Luxoft
A DXC Technology Company

#AIJournal

Document Retrieval Augmented Generation using generative AI:

A comprehensive overview

by **Thomas Cotter**, Data Scientist, Data Analytics Delivery
and **Ciarán Murphy**, Principal Consultant, Data Science and Analytics

Drawing from years of pioneering AI solutions, Luxoft stands at the forefront of transforming industries with cutting-edge innovations. Our expansive AI expertise has consistently delivered unparalleled results for clients worldwide.

In the age of vast amounts of data and increasing demand for immediate, accurate insights, generative AI offers innovative solutions for data interrogation. The term generative AI, or gen AI, has garnered significant attention in various sectors, especially among businesses grappling with extensive document-based workflows. Retrieval Augmented Generation is a promising approach within this domain, promising a transformation in how we retrieve, analyze and utilize extensive amounts of stored data.

What is Retrieval Augmented Generation?

Retrieval Augmented Generation (RAG) is a paradigm that combines two powerful components: information retrieval and generative model capabilities. Imagine a scenario where you have a plethora of documents — too many to sift through individually. Using RAG in business, you can set up a Q&A interface that allows you to query these documents, with the AI generating precise answers in response.

Real-world application: Insurance conversion with Luxoft

One of the ground-breaking applications of RAG generative AI has been in the insurance sector. Luxoft, a DXC Technology company, embarked on an extensive insurance conversion project. The client's challenge was to transfer old insurance policies from a dated system to a modern one. The enormous volume of documents dating back to 2005 posed a significant issue in terms of time and resources. By implementing RAG, Luxoft saw an opportunity to expedite the process.

Documents central to the project, such as product details, are loaded and transformed into numerical representations using an encoder. For instance, specific words are encoded as unique mathematical representations, creating a reference system that is machine-readable and efficient.

How does it work?

1. Encoding and vectorization:

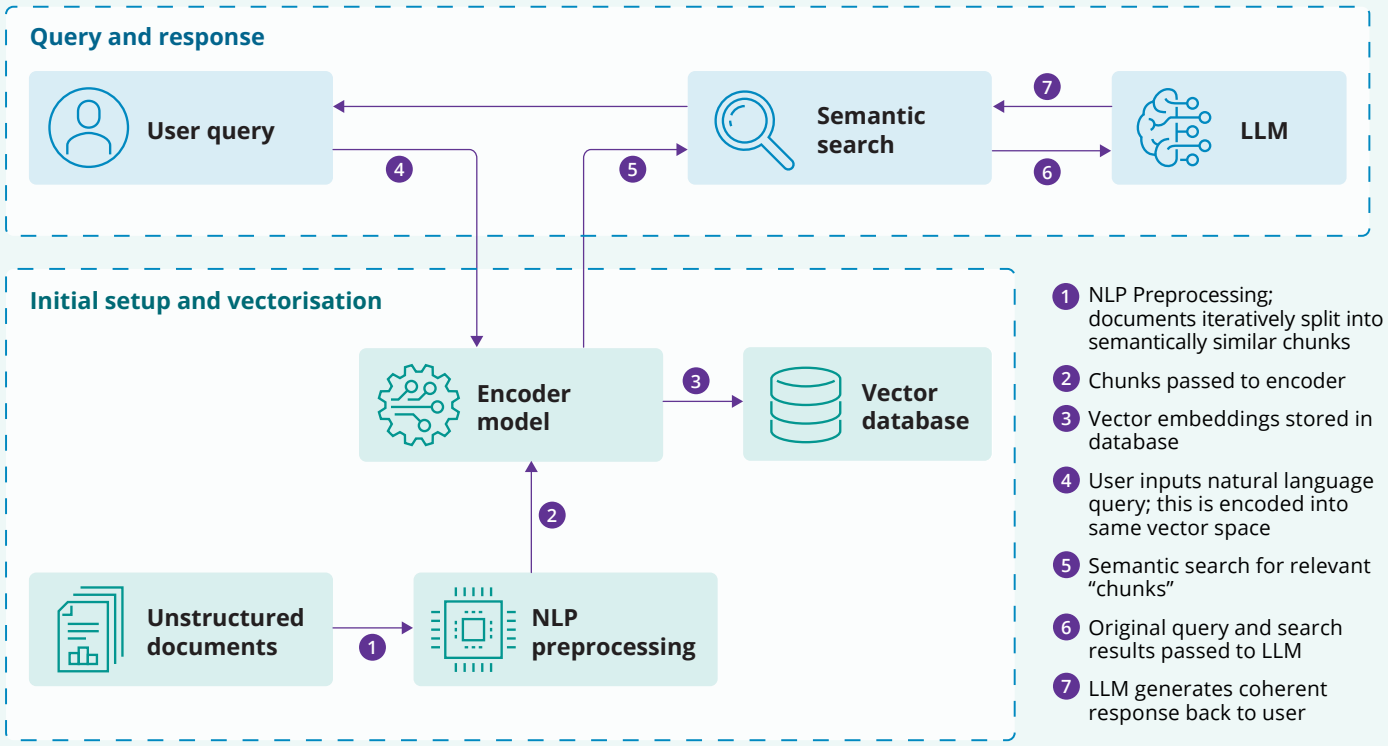
At the core of RAG is the process of encoding. Documents are transformed into a numerical format, where even complex constructs like words or phrases are represented as vectors. This ensures that the content is not just stored, but is context-aware, paving the way for more intelligent data retrieval.

