INTEGRATING GENERATIVE AI TO DRIVE EFFICIENCY IN SPEND INTELLIGENCE AND NEGOTIATION STRATEGY

by

RIA VERMA

Bachelor of Science in Economics and Supply Chain Management, Arizona State University (2019)

and

ANDRES ARTURO AYALA MENNECHEY

Master of Science in Big Data and Business Analytics, IE Business School (2022)

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Signature of Author:..
Department of Supply Chain Management
May 10, 2024

Signature of Author:..
Department of Supply Chain Management
May 10, 2024

Certified by...
Dr. Elenna R. Dugundji
Research Scientist

Certified by...
Dr. Thomas Koch
Postdoctoral Associate

Accepted by..
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Integrating Generative AI to Drive Efficiency in Spend Intelligence and Negotiation Strategy

by

Ria Verma

and

Andres Arturo Ayala Mennechey

Submitted to the Program in Supply Chain Management
on May 18, 2024, in Partial Fulfillment of the
Requirements for the Degree of
Master of Applied Science in Supply Chain Management

Abstract

Generative AI and large language models are transforming industry practices by enhancing productivity for developers and knowledge workers. This project aims to develop a generative AI-based chatbot for our sponsor's category management team. The chatbot utilizes company data alongside natural language to deliver domain-specific insights, such as pricing, supplier spend, risk assessments, and Bill of Materials analysis. This integration lessens dependence on traditional tools like Excel, streamlining negotiations and democratizing data access, thus reducing the learning curve for new hires. Our approach involves a retrieval-augmented-generation model with Langchain's text-2-SQL agent in parallel with Langchain's kuzuQAchain agent on a graph knowledge database. These agents generate SQL or Cipher code based on queries to retrieve and deliver accurate information in natural language. The model's effectiveness was evaluated through precision testing against various prompts. Using instruction tuning and prompt engineering, the chatbot returned accurate and contextually relevant responses for selected use-cases with no hallucinations, but experienced limitations with token handling in complex queries. We recommend embedding the chatbot into existing dashboards, starting with a pilot targeting technically adept new managers to refine data retrieval processes, and planning subsequent expansions to cover additional use cases.

Capstone Advisor: Dr. Elenna R. Dugundji
Title: Research Scientist

Capstone Advisor: Dr. Thomas Koch
Title: Postdoctoral Associate

Acknowledgments

We extend our deepest gratitude to our advisors, Dr. Elenna R. Dugundji and Dr. Thomas Koch, for their invaluable guidance and support throughout this project. Their encouragement and challenges pushed us to expand the boundaries of our research, providing us with numerous resources and opportunities to explore and share this transformative technology further. Thank you to our writing advisor, Ms. Pamela Siska, for her appreciation and encouragement.

We are also immensely thankful to the myriad of individuals who enriched our learning experience. This includes the wealth of knowledge shared by professors, fellow researchers, and professionals from various conferences, research journals, and discussions. Special thanks to the professors at MIT and the numerous companies who engaged with us, sharing insights into their work, and demonstrating their cutting-edge developments. Each interaction has left a significant imprint on us, shaping our academic journey, and influencing our future career decisions. This project would not have been possible without the collective wisdom and contributions of all those involved.

Andres Ayala: I want to thank my family, Luis Arturo, Odette, Alejandra and especially my wife, Michelle. Without you I would not be the person that I am today. I also want to thank my amazing partner Ria Verma for your professionalism and human values, it’s been an amazing journey and I have learned so much from you.

Ria Verma: I want to thank my parents, Sanjay and Jyotsna Verma, and my big sister Tushita Verma, for their endless support and belief that I can take on hard challenges. Furthermore, I am so grateful for my amazing partner and friend Andres Ayala for always encouraging me to push the limits and embrace my passions.

Table of Contents

1 Introduction

1.1 Motivation

Our sponsor company is a global healthcare company that conducts research, development, manufacturing, and distribution of healthcare products. The company operates in several segments including pharmaceuticals and medical devices and equipment.

Procurement spend for the company is several billion dollars annually due to the vast portfolio of items and materials. Currently, they have advanced procurement data collection from various enterprise resource planning (ERP) data systems that are accessed by category managers, primarily through their supply chain procurement business intelligence dashboards. However, these big database sources are largely untapped resources that have not been used to full capacity.

Foundation models and generative AI are currently experiencing a surge in attention and expectations, leading to their peak in the hype cycle. As noted by Gartner, this surge is driven by the technology's substantial impact on enhancing the productivity of both developers and knowledge workers (Gartner, 2023a). This significant boost in efficiency is leading organizations, like our sponsor company, to reevaluate and adapt their business models. Central to this technological advancement is the use of large language models (LLMs), one of the key intelligence tools designed to understand and generate human language. Across supply chains, functions that are particularly data-rich benefit the most from the adoption of LLMs, such as contract management and negotiation, scenario analysis, supplier selection, risk resilience, and data retrieval and interpretation. However, many organizations struggle to overcome the complexities of integrating generative AI into mainstream operations, one reason that projects often fail. Before investing extensive resources into a full-scale AI model or an off-the-shelf solution, the company's approach is to first pilot a generative AI solution with short phases of rapid experimentation to test how the technology can add strategic value.

1.2 Problem Statement & Research Questions

The goal of our sponsor company is to explore and incorporate generative AI technology to improve productivity and efficiency for category managers. There are many areas of opportunity within the space leading to the following research questions:

1. AI Insights: The company currently uses 30+ disparate transactional and master data systems, leading to a substantial amount of unused data from which few insights are being extracted. How can the sponsor company unlock value using large language models and generative artificial intelligence to retrieve and connect insights across data sources?

