# **KPMG Cognitive Search:** Important Product Considerations & How I Can Help

Michelle Schaffer
February 2024
LinkedIn | michelle.schaffer@gmail.com

# Table of Contents

# Need for Cognitive Search Solutions at KPMG

# Overall Goal of KPMG Cognitive Search Solutions

## AIM

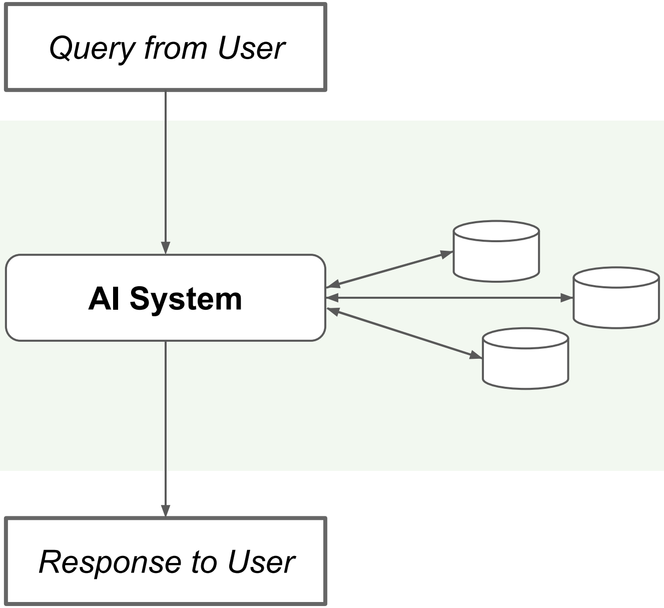
Deliver tangible business value to KPMG by improving audit efficiency, data analysis, and risk assessment [1]

## APPROACH

Create innovative generative AI solutions to **enhance the retrieval, augmentation, and generation of knowledge** across the firm's audit repositories [1]

How can we design our AI solutions to ensure they can:

- Understand the intention of user queries
- Gather appropriate info from various KPMG repositories to provide relevant context for those queries
- Use AI to generate knowledge that not only answers the queries, but ideally goes beyond to provide additional value



# Addressing User Needs of Audit Professionals

## BASIC NEEDS: INFO RETRIEVAL

In response to their queries, users want to get info returned that:

- is accurate
- is relevant
- includes reference links to provide more details

## ADVANCED NEEDS: GENERATIVE AI

Many users are familiar with Chat-GPT and would love to apply generative AI’s powerful capabilities to their day jobs. Beyond just getting info back from queries, we’ll find ways to help KPMG users *create even more value* from it:

### Audit Efficiency

e.g. automate manual, time-consuming tasks such as writing reports

### Data Analysis

e.g. rewrite data analysis code found in a shared repo to process a new dataset

### Risk Assessment

e.g. summarize documents to help audit team pinpoint specific ones to examine more closely

# Important Cognitive Search Product Considerations

# Generative Model Input Options

## MODEL INPUT REQUIRED

To help audit professionals within KPMG, generative AI models would need data input that is:

- Recent - *some models were trained years ago and lack up-to-date info*
- Company-specific - *models trained externally wouldn’t know non-public info about KPMG*
- Relevant - *need to anticipate range of potential user questions*
- High-quality - *need to ensure data is accurate*

When a generative AI model doesn’t have data needed to answer a query, it may provide inaccurate info (“hallucinate”).

## OPTIONS FOR PROVIDING INPUT

Frequently retraining generative AI models on the latest, company-specific data can be expensive and inefficient. [2,3] Instead, retrieval augmented generation (RAG) provides relevant info to models as part of the prompt.

Some benefits of using a RAG approach include: [2,3]

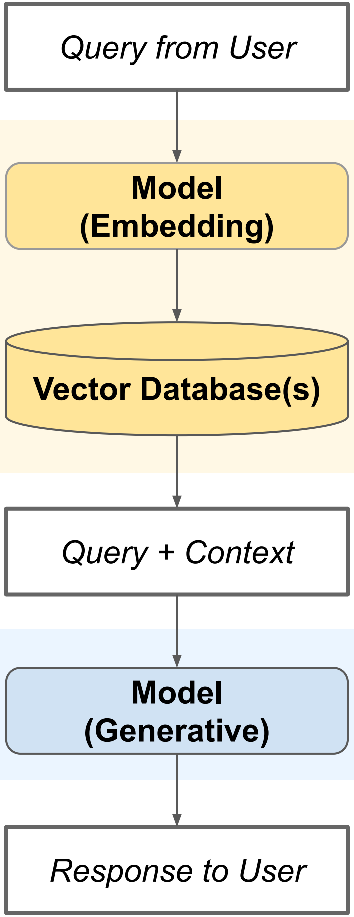
- **Increased accuracy and reduced hallucinations**: models are less likely to make up incorrect answers
- **Improved auditability**: RAG allows models to cite its sources
- **Lower costs**