2. Querying the encoded data:

Business analysts can then interact with the encoded documents. They select specific documents or groups of documents to query and submit their questions. These questions undergo the same encoding process, ensuring they exist within the same vector space as the stored documents.

3. LLM interaction:

Once the relevant sections of documents are retrieved based on the query, they are sent to a Large Language Model (LLM). The LLM then processes this data to generate direct, concise answers to the question posed.



Key **benefits**

Speed:
What would have taken hundreds of staff several hours can be reduced to taking a handful of staff merely seconds to complete. When searching the vector database, it takes just 38.3ms to return the closest matching section of the document for a single question when searching over 1,000,000 embeddings.

User-friendly interaction:
Gone are the days of navigating complex file systems or SharePoint databases. With RAG, users can simply type in a question and get the answer, streamlining the user experience and boosting efficiency.

Flexible querying:
This technique allows for more complex types of queries, which were not previously possible. For example: “Show me all policies in the last 10 years with customers located in Birmingham, Liverpool and London with two or more cars,” and “Compare the reinstatement interest across these products if the policyholder is 85 years old.” It gives users the ability to mix and match various search types, combining complex search criteria into one query rather than having to make several different searches.

Modularity:
One of the most potent advantages of this solution is its modularity. Depending on data privacy needs, businesses can either use a hosted LLM or opt for an on-premises model. This flexibility ensures that the solution is adaptable to varying requirements.

Cloud vs on-premises

Cloud deployment benefits:

- **Cost-efficiency:** Initial costs can be relatively low due to on-demand pricing models
- **Ease of integration:** Hosted models primarily use REST APIs, ensuring seamless integration with existing systems
- **Quick time-to-value:** Rapid deployment means quicker results
- **Upgrades as standard:** At the rate AI is developing, with on-premises, you might be looking at a time-consuming upgrade every three months; something you don't have to worry about with cloud

On-premises deployment benefits:

- **Full data privacy:** Ensuring sensitive data never leaves your network
- **Customization:** Allows for specialized training on domain-specific data
- **Resource utilization:** While upfront costs can be higher, once in place, maintaining and updating the model can be more cost-effective in the long run

Data augmentation over automation

While AI can provide incredible insights, the importance of human judgment cannot be sidelined, especially in critical sectors like insurance and banking. RAG provides an augmentation of human capabilities, not a replacement. It will always be possible to review the original document. The aim is to assist users by providing time- and effort-saving efficiencies, not to replace humans in the process altogether.

Case study: RAG in action

Luxoft undertook a significant project for RAG in insurance, involving old policy and product documents. The conventional method of analyzing these documents was time-consuming, impacting the productivity of business analysts. With the introduction of RAG, Luxoft enabled a Q&A-style interaction with these documents. This transformed a cumbersome process into an efficient and accurate one, promising scalability for similar projects in other sectors, from banking to insurance industry solutions.

Looking to the future

Retrieval Augmented Generation is not just a buzzword; it's an actionable, efficient approach to data retrieval and analysis. As the boundaries of what's possible with AI continue to expand, approaches like RAG will redefine industries, enhance productivity and deliver insights in ways previously unimagined. This project exemplifies the transformative power of generative AI in real-world scenarios. It underscores the importance of leveraging technological advancements for strategic business outcomes. The future of data interrogation is here, and it's intelligent, efficient, and generative.

About **the authors**

Thomas Cotter
Data Scientist, Data Analytics Delivery, Luxoft

Tom is a recent graduate of the University of Nottingham with a master's degree in computer science and AI. He has a strong foundation in data science and deep learning and is passionate about leveraging data to drive insights. With hands-on experience as a data scientist, Tom is already leading innovative projects that create business value as well as being committed to continuous learning and improvement within the field.

Ciarán Murphy
Principal Consultant, Data Science and Analytics

Amassing 10 years of experience in data analytics and machine learning has taken Ciarán across multiple industries, from financial services to e-commerce and utilities. The hard-earned result is a proven track record in implementing end-to-end data science use cases, communicating results and insights, and solutioning. Ciarán has a BSc in Mathematical Physics from University College, Dublin, and an MSc in Theoretical Physics from King's College, London.

Want to find out how RAG can take your business to the next level?
Visit our website and **get in touch** with our Luxoft experts today.

About Luxoft

Luxoft, a DXC Technology Company delivers digital advantage for software-defined organizations, leveraging domain knowledge and software engineering capabilities. We use our industry-specific expertise and extensive partnership network to engineer innovative products and services that generate value and shape the future of industries.

For more information, please visit **luxoft.com**