2. User Perspective: How can generative AI be used to democratize data accessibility and encourage data-driven decision-making for supplier discussions by category managers of all technical levels and work experience?

3. Category Management Experience: How can the company enhance negotiation capabilities and efficiency, such as establishing fair prices and identifying top suppliers for each category, using a generative AI data retrieval tool built upon their proprietary data?

1.3 Scope & Expected Outcomes

The objective of this project is to provide our sponsor company with a generative AI-based proof-of-concept chatbot combining company data with natural human language to provide domain-specific knowledge on pricing insights, supplier spend, risk evaluation, and Bill of Materials details for finished goods.

The expected deliverables include the following:

1. Detailed research and defined use cases of current enterprise solutions with an understanding of limitations and challenges.

2. Proof-of-concept AI model using the company's proprietary data within a secure DataBricks environment, tested against several use cases with spend, pricing, and Bill of Material data sources from their enterprise resource planning systems, with testing for accuracy and sensitivity to prompts.

3. Integration of graph network for more complex questions involving multiple data sets.

4. Chatbot user interface example using masked data for demonstrations.

The expected outcomes include:

1. Reduce time to retrieve data insights by removing the need to click through several tabs and filters on a dashboard or export data into Excel to build pivot tables.

2. Increase negotiation efficiency by allowing managers to focus on more crucial parts of their work.

3. Democratize data mining capabilities by reducing barriers to data interpretation and analysis, thereby lowering the learning curve for new hires.

4. Reduce dependency on data science teams and ad-hoc requests by adding complex data science models, graph networks, and other advanced data structures to the new architecture.

The scope does not include developing and deploying a full-scale AI solution. It is only based on available data within the dashboards and does not account for data errors or missing information in the raw data sets.

2 State of the Practice

The research presented in this chapter is valuable in understanding the potential impact of generative AI within the supply chain. A comprehensive understanding of the foundational concepts, from a basic neural network to its evolution to transformers and large language models (LLMs), is fundamental for grasping generative AI capacities and limitations. A crucial part of this research is to understand the most successful deployments of generative AI models in supply chain, including an evaluation of their methodologies to measure success and the limitations that these models are currently facing. With this information, we designed a strategic approach that allows us to pilot a state-of-the-practice generative AI solution that generates value to our sponsor company.

2.1 Rise of Generative AI

Technical advancements in generative AI and large language models have ushered in a new wave of human-computer interaction and predictive capabilities for corporations and individuals. Discussions about the applications of LLMs and generative pre-trained transformers (GPT) models, among others, have become dominant and it is believed that in the near future, nearly all professionals will be interacting with this technology (Dhamani & Engler, 2024).

Generative AI is expanding rapidly, but artificial intelligence has been around for decades. Artificial intelligence (AI) refers to "any technique that enables computers to mimic human behavior" (Amini, 2023). This is a broad term that encompasses machine learning and deep learning. Deep learning can sometimes be used interchangeably with neural networks, but specifically refer to a network with many layers, hence "deep" learning (Hardesty, 2017).

Deep learning and neural networks began as early as the 1940s (Hardesty, 2017). In the last decade, the increase in big data, enhanced hardware, processing power of graphic processing units (GPUs), and availability of open-source software have enabled greater access to techniques that can support these complex models (Amini, 2023).

Recurrent neural networks (RNNs) were often used for many tasks that have now been replaced by the transformer architecture (also known as transformers). Transformers first rose to prominence in 2017 based on the article "Attention is All You Need" written by a group of Google researchers and has since served as a blueprint for several other architectures (Vaswani et al., 2017). In 2018, Google released its large language model Bidirectional Encoder Representations from Transformers (BERT), the same year that OpenAI released GPT. The latest version at the time of this paper, GPT-4, was released in March of 2023. ChatGPT, an "advanced AI conversational chatbot powered by one of OpenAI's large language model (LLM) frameworks...[harnessing] the power of natural language" launched in November of 2022 and has generated more buzz on the potentials of chatbots and generative AI in business (Sullivan, 2023).

2.2 Brief Overview of Neural Networks & Transformer Architecture

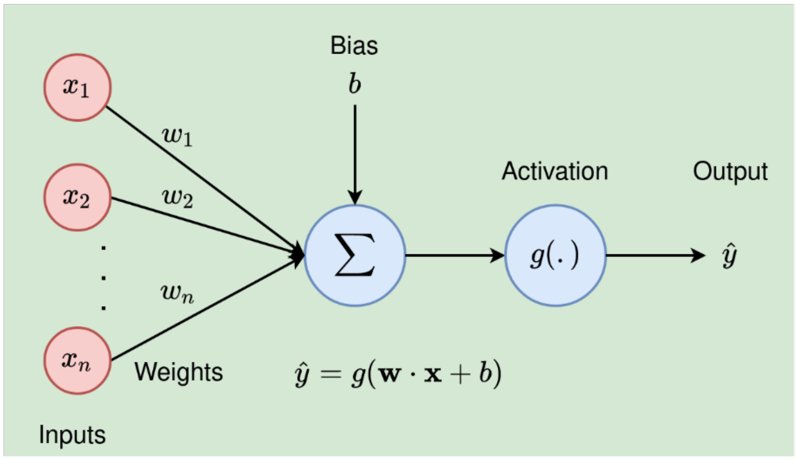
The scope of this literature review will be to evaluate different neural network architectures and will not focus on the intricacies of the mathematical models behind neural networks. The increase in application programming interfaces (APIs) and coding packages creates a layer of abstraction so users can understand and implement generative AI solutions in practice, even without extensive computer science experience. This paper assumes the reader has a high-level understanding of AI and ML.