# How Retrieval Augmented Generation (RAG) Works [2,3]

## STEP 1) SYSTEM GATHERS CONTEXT RELEVANT TO QUERY

- Analyzes user query to determine what user is asking for
- Retrieves additional context, e.g. reference articles, to augment query
- Passes original user query plus extra context into model prompt

## STEP 2) GENERATIVE AI CREATES RESPONSE



# RAG: Using Semantic Search to Find Additional Relevant Context [2,3]

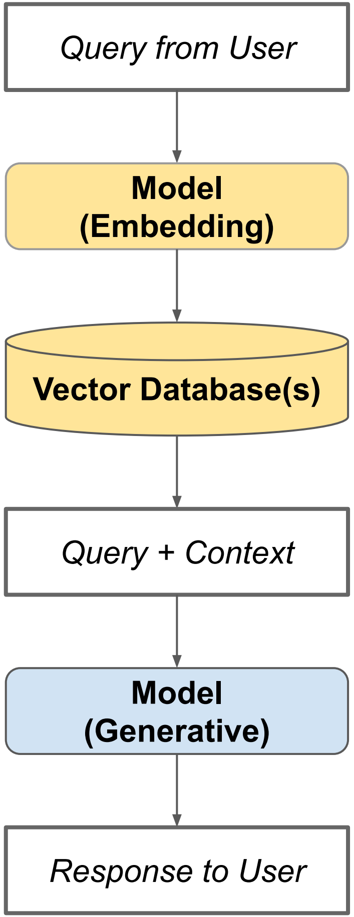
Traditional search relies on matching key words or phrases.  Semantic search is an improved approach because it relies on meaning instead.  It tries to understand the intention of a user query and find info saved in knowledge repositories that's relevant to it.

Embedding models are an important component of semantic search because they capture the meaning of the content.  The models translate content into numerical representations (or "vectors") that can be processed by AI systems.

Processing by an embedding model is needed both:

- ○ during original set-up when saving all content into vector database(s)
- ○ later at run-time when analyzing the incoming user query

During retrieval augmented generation, the system compares the vector representation of the incoming user query with the vectors already saved in the database to find the content that's the most similar.



# RAG: Embedding Model Considerations

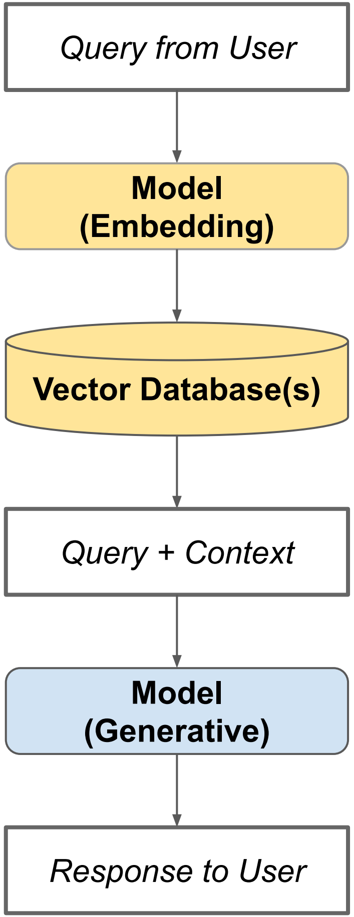
## SELECTING AN EMBEDDING MODEL

Several factors play into choosing a model including finding one that:

- aligns with our use case, e.g. familiar with audit terminology [4]
- is effective (need to determine benchmarks to look at, e.g. MTEB) [5]
- fits within our budget

## BREAKING CONTENT INTO SEGMENTS

When translating a long document into vectors, it's necessary to divide the full text into segments (or "chunks"). It's essential to optimize the "chunking" process, e.g. not break up text in the middle of a sentence. [4]



# RAG: Data Considerations

## DATA STORAGE

During RAG, the system identifies and ranks info that is most relevant to the user query. Though all content is ideally saved within vector databases, it's also possible that it could be kept in a variety of other types of data stores distributed throughout the company. [6]

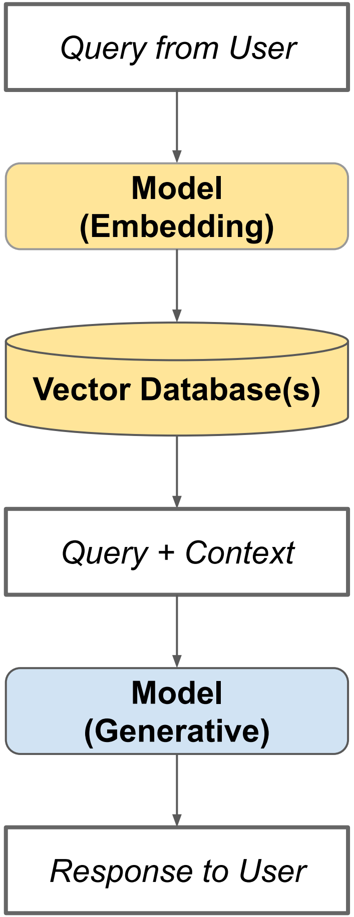
It is important to evaluate a range of criteria when deciding whether to use a vector database, e.g.: [7]

