and outdated data becomes less relevant.



## Human Readable

The integration of generative AI with a unified data platform enables greater transparency and provides the ability to reference and examine specific URIs of the private business data used to create the generative AI answers. As a result, the system becomes easier to debug and allows the easy tracking and review of the actual prompt submitted to generative AIs. This creates human readable audit trails that are required in regulated environments, which can be used to further train the generative AI system.



## Generative AI Model Independence

By using a data management platform as a main data repository, organizations can use the same memory and data against multiple generative AI models. They can swiftly switch between different generative AI models, and this doesn't require a complete re-indexing of the data upon changing the model.



## Tunable Use-Case Dependent Relevancy

Organizations can customize the model according to their business needs and specific use cases by providing just the right semantic data. Businesses can build a solid enterprise data architecture instead of relying on a one-off AI initiative—and achieve a single, unified view of their data. Data can be integrated with any platform, helping to promote data consistency and reducing data silos.



## Adheres to Enterprise Standards

Organizations can enhance their data privacy and control the data generative AI is using. AI will only be fed data that the user's roles or query's rules permit, which means that it tightly couples role-based or query-based security to the user's answer, and the short-term memory never gets data it shouldn't. These AI systems adhere to enterprise standards for governance, lineage and provenance. This helps them operate within the established parameters, company protocols and security measures, thereby helping to safeguard enterprise data.



## Enhanced Data Quality Tools

Good quality data is the foundation for good generative AI answers. Companies can implement processes such as data harmonization, data deduplication and data mastering to promote data consistency from widely varying sources and reduce the amount of unnecessary data fed to the generative AI model. By using aggregation and analytics to detect data biases and hallucinations in generative AI's responses, businesses can improve data quality during the entire data journey.

# Knowledge Graph Database vs. Vector Database

The unpredictable nature of LLMs can be solved through retrieval augmented generation, grounding LLM's responses on facts. There are two complementary solutions to this: knowledge graph databases and vector databases.

Knowledge graph databases, like MarkLogic, in conjunction with generative AI tools, perform better with textual data thanks to their flexibility and rapid reusability of data. This provides better context, explainability and accuracy.

Vector databases are more beneficial for indexing and retrieve non-textual modes of data like audio, images, videos, real-time fMRI feeds, etc. However, any textual metadata of audio, images and video files, including descriptions and transcripts, also benefit from being stored in the semantic graph search platform.

Description	Unified Data Platform	Vector DB
Provides external long-term memory	Yes	Yes
Human readable	Yes	No
GenAI model version independent	Yes	No
Use case tunable relevancy	Yes	No
Clustering	Yes	Yes
Similar query	Yes	Yes
Semantic / graph queries	Yes	No
Term co-occurrence queries	Yes	No
Customizable relevancy	Yes	?
Customizable similarity search	Yes	?
Customizable clustering parameters	Yes	?
Data harmonization, data deduplication / smart mastering	Yes	No
Role and query-based security	Yes	No
Governance, lineage, provenance	Yes	?

# Use Cases: MarkLogic Capabilities & Generative AI

Use Case	Description	MakLogic Capability
Search	Results are ranked by relevance to a query string	Built-in search function
Clustering	Text strings are grouped by similarity	Built-in clustering function
Recommendations	Items with related text strings are recommended	Built-in similarity functions
Classification	Text strings are classified by their most similar label	Built-in SVM functions and Semaphore model-driven classifications; ONNX based model execution
Anomaly detection	Outliers with little relatedness are identified	Built-in clustering will help identify outliers
Diversity measurement	Similarity distributions are analyzed	Search scores for diversity; Semaphore classification scores

## Fine-Tuning AI Models: Is It Worth it?

LLM fine-tuning describes the adjustment of a model that's already been trained to perform a specific task or the improvement of its current performance on a particular set of data.