2.2.1 Neural Networks

As its name suggests, neural networks are inspired by the complex biological network of the brain. Just as the brain is composed of neurons, dendrites, and axons to generate decision outputs, the neural network model consists of units and connections to perform logical computations and generate a mathematical output (Géron, 2022). In its simplest form, a neural network consists of different inputs that the model associates with a term known as weights that represent the relevance of the input node. The model takes these weights and computes a weighted sum of the inputs. Finally, the "neuron" uses a mathematical function known as the step function that takes the weighted sum of inputs and generates an output, as seen in Figure 2-1.

Figure 2-1

Simple Neural Network



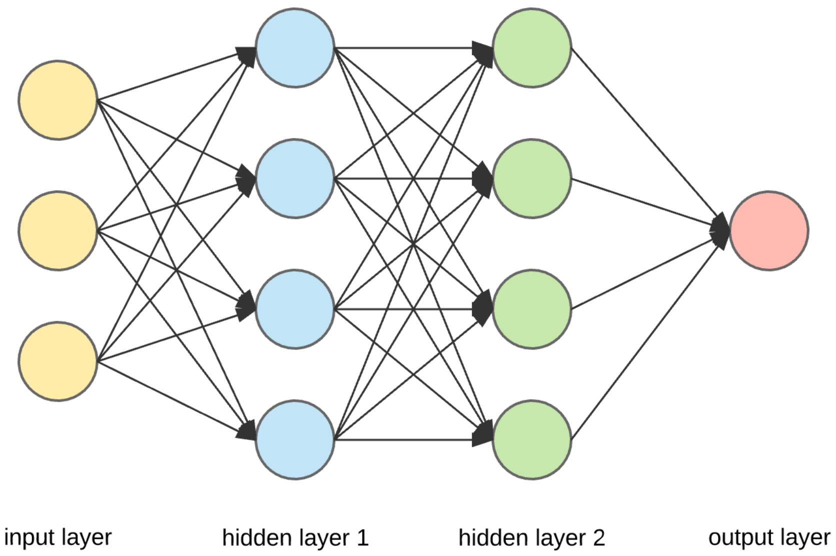
Note: From “The Basics of Neural Networks (Neural Network Series) — Part 1” by Kiprono Elijah Koech.

Neural networks are very versatile, as the input of one neuron can become the output of another neuron. This interlace of connections makes neural networks adaptable to a vast variety of complex tasks, such as image classification, speech recognition or even playing chess (Géron, 2022).

A multi-layer neural network consists of an input layer, a stack of neurons known as a hidden layer, and an output layer as seen on Figure 2-2. When a neural network has several stacked hidden layers they are referred to as deep neural networks (Géron, 2022).

Figure 2-2

Hidden Layers of a Neural Network



Note: From “Everything you need to know about Neural Networks and Backpropagation — Machine Learning Easy and Fun” by Gavril Ognjanovski.

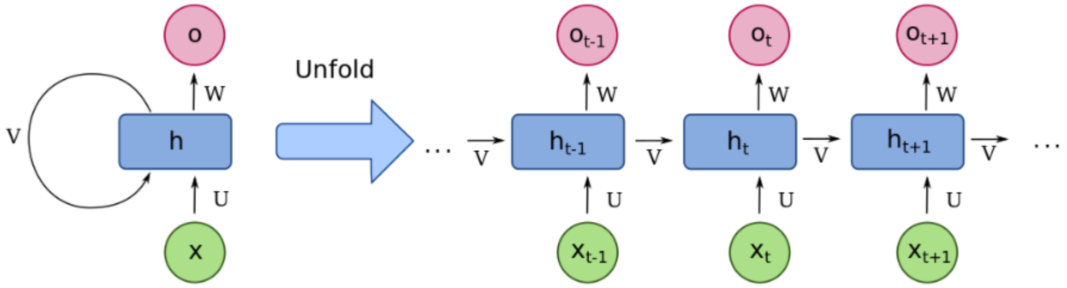
Neural networks aim to minimize a predefined loss function by tuning the weight parameters and measuring the error of the model’s forecast. The model trains in different batches of the dataset and provides an output. The outputs are compared to the target value to calculate the errors of the batch. Then, through a process called backpropagation, the model deciphers the contribution to the error for each connection. Finally, the neural networks follow a process known as gradient descent to correct the weight parameters and reduce the error to an acceptable range (Géron, 2022).

2.2.2 Recurrent Neural Networks (RNNs)

Recurrent neural networks are neural networks that have a connection that feeds the neuron back to itself, as seen on Figure 2-3 (Géron, 2022). This type of neural network tries to model sequences in time steps (Castellanos, 2022).

Figure 2-3

Recursive Neural Network



Note: From “Deep Learning: Recurrent Neural Networks” by Pedro Borges.

RNNs suffer from high complexity issues that demand large computational power. It can take months to train these models. Additionally, they are not able to handle short and long-term relationships of a sequence easily. (Castellanos, 2022).

2.2.3 Transformers and Attention Models

Transformers are neural network models that try to learn the long and short relationships of a sequence based on the idea of an attention layer. The groundbreaking paper, "Attention is All You Need" (Vaswani et al., 2017) outlines the concepts that make transformers much more capable of understanding the complex relationships that exist within human language. Transformers can do the following:

1) Take the surrounding context of the word into account.

2) Take the order of the words into account.

3) Generate output with the same length as the input. (Vaswani et al., 2017).

