FACTSET

factset.com

AI IN ACTION

How Retrieval Augmented Generation Enhances GenAI Accuracy for Financial Research Data

TABLE OF CONTENTS

How LLMs Do—and Do Not—Work	4
Hallucinations	6
Retrieval Augmented Generation	8
The Relevance of Data Formats	11
Semantically Rich Metadata	15
Infusing Domain Knowledge	16

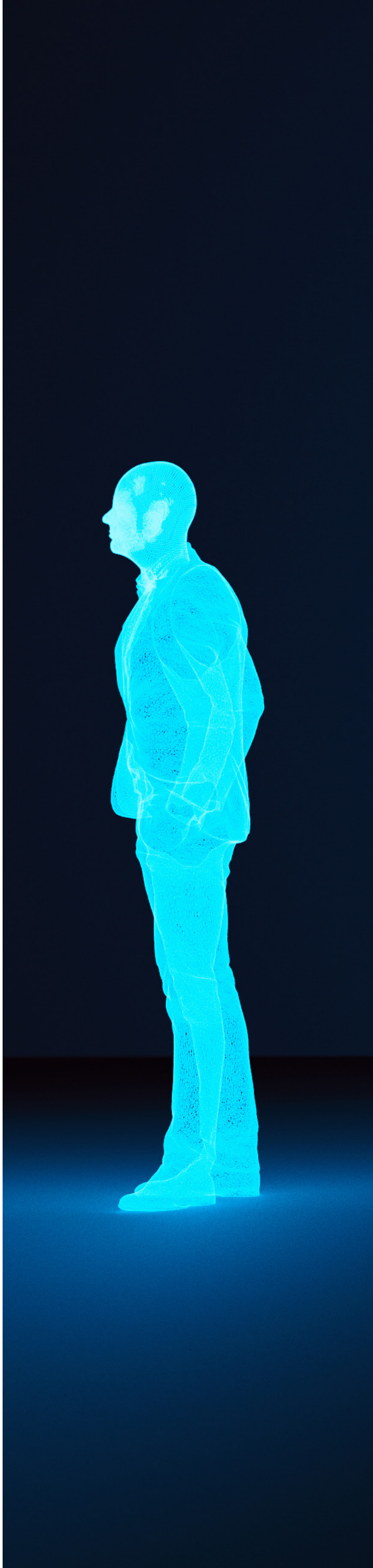
INTRODUCTION

When properly engineered with high-quality data, the Large Language Models that drive generative AI can help firms across the financial sector materially strengthen workflows, collaboration, and compliance. This applies to the buy-side and sell-side as well as wealth managers, private equity firms, and corporations.

The purpose of this e-book is to summarize how LLMs work and highlight some of the upside potential and downside limitations. Because data accuracy is table stakes for financial professionals, we will discuss the problem of AI hallucinations and emphasize an effective programmatic solution to stem them: retrieval augmented generation (RAG).

Based on knowledge gleaned from years of using AI and LLMs across FactSet's products and services, our overall intention is to offer perspective that helps you realistically generate the most value from your investment in AI.

Large Language Models that drive generative AI can help firms across the financial sector materially **strengthen workflows, collaboration, and compliance.**



01 How LLMs Do—and Do Not—Work

Understanding the operational basics of LLMs will help inform your firm's strategic planning and tactical execution.

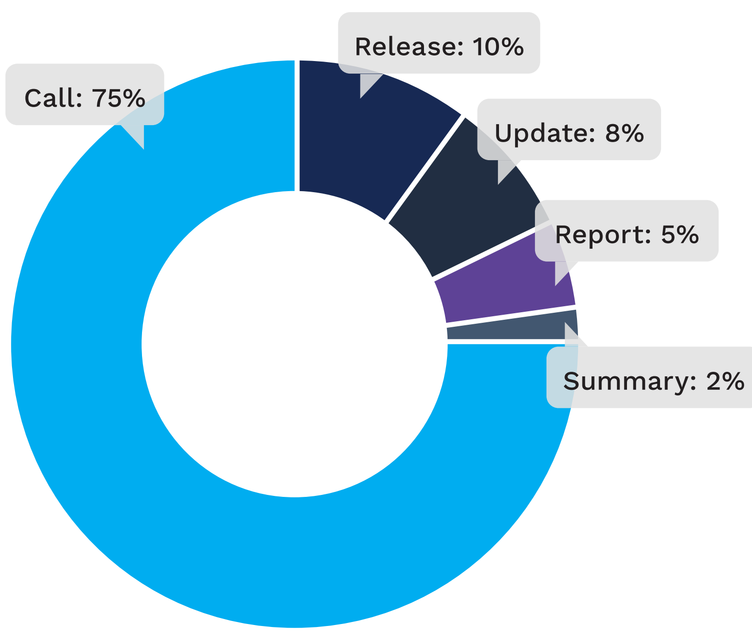
At a basic level, when a user types a prompt (i.e., a question they need answered or a direction to complete an action), an LLM will iteratively select the most probable next token (a unit of text, such as a word or part of a word) from a distribution for its response. The LLM considers both the original prompt and each previously generated token to maintain coherent context.