### Advantages:

- Enhances the LLM's output for specific tasks
- Doesn't require extensive data sets

### Disadvantages:

- Requires a lot of technical expertise
- Data/IP leakage and security concerns
- Expensive implementation for enterprises
- Currently offered on a limited basis by OpenAI or Google for their larger models
- Best results are based on trial and error

# Applying Retrieval Augmented Generation (RAG) in the Organization

Retrieval Augmented Generation allows companies to optimize LLMs' answers with relevant information without modifying the model. With RAG, users are augmenting prompts by incorporating enterprise data, guiding the model to generate responses based on a specific context. By leveraging RAG with proprietary data, users can:

- Enhance data privacy with internal data control
- Customize the model according to their business needs
- Improve accuracy with unbiased data results and significantly less hallucinations
- Control the content generation and achieve data transparency
- Achieve consistent results and trust with the LLM model

## Navigate the Risks of Generative AI

The rapid adoption of LLM tools brings both opportunities and challenges for data protection. The use of LLMs and AI tools can increase the risks of data breaches, and safeguarding sensitive information is becoming a mission-critical task for businesses. The convergence of AI and LLMs necessitates a more holistic approach to data protection, including compliant data handling practices, access controls and robust encryption.

The right balance should be found between exploiting the power of generative AI models for business purposes and protecting the rights of individuals and businesses.

**MarkLogic provides additional layers of security, helping to elevate data protection, quickly implement data policies and support governance.**

# Conclusion

Although the potential of AI is substantial, the field of AI is far-reaching and complicated. Many organizations will require the help of technology partners to navigate them along their AI journey. With access to the proper technology and expert guidance, companies can invest in generative AI technology confidently, allowing them to harness the power of AI in the future while gaining both a short-term and sustainable long-term competitive advantage.

Some questions to ask yourself: :

- Are you considering using generative AI?
- Do you want to combine your private / proprietary enterprise data with generative AI?
- Do you want to reduce generative AI problems and issues (hallucinations, bias, transparency, governance, etc.)?
- Have you considered the security and privacy implications of using generative AI?
- Do you want to deploy your own private generative AI?
- Do you want to build out long-term memory that is independent of the generative AI model you use?

If you can answer all questions, then you will go a long way to solving some of the fundamental issues standing in the way of generative AI, offering you transformation innovation in your business. Implementing a unified data platform as a solution means that you will be able to trust your AI more, increase quality and accuracy to the responses of your prompts and save your business time and money pre- and post- generative AI processing.



## About the Author

Imran Chaudhri enjoys working with customers to understand their strategic goals and create digital solutions that can incrementally solve tough problems. He has been bringing his many years of successful AI and agile start-up-based problem-solving experiences to many large companies in the healthcare and life sciences. Imran joined Progress to focus on bringing enterprise-quality AI and NoSQL solutions for incrementally integrating data silos, making the data agile and then creating knowledge-based operational and analytics solutions on this semantic agile data.



**Get started with incorporating your business data with LLMs**






### About Progress

Progress (Nasdaq: PRGS) provides software that enables organizations to develop and deploy their mission-critical applications and experiences, as well as effectively manage their data platforms, cloud and IT infrastructure. As an experienced, trusted provider, we make the lives of technology professionals easier. Over 4 million developers and technologists at hundreds of thousands of enterprises depend on Progress. Learn more at [www.progress.com](http://www.progress.com).

2024 Progress Software Corporation and/or its subsidiaries or affiliates. All rights reserved.  
Rev 2024/01 RITM0217217

### Worldwide Headquarters

Progress Software Corporation  
15 Wayside Rd, Suite 400, Burlington, MA01803, USA  
Tel: +1-800-477-6473

-  [facebook.com/progresssw](https://facebook.com/progresssw)
-  [twitter.com/progresssw](https://twitter.com/progresssw)
-  [youtube.com/progresssw](https://youtube.com/progresssw)
-  [linkedin.com/company/progress-software](https://linkedin.com/company/progress-software)
-  [progress\\_sw\\_](https://instagram.com/progress_sw_)