The positional encoding refers to the combination of the standalone embedding for each word and its order in the sequence of words, or where it exists in the sentence. Each positional encoding is then used to create the contextual embedding, which takes the weighted average of each standalone word in the sentence and its relevance to that positional embedding. Essentially, the contextual embedding assigns a certain amount of "attention" each word has to one another, which is found based on a cosine similarity. The combination of these self-attention layers results in the "multi-head attention" layer which is used to attend to multiple patterns in a sentence, like tense, tone, and relationships between words. The final process for each of these embeddings is to return them to a dense layer of the same size as the input (Ramakrishnan, 2024). All of these processes make the transformer architecture exceptionally powerful and has led to its rise in prominence amongst deep learning architectures for numerous purposes.

2.3 Foundations of Generative AI

2.3.1 Large Language Models

The core of generative AI leverages foundation models which are "massively pre-trained models on a huge corpus of unlabeled internet data"(Chandrasekaran, Miclaus, & Goodness, 2023). They can either be transformer-architecture-based models (like most large language models) or diffusion-based models (like most computer vision models)" (Chandrasekaran et al., 2023).

Simply put, users can pose a query or prompt to an application built upon a foundation model to perform complex tasks. "Foundation models for natural language processing are large language models...trained on information that includes text languages, coding languages, images with descriptions, and weblogs optimized to predict information based on the nature of the prompt" (Khorana, 2023).

LLMs compute the probability of an upcoming word given a sequence of words. (Castellanos, 2022). The model requires deep neural networks that learn from a large amount of data. These LLMs have gone through numerous improvements in the past few years and can be used out of the box or fine-tuned to fit specific needs.

2.3.2 LLM Models and Solution Approaches

There are several approaches to building a functioning application based on a large language model, with different use cases and challenges. Choosing an approach means balancing many trade-offs including control, accuracy, and efficiency.

The first decision is between using a proprietary LLM, an open-source foundation model, or a fully customized model. See the advantages and disadvantages and examples in Table 2-1.

Table 2-1

Pros and Cons of Different Large Language Model Sources

LLM	Definition	Pros	Cons	Source
Custom LLM	Model trained from scratch, typically in-house.	Max control. Unique tasks. Domain-specific info.	Extremely resource intensive, billions to trillions of tokens needed to build the training dataset, including writing the prompts and expected answers	databricks webinar-tang_disrupt
Proprietary LLMs	Proprietary LLMs are owned by a company and can only be used by customers that purchase a license. <u>Examples:</u> OpenAI AuzreOpenAI Bard/ Gemini	Trained on vast datasets (trillions of tokens). Higher accuracy levels. Larger range of use cases. Simpler deployment, accessible with APIs.	Security risks. Less customization than open-source models. Expensive.	team_open_2023
Open-source foundation models	Model where the code and underlying architecture are accessible to the public, meaning developers and researchers are free to use, improve, or otherwise modify the model. <u>Examples:</u> LLaMA V2 Hugging Face MosaicML	Cheaper than OpenAI. Flexible to customize with better control over costs, easier to evolve specific to company goals. Better control over security and privacy (can run on-premises). Fewer barriers to enter/exit, easy to swap and test new solutions. More transparency on how they are fine-tuned and trained.	Greater investment needed in data engineering, tooling integration, and infrastructure. Constant model iterations/versions needed. Varied licensing requirements. Higher need for skilled fine-tuning and model development personnel. Can have a lower accuracy.	robuck_feature_2023 amos_quick_2023 team_open_2023
Out of the Box Solutions	Solutions that are made by companies with low-code options such as drag-and-drop	Low development requirements. Fastest to deploy to market.	Costly. More rigid; difficult to customize to individual needs, like adding a GNN. It doesn’t negate the need for data preparation which	

	features and fixed interfaces. Examples: CustomAI GPT		will still require coding solutions/ data science.	
Azure Cognitive Search with Azure OpenAI Service	It gives users access to AI tools (like GPT) through Microsoft's Azure platform, making it easier to add smart features like natural language understanding to apps, websites, and other digital products.	Uses cognitive search to index the data across large knowledge bases. Easily ingrained with existing Azure tools like Azure Data Factory, DataBricks, Azure Data Warehouse. "State-of-the-art".	Costly. Longer time to acquire the subscription.	Castro, 2023

The most difficult method is building an LLM from scratch. While it allows maximum control over the data the model is being trained on, it is extremely resource-intensive and requires billions to trillions of tokens for training (Tang & Koleva, 2023). Due to the complexity of the technology and the need for high levels of computing power, a few players are currently dominating the proprietary LLM market. Typically, developers can access their APIs to build models on top of, paying per token or input.

As of 2023, the key players include OpenAI (GPT), Meta (Llama), Google (Palm and Bard), as well as several open-source contributors.

The most prominent player in the industry and the makers behind ChatGPT is OpenAI. Microsoft partnered with OpenAI in January 2023 enabling a significant sharing of resources, financial support, and access to Microsoft servers. This partnership also opened up OpenAI's integration to AzureOpenAI, Bing, and the development of Microsoft CoPilot (Edell, 2023). OpenAI launched GPT-4 in March of 2023 with significantly fewer hallucinations (misleading, fabricated, or inaccurate answers (Dwivedi et al., 2023)) and higher levels of accuracy than GPT 3.5 (OpenAI, 2023). OpenAI also launched customizable GPT models which allow users to create applications that can search the web, analyze data, or even generate images.

Another popular model is Meta's LLaMA (Large Language Model Meta AI) model, which can be downloaded locally and accessed without using an API, reducing the concern of sending private data over the internet. It also allows for greater customization and flexibility, at the tradeoff of increased setup and operational costs (Robuck, 2023).