For a simple example, given the sentence "The company's performance exceeded expectations as per their recent earnings ________," the LLM will generate the list of words in Figure 1 to finish the sentence and select the word most statistically likely to match the context of the request.

It does this behind the scenes; the processing is not visible to the user. Based on its training, the LLM determines that "call" is the most likely next word.

Because LLMs are intentionally built with some allowance for variability, inconsistent responses to the exact same prompt can happen. For example, in this scenario the LLM might return "statement" instead of "call."

FIGURE 1:

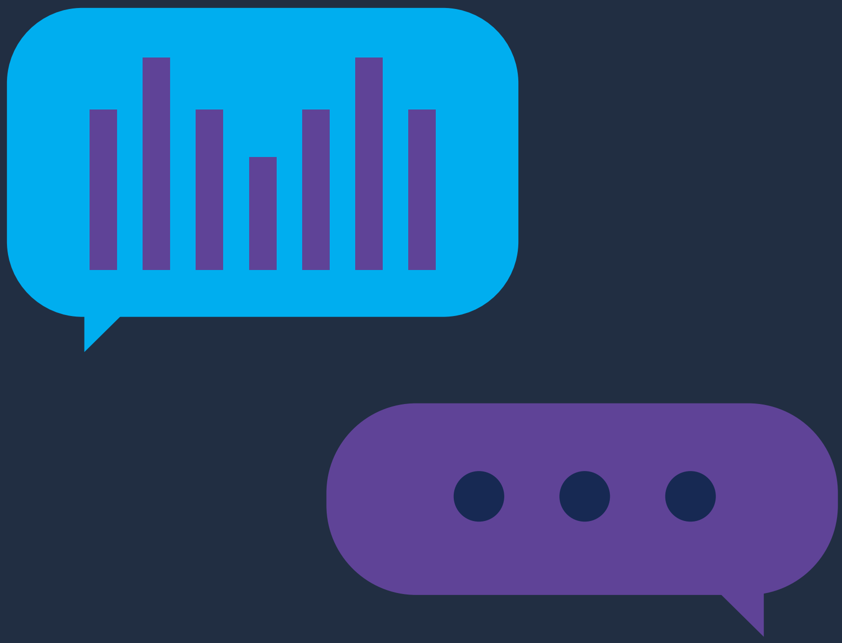


Programmers can control how conservative or creative an LLM's responses are by setting a temperature parameter. Temperatures range from 0 to 1.

A lower temperature value such as 0.2 would cause the LLM to stick to the most likely or predictable outcomes. For example, if asked to complete the phrase "Top line improvement is due to _________," it might almost always say "sales" or "revenue."

A model with a higher temperature value such as 0.7 would produce more varied and creative responses, balancing predictability and innovation. For the same prompt, it might generate responses such as "expansion," "demand," or "innovation."

Of course, users of the popular LLM ChatGPT know that responses to prompts can be much more elaborate than the above examples. But the fundamental idea is that an LLM is not looking up data in a database, it is not searching the web (unless the model was enhanced with that capability), and it is not cognitively understanding the text like a human does. Rather, it is using statistical correlations or patterns it has learned from the large datasets of text on which it was trained.



02 Hallucinations

LLMs have a remarkable ability to generate useful, relevant content and to iterate responses as users fine-tune their prompts. But there are limitations.

Given accuracy is vital in the fact-based, data-centric decision-making processes of the financial environment, hallucinations are perhaps the riskiest limitation of generative AI to resolve.

Hallucinations are basically errors in fact and logic; instances when a model generates coherent, plausible-sounding text that is actually inaccurate, misleading, or completely fabricated.