- performance
- cost
- frequency of data updates

## OTHER DATA FACTORS

Additional considerations related to data use include:

- data privacy and security
- data quality
- data pre-processing (e.g. audio transcription)



# Generative Model Optimization

## SELECTING A GENERATIVE MODEL

The model used will make a significant impact on the success of generative AI systems. Selection criteria should include: [8]

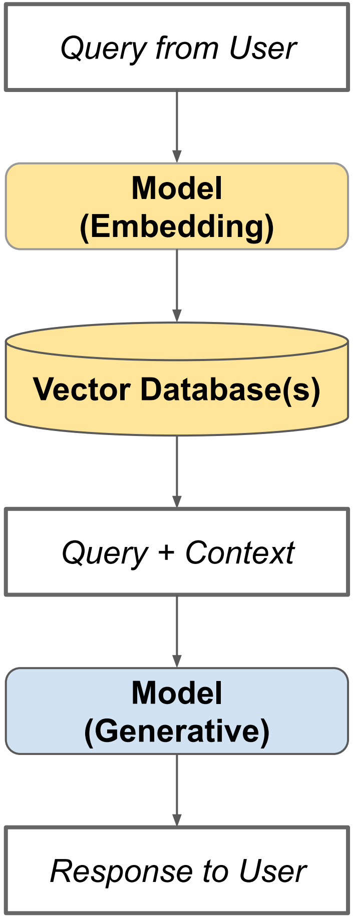
- licensing and source (e.g. open source vs closed)
- complexity (e.g. size and architecture)
- performance
- cost

Regulations and responsible AI best practices must factor into the model choice as well:

- legal restrictions (e.g. Chat-GPT was banned for some time in Italy [9] )
- privacy concerns (e.g. whether customer data will be used to train foundation models)

## PROMPT ENGINEERING

In addition to providing the user query and relevant context, other input can be given to the generative model to ensure the best results.



# Development Considerations

## ROI

It’s critical to gauge return on investment:

### Costs

- “**All-in expenses**”: How much will solutions cost including development, ongoing support, and upgrades? Beyond tech resources, what non-technical expertise will also be needed, e.g. UX design, etc?

**vs**

### Benefits

- **Reach**: how many users will be able to leverage the new capabilities?
- **Impact**: to what extent will new capabilities enhance their productivity?

## MEASURING SUCCESS

What evaluation criteria can we use (e.g. information retrieval golden source comparison [4])?

## PHASING

Are there quick wins to develop now?  Long-term strategic investments to work towards?

# How I Can Help

# Highlights of Some of My Relevant Tech Product Management Experience

## SUCCESS IN CREATING LARGE-SCALE ENTERPRISE PLATFORMS TO ENABLE AI

Working at Chase and Capital One for more than four years gave me valuable experience in constructing large-scale, cloud-based platforms for mission-critical AI operations e.g.:

- Our feature publishing tool was used by multiple groups across Capital One, including fraud and customer-facing technology teams that generated as much as **$1.2 billion in total business value per year**.
- Our retrieval systems were developed to support important Capital One business use cases, e.g. providing years of client details to ML models so forecasters could **predict millions of dollars in potential credit card losses**.

## BACKGROUND IN BUILDING GENERATIVE AI SOLUTIONS

After leaving Chase, I helped a start-up for five months as an AI product manager. I joined forces with their data scientists on several AI capabilities including a large language model-based support chatbot:

- System was designed to use **retrieval augmented generation**. Prototype tests showed encouraging results.
- We explored options for bolstering quality controls, cutting AI costs, and implementing future upgrades.

## SKILL IN USER RESEARCH

At Chase, I led a firm-wide product strategy initiative where I interviewed many AI practitioners across all divisions. After identifying gaps in our enterprise offering, I worked with stakeholders to use these requirements to **guide millions of dollars in product investments**, including a new innovative tech project covered by a non-disclosure agreement.

# Research Sources

[1] KPMG Cognitive Search Product Manager job description

[2] Kim Martineau. What is retrieval-augmented generation? IBM. Aug 22, 2023

[3] Retrieval Augmented Generation (RAG). Pinecone. Aug 3, 2023

[4] Ben Lorica. Navigating the Nuances of Retrieval Augmented Generation. The Data Exchange. Oct 26, 2023

[5] Choosing an Embedding Model. Pinecone.

[6] Vector databases. Cloudflare Docs.

[7] An (Opinionated) Checklist to Choose a Vector Database. Pinecone. Sep 13, 2023

[8] Which LLM to choose for your use case? UbiOps. Feb 1, 2024

[9] Shiona McCallum. ChatGPT banned in Italy over privacy concerns.  BBC. Apr 1, 2023