Hugging Face is an open-source library providing thousands of pre-trained models, offering "automation and governance tools and curated datasets, as well as model APIs and generative AI applications, targeting enterprise needs" (Chandrasekaran et al., 2023). The biggest benefits of using an open-source foundation model over a proprietary LLM are the increased flexibility for customizing, better control over security and privacy, and ease of testing out different models due to lower barriers to entry and exit. However, they generally require a greater investment in technically skilled resources and more highly skilled fine-tuning (Ramos & Chandrasekaran, 2023). Currently, the accuracy is lower for these open-source models compared to the enterprise solutions.

There are several APIs, wrappers, and tools that abstract the need to delve directly into the LLM, such as LangChain. LangChain is an open-source Python library that works as a framework to develop applications with LLM models (Topsakal & Akinci, 2023). LangChain orchestrates the application through an API connection to the LLM models, which can be private such as OpenAI or open-source such as LLaMA. It also allows the user to include a prompt template where the developer can type the instructions or inputs to the LLM to customize the format of the responses. Additionally, LangChain has "agents" that are tailored solutions for specific tasks.

For example, a SQL agent takes a prompt, interprets it using natural human language, creates a SQL query to pull the relevant data, and returns the answer to the user's question (Dieruf, 2023).

2.3.3 Optimizing Generative AI Models

A webinar given by Colin Jarvis (OpenAI) and Luv Kothari (ScaleAI) explained the key components needed to customize a generative AI solution for a company's use case. There are three main methodologies in the optimization of a generative AI model: prompt engineering, retrieval-augmented generation (RAG), and fine-tuning (AI, 2023).

Prompt engineering, the most basic, involves asking differently phrased questions or prompts to elicit more specialized information from the model. It's beneficial for fast, on-the-fly model guidance, but not easily controlled or scaled (Tang & Koleva, 2023).

Retrieval-augmented generation (RAG) involves giving the model extra data and information to base its answer. In RAG, the data or documents are embedded and/or stored in a knowledge base, which can be in the form of a vector database or another data storage. The user question is interpreted by the LLM which uses a retriever component to search the database for relevant information that enhances the LLM's response to the question. The final output to the user's query is typically composed of both the model's pre-trained knowledge and the specific information provided. This method introduces new information to the model to update its knowledge store and can also reduce hallucinations by forcing the model to restrict data retrieval to a specific knowledge source. However, it is not as useful when scaling to a broad domain due to the retrieval process, and the addition of more context and tokens can also increase the price of the solution. It's often better to use RAG when there is a short-term solution to solve, and accuracy is prioritized over computational efficiency and cost (AI, 2023).

Fine-tuning involves training the model further on a smaller highly domain-specific data set to optimize a model for a specific task. It is best used for customizing the structure of responses based on complex instructions derived from knowledge that already exists in the model instead of introducing new information like with RAG. It can improve the model's efficiency and reduce the number of tokens needed to execute a task. In many cases, a fine-tuned GPT-3.5 model can outperform a GPT-4j model; in one example, a fine-tuned OpenAI model was able to execute a language-to-SQL task 84% better than a broad model after being fed a series of examples of specific SQL scripts (AI, 2023). Many enterprise solutions today use a combination of these methods to achieve the highest result.

Table 2-2

RAG, Prompt Engineering, vs Fine-Tuning

Model Approach	Definition/ Use Cases	Pros	Cons	Source
Prompt engineering	Crafting specialized prompts to guide model behavior. Used for quick- on the fly- model guidance.	Fast, cost effective, no training involved.	Not as controlled as fine-tuning. Harder to get better accuracy. Time intensive.	databricks webinar-tang_disrupt
Retrieval-Augmented - Generation (RAG)	LLM with external knowledge retrieval through knowledge base/ vector database. Used for Dynamic datasets, companies with lots of personal data.	Better accuracy than prompt engineering. More up-to-date data. Domain-specific or private data. Brings organizational context without the complexity and cost of fine-tuning. Reduced hallucinations due to grounding on organization data.	May increase latency (more time to answer a question), impacting the viability. Needs to incorporate vector databases and embedding models which involves a new architecture. Potential security risks.	databricks webinar-tang_disrupt choose an approach for embedding - chandra_how_2023
Fine tuning	Fine-tuning is a classic approach for "transfer learning" aimed at transferring knowledge from a pre-trained LLM to a model tailored for a specific application. Adapting a pre-trained model to specific datasets. Used for domain or task specialization.	Granular control. High specialization.	Can risk hallucinations by over-fitting / over-training the model. Involves hyperparameter tuning; needs more experienced technical developers. Computationally resource intensive; needs thousands of examples and necessitates GPUs.	databricks webinar-tang_disrupt

2.3.4 RAG Data Ingestion: Knowledge Graph vs Relational Databases

A common application enabled by LLMs is creating contextual dialog applications such as chatbots (Hadi, Tashi, et al., 2023). To build a chatbot for private data using a RAG approach, the data is ingested from a source database to the LLM, which then retrieves the relevant information and generates the correct answer. There are several ways of ingesting the data to the LLM, but the approach depends on the database where the data is stored. Some example databases are using CSV files or relational databases consisting of rows and columns. Yet there is another way of ingesting data to LLMs through a database known as a knowledge graph (Langchain, 2024).