The phenomenon happens because, as described earlier, LLM models predict words based on patterns learned from training data. Models do not possess an understanding or a knowledge base that ensures factual accuracy based on research. And today's models do not inherently validate their outputs, although engineers can fine-tune prompts and implement a procedure to flag quality issues in more advanced settings.

LLM makers have been iteratively improving models to mitigate hallucinations. Bolting on fact checkers, grounding answers with fetched data (with methods like web browser plug-ins), and improving the training data and reinforcement learning process are examples of useful enhancements.

Given accuracy is vital in the fact-based, data-centric decision-making processes of the financial environment, hallucinations are perhaps the riskiest limitation of generative AI to resolve.

Because the technology is evolving rapidly, software engineers can leverage research on various models, including LLMs from Meta, Anthropic, Google, OpenAI, and open-source models. The research assesses models' factuality as well as criteria such as logical abilities and avoidance of harm.

If your organization allows the use of publicly trained models such as ChatGPT, there are strategies to help minimize hallucinations. For example, you could use the most current model and write precise prompts that provide the full context of the requests. Writing effective prompts is a useful skillset.

For environments where hallucinated content would be a material risk—professions such as law, medicine, finance, or coding, for example—it's essential to carefully compare outputs with a trusted, validated source. Consider developing guidelines to assess the level of content risk, such as in Figure 2.

FIGURE 2: A RISK-BASED APPROACH TO VALIDATING GENERATIVE AI OUTPUTS

03 Retrieval Augmented Generation

Perhaps the most important strategy to vastly reduce hallucinations is retrieval augmented generation (RAG), available to engineers building software products on top of an LLM. RAG is the programmatic version of providing context in a prompt and is akin to grounding answers in facts.

Augmenting responses from an LLM with RAG provides the following key benefits:

- No need to re-train or fine-tune the LLM
- Better accuracy and fewer hallucinations since answers are derived from proprietary data
- Improved auditability with the source of an answer
- Enablement of up-to-date knowledge and user-based security

With RAG, the LLM first searches for an answer to a user's question in a trusted database. For example, a customer service chatbot would look through existing help documentation. It would combine the best matches from its database with the text of the user's question and then use generative AI to present the response in a conversational format.

To highlight the difference in outputs with and without RAG, below is an illustrative example:

Without RAG

What is FactSet's dividend yield as of February 7, 2025?

As of my last update, FactSet Research Systems Inc. has periodically paid dividends, but the exact dividend yield can fluctuate based on the stock price and dividend announcements. Historically, FactSet's dividend yield has been relatively modest, reflecting its status as a company that often reinvests earnings to fuel growth. For precise historical yields, you would typically refer to annual reports or financial databases.

With RAG

What is FactSet's dividend yield as of February 7, 2025?

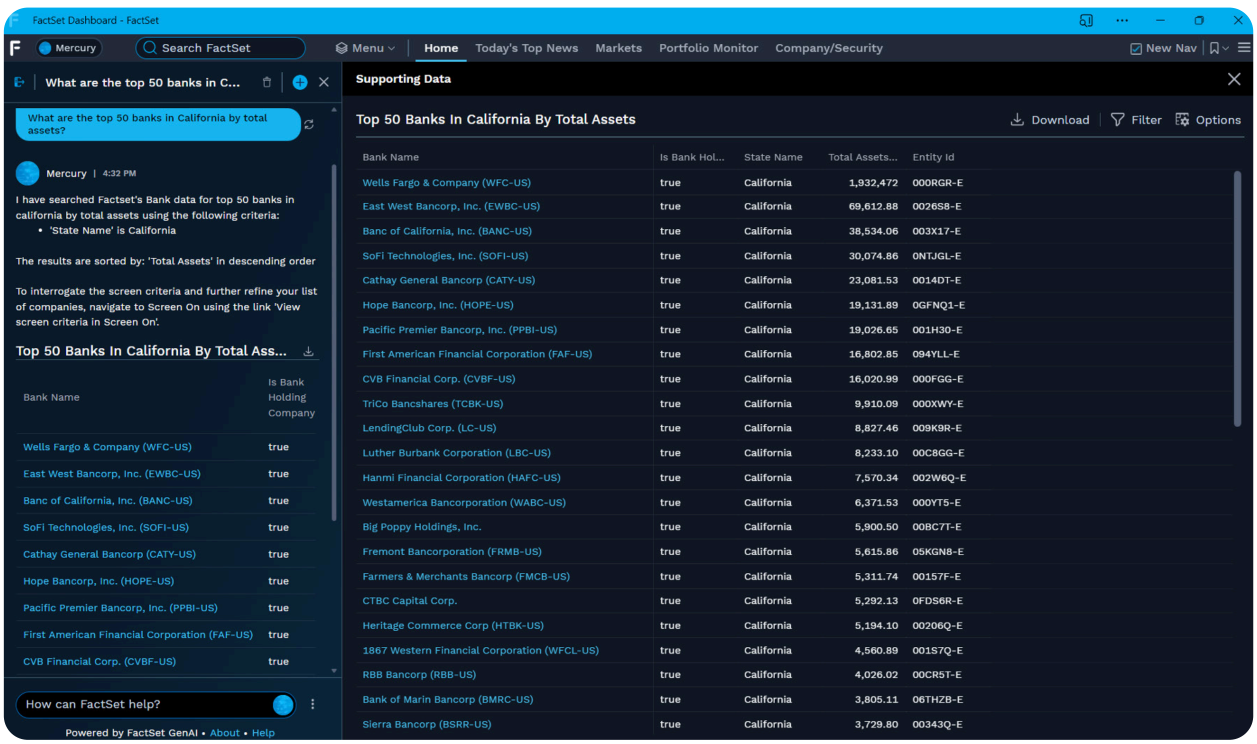
As of February 7, 2025, the dividend yield for FactSet Research Systems is 0.90%.

