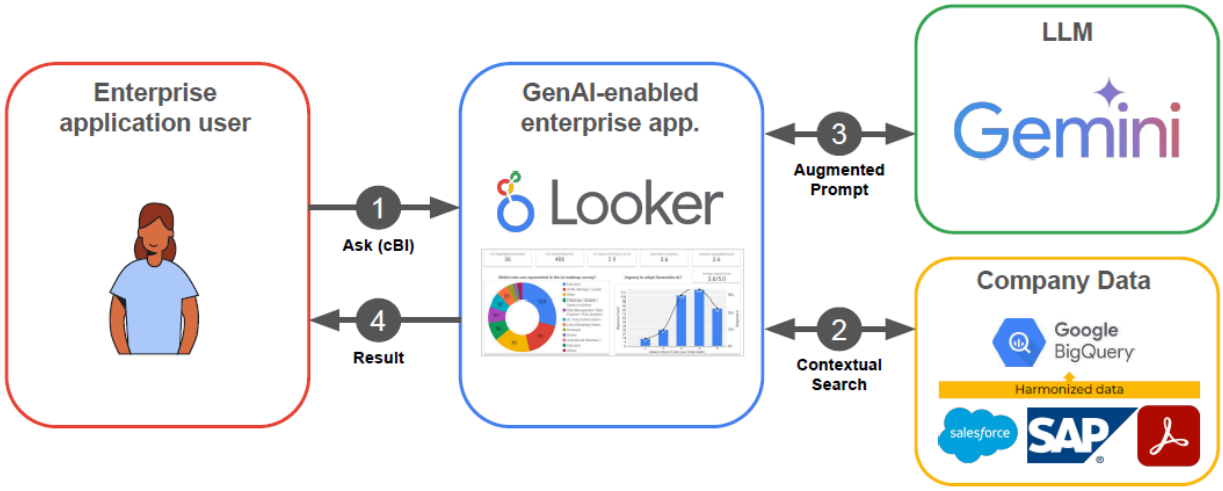
Enabling Enterprise AI with Retrieval-Augmented Generation (RAG)

By: Blake Duggan

Very grateful for the **Method360, Inc.** Data Insights and AI engineering team working in close partnership with **Google Cloud** engineers to deliver cutting-edge enterprise AI solutions. Our teams are going deep on ***Retrieval-Augmented Generation (RAG)*** to enhance the capabilities of Large Language Models (LLMs) like Gemini for enterprise use cases. What I've learned is that unlike traditional LLMs that rely solely on pre-trained knowledge, RAG-enabled AI models leverage external knowledge sources (i.e., 3rd party data sources, and proprietary enterprise systems like **SAP** or **Salesforce**) to provide more accurate, up-to-date, and contextually relevant responses. ***Simply put, good RAG modeling helps LLMs surface better answers.***



RAG-enabled AI apps. integrate company data (i.e., ERPs) with LLMs to surface contextualized insights

A RAG modeling approach seeks to curate knowledge repositories into a network of enterprise data sources, ranging from structured databases to unstructured documents and news feeds. This consolidated set of information is then transformed into numerical representations using embedded language models and stored in a **vector database,** which then facilitates efficient search and retrieval based on semantic similarity. When a business user poses a query to a RAG-powered AI application, the query is transformed into a vector and matched against the

constellation of vectors stored in the database. This retrieval process identifies the most relevant information from the AI knowledge repository, which is then combined with the original query and fed into the LLM. The LLM then leverages **both** its pre-trained knowledge and the dynamically retrieved contextual information (say the shipment status of a sales order, as of current time) to generate a comprehensive and accurate response.

Sounds cool right? (It is.)

A RAG approach to Enterprise AI architecture offers several compelling advantages:

1. It ensures generated responses are **grounded** in the most current and relevant information available within the organization's knowledge base.
2. It allows for **continuous updates** to the knowledge repository without the need for costly retraining of the LLM.
3. It enables the AI system to provide **citations** for the sources of information used in its responses, enhancing transparency and trust.

However, RAG approaches can also present certain complications:

1. Integrations can be more **expensive** than using LLMs alone due to the need for additional infrastructure and processes for data integration and system/prompt management.
2. Determining the optimal way to **model structured and unstructured data** within the knowledge repository remains an ongoing area of “art and science”.
3. As an emerging data engineering capability, investments in enterprise-scale AI requires business and technology leaders to invest in **talent, coaching and architecture modernization.**

Despite these challenges, we believe Retrieval-Augmented Generation is central to delivering enterprise-grade AI solutions by improving the quality of GenAI output by allowing LLMs to **tap into existing data assets without expensive retraining,** and to enable sophisticated multi-step prompting, nuanced insights, and recommendations tailored to individual needs.

Direct message (Blake Duggan) if you'd like to learn more about our teams approach or, how to initiate a scalable and efficient data, insights and enterprise AI program.

About us: mXa, on the 20+ year foundation of Method360, was founded to intentionally serve fast-growth companies and the unique challenges they face. We understand that inorganic and organic growth provokes change, ambiguity, and uncertainty that can deeply burden the organizations involved. By seeking to understand the human element in M&A and fast growth environments, mXa embraces a unique, contrarian approach in advising clients that seeks to realize maximum value for them in alignment with business objectives.

Interested in learning more about our capabilities or discussing your story? We're here to help.