In essence, a graph knowledge base consists of nodes and edges that capture relationships between nodes. Knowledge graphs are excellent for contextualization of data simply and efficiently in terms of computational memory (Yse, 2023). Even though several papers have written about the excellent performance of LLMs in using SQL (programming language for relational databases), there does not seem to be a significant amount of papers replicating this evaluation with cipher (programming language for graph knowledge databases) (Salihoglu, 2024). There is a large opportunity for this type of database; companies such as Microsoft are creating graph knowledge databases using LLMs that exhibit significant contextualization improvements when performing Q&A (Larson & Truitt, 2024). The capacity of graph knowledge databases to contextualize data in an efficient way creates a great opportunity for large language models to work with complex data and gather enriching insights efficiently (Yse, 2023).

2.4 Use of AI in Business

A study by McKinsey Digital estimates that generative AI has the potential to add $2.6 trillion to $4.4 trillion to the U.S. economy by impacting the productivity of retail, high-tech, banking, and life sciences, among other industries. In the retail and consumer packaged goods industry, the estimated impact ranges between $400 billion to $660 billion per year. These numbers are enhanced by 35%-70% when the applications of generative AI are embedded along other machine learning and business analytics models(Chui, Hazan, Roberts, Singla, & Smaje, 2023).

Generative AI capabilities of processing natural language, performing speech recognition, generating images, and producing code, foment a broad spectrum of applications for business management (Bi, 2023). These applications are not only related to automating processes but also creating innovative and disruptive outcomes that are expected to revolutionize functions such as customer interaction, marketing ideas generation, software development and maintenance, and R&D (Chui et al., 2023). According to McKinsey, generative AI initiatives should be evaluated not only in terms of cost reduction but also in terms of labor productivity potential. Both schemes are complimentary and may generate financial benefits (Chui et al., 2023).

There are many growth opportunities in the supply chain space, especially in increasing the efficiency of the workforce. A study conducted by Harvard Business School and Boston Consulting Group (BCG) tracked 758 BCG consultants' work productivity and accuracy using 18 realistic consulting tasks. The results showed that those who used GPT-4 completed "12.2% more tasks, 25.1% quicker [with] 40% producing higher quality results." However, the same group was "19% less likely to produce the correct solutions compared to those without AI" (Dell'Acqua et al., 2023).

Cases like this challenge the idea of whether having speed and increased productivity is favorable overtaking a longer amount of time but getting more accurate answers. As the foundation models continue to advance, there is a great potential for continued improvement and mitigation of the accuracy issues, underscoring the benefits of increased efficiency with fewer risks. For example, GPT-4 has already been shown to be faster than GPT-3.5 (OpenAI, 2023).

2.4.1 Use Cases in Procurement & Supply Chain

The advancements in generative AI are already being implemented in procurement and supply chain. Some examples mentioned in an interview conducted by the Chartered Institute of Procurement and Supply Chain with executives of Shell and Maersk are digital assistants, standardization of contracts, user training and upskilling, question-answering capabilities, contract understanding, spend analysis, supplier selection, and negotiation support in procurement. Generative AI can improve productivity in these areas and allow procurement professionals to focus on more strategic aspects of their jobs (CIPS, 2022).

In a report by Frost & Sullivan, within procurement, AI and ML have can be useful for the following tasks: predicting price changes, classifying spend into various categories, vendor matching based on invoice and purchase order history, capturing supplier or market data through web posts and alternative data channels using natural language processing (NLP). Solutions leveraging generative AI can be especially useful to "analyze purchase history data and provide insights on material procurement areas to enable the company to manage risks, create saving opportunities, and optimize buying power" (Sullivan, 2021).

Walmart has already adopted a generative AI tool to automate their supplier negotiations. Instead of a buying team, the AI tool that is trained on Walmart's specific needs communicates with humans to close the deals. Walmart said "it's successfully reached deals with about 68% of suppliers approached, with an average savings of 3% on contracts handled via computer since introducing the program in early 2021."(Sirtori-Cortina & Case, 2023). Other companies like Amazon also invest in automated vendor discussions, but companies like rival Target do not. The study mentioned that a common fear about using automation for account management and procurement is that the machines and algorithms aren't able to benefit from human- vendor relationships, which are often pivotal to negotiation strategy (Sirtori-Cortina & Case, 2023). Nishant Srivastava, a solutions architect at Microsoft, explains how using OpenAI's question-answering models for contract interaction, evaluation, and analysis can accelerate the productivity of procurement managers by helping them focus on strategic aspects instead of doing unreliable manual procedures (Srivastava, 2023). An LLM solution such as

ChatGPT that has been fine-tuned on procurement contracts can then be used to query relevant insights in a fast and efficient way. The solution becomes a tool that empowers procurement managers with better negotiation capabilities. The manager could answer queries such as "What were the termination conditions of the last 3 contracts for supplier ABC" or "What are the lead times established in Contract XYZ for SKU1234?". (Srivastava, 2023).

Another promising use case of generative AI in procurement is in supplier selection. As explained by Deloitte, selecting the right supplier is not an easy task. It requires the evaluation of structured and unstructured data related to products, supplier historic performance, prices, payment options, geography, and risk profile, among others. Generative AI can assist with this task and adapt to specific procurement requirements (Rajani & Deng, 2023). In an interview with a former Chief Data Scientist of one of the world's largest aircraft manufacturers, a similar use case was developed using deep neural networks trained on aerospace parts descriptions and supplier data. The model managed to rationalize complex information and increase the pool of supplier options, allowed to determine a fair value for the parts, and enhanced the negotiating capabilities of the company. This project accounted for $1.7 billion in savings for the company. Although this project was developed 3 years before generative AI was available (2022), it is a successful example of the applications of deep learning models in procurement. The interviewer explained that if generative AI had been available at the time, the project would have required less time, and most likely greater insights may have been gathered (Vu, 2023).