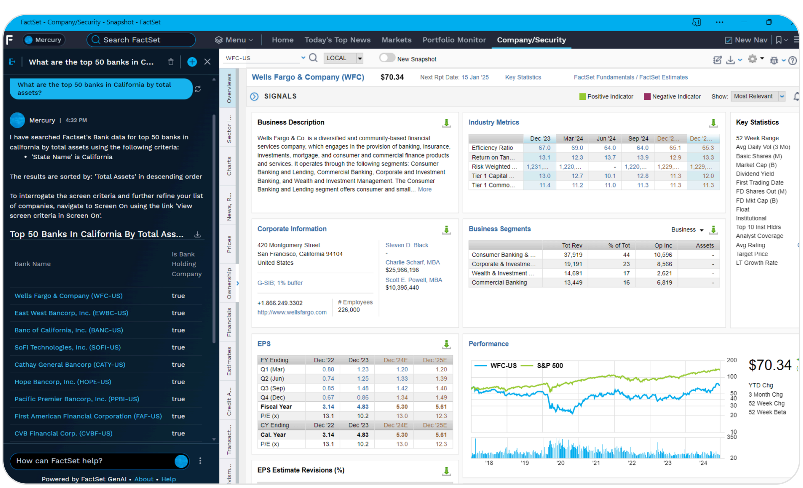
Another example: Asking a LLM to identify the key risks of an investment in a specific company stock. A LLM without RAG and without current training on that company would likely reply that no specific context about the investment risks of the company has been provided. The best it could do, with additional prompting, is summarize the typical risks to consider when investing in equities.

A LLM using RAG within a governed, regularly updated data source will summarize the specific risks to an investment in that company stock and link to each risk's specific source, such as a 10-Q. It could also provide a view of that company's financial highlights and keys stats on trading, valuation, and estimates, for example.

LLM-enhanced products throughout FactSet are using RAG to ground answers and link users to the source. Responses to fact-based requests are retrieved from our governed databases, not generated from the Large Language Model's training data. (FactSet has its own secure and private instances of Large Language Models and does not share client, confidential, or proprietary data with public generative AI models.)

The FactSet Mercury-powered beta AI chat for junior bankers uses RAG within our governed datasets to efficiently deliver accurate, source-linked results. For example, users who prompt "What are the top 50 banks in California by total assets?" would see this format:





FactSet Mercury returns the answer in chat and, as shown above, links each bank to additional context or information sources—in this case, an overview of the top bank on the list. From there, users can click for a full report of the bank (shown to the left), including its entity structure, comps analysis, supply chain, capital structure, and more. Those are all examples of how RAG has enabled accuracy, auditability, and up-to-date knowledge from governed datasets.

04 The Relevance of Data Formats for RAG

To further understand how RAG is used to derive accurate answers from data, it's important to highlight that financial research data is primarily categorized into two types: unstructured and structured.