2.4.2 Evaluating Performance

The transformational capability of generative AI requires a tailor-made performance metric for each enterprise goal and specific use case. Therefore, a strategic approach must be followed when defining the performance metric of any generative AI project. The first step is to establish the goal of the generative AI application and make sure that it is in alignment with the enterprise mission and department goals. This will promote the adoption and funding of the project (Gartner, 2023b). Once the goal is stated, the generative AI value driver must be defined and as a consequence, the performance metrics will become the link between the goal and the value driver (Gartner, 2023b). For example, if the goal of the generative AI deployment project is to increase the productivity of the category managers, the value driver of the solution would be to diminish the time required to gather precise and useful insights from a procurement database. The performance metrics that connect the goal to the value driver would be time to answer a question, number of clicks required to find the answer, and accuracy of the answer found. (Gartner, 2023b)

Once the performance metrics are defined a baseline for success must be defined (Coveo, 2023). Previous measures of key performance metrics (KPIs) may be used as part of the benchmark but new measurements may be adapted to evaluate the success of the generative AI implementation. For productivity, task completion tests and surveys are a common practice to evaluate success. The performance metrics may also be segmented by employee level (years in the company) or domain knowledge of the task to be completed. These segments are also tailor-made to the specific use case and contribute to measuring the impact of the initiative as a value driver of the established goal (Dell'Acqua et al., 2023).

One of the most relevant KPIs to keep in mind is the return on investment (Fernando, 2023). Measuring the financial outcome of deploying a project is not an easy task, but is certainly a very important metric, as most AI projects are not profitable yet (Ashoori, Goehring, Humphrey, Naghshineh, & Rodenbeck Reese, 2023). The financial outcome of a generative AI project is key for the mid-long term evaluation of the initiative as investors' pursuit for returns will keep generating pressure on unprofitable projects. This is a very important concern. Gartner research estimates that 50% of the companies that have deployed large language models initiatives will be forced to desert those projects by 2028 in part due to costs that surpass the financial benefits (Wiles & Sallam, 2023).

2.4.3 Limitations of Generative AI

Businesses must act quickly to capture the benefits of generative AI, but certain risks need to be consciously considered when deploying this technology. Firstly, one of the most relevant risks to focus on is content quality. There is a constant threat of hallucinations that these models can produce when there is no feasible answer. These models generate different answers to the same prompt question creating the need for a human user to assess the accuracy of the answer. Secondly, the security aspect is critical given that sensitive data leakage may occur while accessing the large language models owned and operated by a third-party company like OpenAI that can assure data privacy. Thirdly, the data on which the model is trained could have biases that will be introduced and included in the output. Finally, the workforce's reluctance to adopt these disruptive solutions as a tool due to fear of being replaced creates a barrier to the adoption of generative AI models (Chui et al., 2023).

Hallucinations, Confabulations, and Consequences

Even though generative AI's ability to create coherent responses and conversations has been astonishing, they do not abstain from errors. When this type of model generates misleading, fabricated, or inaccurate answers it is commonly referred to as hallucinations (Dwivedi et al., 2023). Nevertheless, this term does not precisely describe the outcome that is being generated by the model. A hallucination in clinical terms is a false sensory perception that leads to ambiguous behavior. On the other hand, a confabulation is a fabricated logical statement(Hatem, Simmons, & Thornton, 2023). For example, citing a reference that does not exist is a clear example of confabulation, and it is also an example of one of the types of errors that generative AI may produce. Therefore, the most adequate linguistic term to describe the inaccurate information of large language models' errors is confabulation (Hatem et al., 2023).

Confabulations are extremely dangerous due to the over-confidence that users have on these tools (Fui-Hoon Nah, Zheng, Cai, Siau, & Chen, 2023). Using these models without a keen understanding of their limitations, specially of hallucinations, may create terrible outcomes. For example, the model may be suggesting a user to take an incorrect medication without having a professional evaluation (Sallam, 2023).

Generative AI and large language models failures are agnostic of topics or categories and a large list of failures in terms of reasoning, fact reliability, mathematics, and coding, among others have been developed (Borji, 2023) as proof of the issues that may arise from blindly accepting the answers provided by the model. For reference, ChatGPT can reach high scores close to 90% and 95% on the United States Medical Licensing Exam and 84% on the Law Examination of Constitutional Law, but for the Ophthalmology test, it scored approximately 50% and a surprising 27% in Taxonomy (highly mathematical) (Shahriar & Hayawi, 2023). This evidences that the model still has mathematical and reasoning challenges that it cannot solve.

Ethics and Privacy

Generative AI and large language models have been trained using large amounts of data from the World Wide Web. This means that the data on which these models were trained were sometimes unethically biased and therefore the output may also show biased results. The unethical use of private publications or academic papers without the ability to justify citations is a prevalent issue for these models (Shahriar & Hayawi, 2023).

Private data security is one of the most delicate aspects of using large language models. This issue affects individuals, companies, and governments, as the rising use of tools like ChatGPT exposes confidential data to be leaked into the model (Fui- Hoon Nah et al., 2023). This has made generative AI applications an important target for hackers looking to steal sensitive data (Chui et al., 2023). Different techniques such as reverse psychology and prompt injection attacks have been proven to work in generative AI applications to unethically surpass the cyber-security of the application (Gupta, Akiri, Aryal, Parker, & Praharaj, 2023). There have also been reported bugs in the applications that temporarily exposed the chat history of users (Porter, 2023), exposing private information and opening concerns on the security levels of the tool.