Both formats are relevant in retrieval augmented generation for several main reasons:

1. **Efficiency**
   Utilizing structured data allows for quicker and more accurate retrieval of relevant information, which can improve the overall efficiency of the generation process.

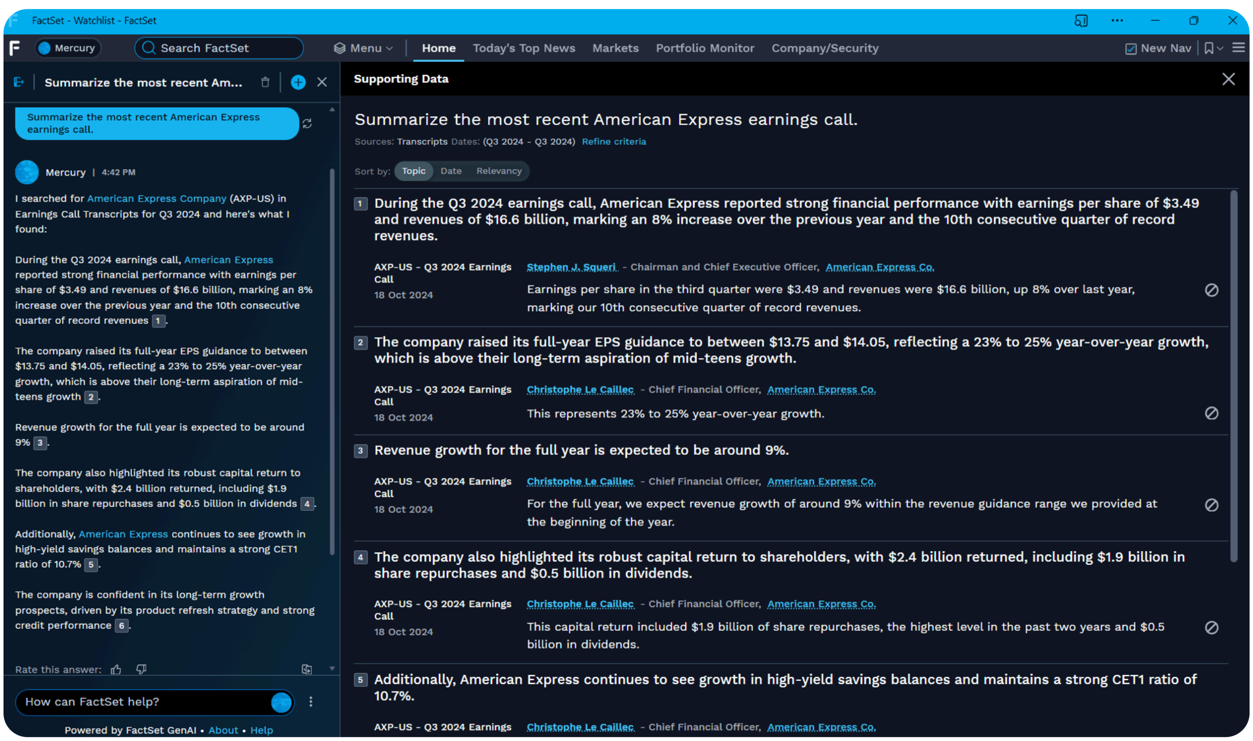
2. **Accuracy**
   Structured data often contains less noise and more reliable information, which can enhance the accuracy and reliability of the generated content.

3. **Depth of Insights**
   Incorporating unstructured data can provide richer, more nuanced insights and context, which can lead to better, more holistic responses to complex queries.

4. **Versatility**
   Combining structured and unstructured data can enable a robust system that leverages the strengths of both types of data. That system can be used for a broader range of applications.

05 Unstructured Data RAG

Unstructured data commonly found in financial settings includes emails, documents, news articles, and transcripts. These text-data sources don't follow a specific format or structure but are rich in information, primarily (but not exclusively) for qualitative insights. For example, FactSet Mercury generated the following summary from a quarterly earnings transcript (unstructured data):



Using the RAG model for unstructured data such as text involves combing through a vast collection of pre-indexed documents from trusted knowledge sources such as proprietary and third-party data. (Pre-indexing maps content for fast retrieval within the LLM.) The objective is to identify the most relevant documents for users' prompts and generate responses to users' questions in the context of those documents.

In the context of an earnings transcript (unstructured data), RAG would operate as follows:

- **Data collection:** As an earnings transcript is mainly composed of text, data collection involves capturing the spoken or written content from company earnings calls or reports.
- **Data cleaning:** This step removes noise and irrelevant information from the data. It may include eliminating duplicates, correcting errors, and filtering out unnecessary text such as advertisements.
- **Data chunking:** When working with larger datasets, it's beneficial to divide data into smaller, more manageable segments before providing it to the LLM. This approach both simplifies data handling and ensures only relevant information is retrieved during queries.
- **Indexing and retrieval:** The RAG system indexes the unstructured data to enable efficient retrieval. The retrieval phase comes into play when a user queries the system with a question such as "What were the main revenue drivers mentioned in the last earnings call?". The system uses algorithms to search the unstructured transcripts for relevant segments that mention revenue drivers.
- **Relevance scoring:** The retrieved segments are then scored based on their relevance to the query. The scoring could involve factors such as keyword matching, semantic similarity, or context understanding.
- **Generation:** The system synthesizes the retrieved pieces of information to produce a coherent response. For example, it might generate output such as "The main revenue drivers mentioned in the last earnings call included increased sales in the North American market, the launch of new product lines, and higher online sales."

06 Structured Data RAG

Structured data, in contrast, is often used in quantitative analyses and is therefore organized in a defined manner, often in rows and columns. Database tables are a classic example of structured data, such as the column of asset amounts for the top 50 banks in California discussed earlier.

However, the application of RAG using structured data poses unique challenges. Unlike text data, structured data cannot be pre-indexed. This necessitates a different approach for retrieval and integration within the LLM.

As an example, let's consider again a junior banker using FactSet Mercury to prompt the top 50 banks in California by total assets. While the request appears straightforward, the chatbot must take the following intricate steps to respond.

- **Understand the question:** Given the chatbot is integrated with both structured and unstructured data, it needs to understand whether this question requires structured data or unstructured data or a combination of both. Since the data required to answer the example question is in tables, we classify it as a structured data question.
- **Identify required data elements:** The chatbot must discern the different elements of data that answer the question. In our example, that includes banks, asset sizes, and locations in California.
- **Determine data source:** The chatbot must know where to find this information. Depending on the setup, this could be a table (or tables) in a database, multiple tables in multiple databases, or a data-provisioning service or API.
- **Retrieve data:** After locating the data source, the chatbot needs to retrieve the relevant data. This could involve executing a database query either directly or through a data-provisioning layer.
- **Perform necessary operations:** To present the top 50 banks in California, the chatbot must sort them based on their assets and select the top 50. This involves retrieving data and applying the correct sorting and filtering logic.
- **Generate a user-friendly response:** Finally, the chatbot must present the information in a clear, concise manner. Depending on the user's prompt and the chatbot's capabilities, the format could be a simple text response, a table, or even a visual representation such as a chart.

The California bank example underscores the complexity and sophistication required when dealing with structured data. It's not just about understanding language. It's also about effectively interacting with and processing data from various sources to provide accurate, relevant responses.

07 The Role of Semantically Rich Metadata in RAG Solutions

Most enterprise databases weren't designed with LLM capabilities and requirements in mind. In the AI environment, that poses a significant challenge because of the guidance LLMs require to navigate and interpret the vast amounts of data and metadata in databases. (Metadata provides information about other data such as titles, authors, dates, and keywords.) However, metadata—which is crucial for understanding the content and context of the stored information—may be incomplete, missing, or incompatibly formatted.

To bridge the gap, it's essential to provide LLMs with semantically rich metadata, which does more than just provide the basic information of metadata. It enables LLMs to effectively map user questions to the correct data sources and specific fields with the granular data the user seeks. Examples of semantically rich metadata include keywords such as a financial report, an abstract (short summary) of the report, and a geolocation such as New York, NY, USA.

For instance, in response to the prompt for the top 50 banks by assets in California, the LLM must correlate the query with data fields such as bank name, bank ID, state name, and asset value.

By enhancing metadata that way, the LLM in that scenario can more accurately identify, retrieve, and interpret the specific pieces of information to answer the user's query. This process involves both recognizing key terms in a question and understanding their relevance and relationship to the data fields within the database.

Thus, the effectiveness of an LLM in an enterprise setting hinges significantly on the quality and compatibility of the metadata provided alongside the data model.

By enhancing metadata, the LLM can more accurately identify, retrieve, and interpret the specific pieces of information to answer the user's query.

08 Infusing Domain Knowledge

It's beneficial to integrate a knowledge base—a centralized repository for information, data, and rules that an artificial intelligence system uses to make decisions, generate responses, and solve problems.

Going beyond fact-based prompts to list the top 50 banks by assets, LLMs also need to accommodate vague or open-ended questions that users might ask. For example, "Are large banks more profitable than small banks?".

Such questions don't always seek specific factual data but rather a form of analysis or insight. Hence, the chatbot must navigate through a set of assumptions to provide a meaningful answer.

To effectively respond to those types of inquiries, it's beneficial to integrate a knowledge base—a centralized repository for information, data, and rules that an artificial intelligence system uses to make decisions, generate responses, and solve problems.

Knowledge bases typically contain facts about the world, relationships among data, and heuristics. Heuristics are rules of thumb or simple strategies for making decisions and solving problems quickly, without the need for exhaustive calculations or comprehensive data analysis.

The knowledge base infuses the domain-specific knowledge, helping the chatbot understand and define key concepts such as what constitutes a large or small bank and the metrics that determine profitability.

Using this external knowledge base, the chatbot can correctly interpret the data and enhance the relevant pieces of data or information to construct an answer. This approach handles a broader range of queries that require subjective analysis or drawing conclusions from data.

CONCLUSION

Artificial intelligence offers firms across the financial sector a real opportunity to modernize their target operating models and unlock new levels of growth. Specifically, a Large Language Model using the RAG approach is a powerful tool to improve accuracy and auditability and help stem hallucinations. Altogether, the technology enables financial professionals to access and analyze high-quality data, streamline workflows, and deliver differentiated insights to colleagues and clients.

Technology and business leaders who know how to leverage AI will be well-positioned to help their firms thrive throughout market and economic cycles. But you don't have to navigate this new frontier alone.

Built on a foundation of industry-leading data and decades of experience understanding and meeting the most complex workflow needs of the investment community, FactSet has the expertise to guide you through your AI journey, helping you identify use cases and design and implement high-impact solutions.

Our pioneering GenAI tools deliver real, relevant value to users across the buy- and sell-side, elevating search intelligence, automating workflow tasks, and enabling financial professionals to focus efforts on high-value work. Learn more about our innovative AI capabilities to help you focus on what's important for your business.

The following FactSet experts contributed content and insights to this e-book.

Ryan Roser, Head of AI and Machine Learning

Yogi Miraje, Lead Machine Learning Engineer

Shreyas Valmiki, Principal Software Engineer

Lucy Tancredi, Senior Vice President of Strategic Technology Initiatives (retired)

AI TERMINOLOGY

Descriptions of AI and related technologies vary across operating environments.
For the purpose of this e-book here's a summary of key terms discussed.

Artificial Intelligence
A branch of computer science that creates and trains systems or models to imitate how humans perform intellectual tasks.

Context
Context is relevant information, circumstances, or conditions that surround a piece of data. Context helps an AI system understand the nuances and specifics of the information it is processing, allowing it to generate more accurate and appropriate responses for users.

Generative AI
Generative AI models generate new content such as text, images, videos, and audio. Large datasets are used to train a model to learn patterns and structures. The model then can create new content based on the patterns it recognizes from its training.

Large Language Models
LLMs process and generate human text at an advanced level. They are trained on enormous amounts of text data to learn the statistical patterns, grammar, syntax, and meaning of human language. LLMs have a remarkable capability to process and generate text that is coherent and contextually relevant. They can also translate text to other languages, summarize articles, answer questions from documents, and enable chatbots.

Models
Mathematical tools that help computers become more capable. When trained with accurate, high-quality data, AI models can enable computers to recognize images, make decisions, solve problems, and communicate.

ABOUT FACTSET

FactSet (NYSE:FDS | NASDAQ:FDS) helps the financial community to see more, think bigger,and work better. Our digital platform and enterprise solutions deliver financial data, analytics, and open technology to more than 8,200 global clients, including over 218,000 individual users. Clients across the buy-side and sell-side as well as wealth managers, private equity firms, and corporations achieve more every day with our comprehensive and connected content, flexible next-generation workflow solutions, and client-centric specialized support. As a member of the S&P 500, we are committed to sustainable growth and have been recognized amongst the Best Places to Work in 2023 by Glassdoor as a Glassdoor Employees' Choice Award winner. Learn more at www.factset.com and follow us on X and LinkedIn.