Explainability

Machine learning and artificial intelligence solutions are greatly trusted when the output of the model is explainable and interpretable. Both components are critical aspects of transparency of a model’s behavior (Balasubramaniam, Kauppinen, Rannisto, Hiekkanen, & Kujala, 2023). Large language models are created using complex deep neural networks with over 175 billion parameters that derive from a diverse and vast amount of data gathered from different sources. The complexity of these models makes explainability and interpretability a very complicated task to achieve. Not having the capacity to understand the reasons behind a decision generated by a large language model amplifies the ethical, transparency, and trust issues of this solution (Hadi, Qureshi, et al., 2023).

3 Methodology

This section outlines the methodology used for developing and evaluating the generative AI-based chatbot for our sponsor's category management team. It describes the selection and implementation of large language models and retrieval-augmented-generation techniques, integrating Langchain's text-2-SQL and kuzuQAchain agents into a graph knowledge database. The methodology also includes precision testing, instruction tuning, and prompt engineering to ensure the accuracy and relevance of the chatbot. These steps address key AI implementation challenges such as data hallucinations and complex query management, emphasizing the chatbot's adaptability and scalability in meeting business needs.

3.1 Technical Approach

The goal of this project is to test the question-answer capability of generative AI with real company data and demonstrate the applications with the category managers to understand the value and potential adoption risks. The intent of focusing on a pilot with only a few use cases and data sources is to test the hypothesis that a generative AI solution would provide benefits and is worth the significant investment in time and resources to deploy the tool in production. Creating a minimal viable product will help to circumvent the constraints that the engineers can face.

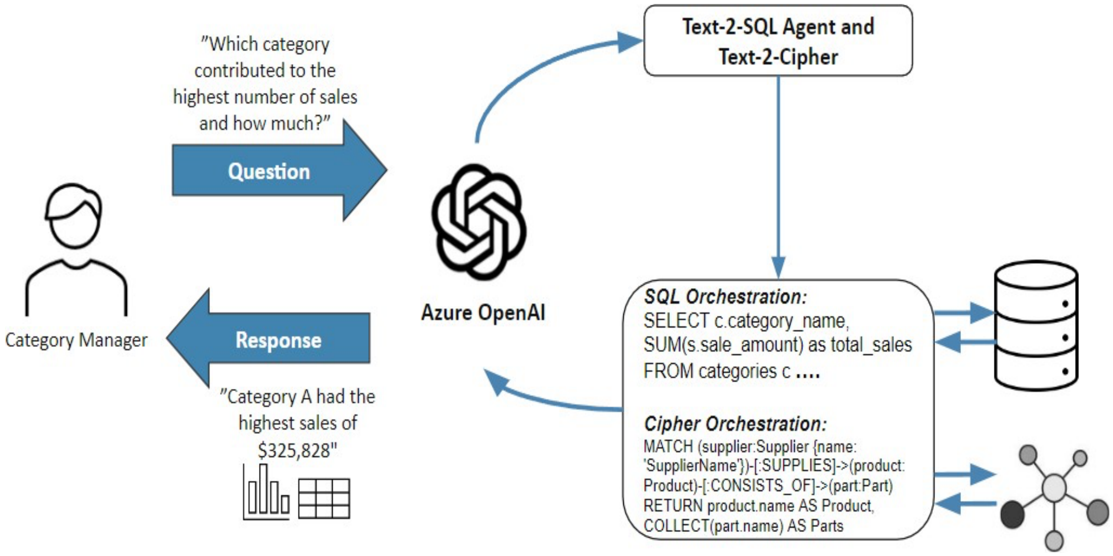
After evaluating the advantages and disadvantages of several foundation models, we decided to proceed with using Azure OpenAI's LLM (GPT-3.5 Turbo). The sponsor company is currently using Microsoft Azure's Cloud Platform, including Azure DataBricks and Azure Data Warehouse as their data storage platform. After providing our justification to the company's internal LLM governance team, we obtained access to the company's Azure OpenAI API through Azure DataBricks, enabling our models to pull directly from the data within their data containers. This addressed the security/privacy issue, as Microsoft platforms are closed environments that prevent the data from being exposed, unlike using OpenAI's platform or another external company that provides a proprietary LLM.

After evaluating fine-tuning, prompt engineering, and Retrieval Augmented Reality (RAG), we decided to use a RAG approach to feed our company’s data on supplier spend information to our model’s knowledge base. This approach is less computationally expensive and, as shown in our research, can reduce the frequency of hallucinations. We leveraged the current relational database architecture as a structure to feed the LLM. The chosen text-2-SQL agent from LangChain (after testing several different open-source options) uses generative AI to turn the natural language queries posed by the user into structured query language (SQL), which can then be run against the related knowledge base (in this case, spend data) to retrieve an answer. This involves a combination of instruction tuning and prompt engineering to focus the model on answering specific use cases.

In parallel, we built a knowledge graph database from the sponsor company’s relational database using Kuzu and NetworkX Python libraries. We used this database structure to feed the LLM model and evaluate whether we could achieve better accuracy in more complicated queries than with the relational database. To create the graph knowledge base, we combined different datasets from the relational database. We created a directed graph of the Bill of Materials (BOM), connecting all finished goods with their respective raw materials and combining this information with the vendor information of each raw material. After constructing the knowledge graph database, we deployed a text-2-Cipher agent named kuzuQAchain (Cipher is the query language for graph networks). The agent works similarly as the text-2-SQL agent. It takes the input prompt and generates Cipher code that looks for the answer inside the graph knowledge database. With this, it’s possible to answer question such as "If supplier ABC has a disruption, which BOMs will be affected and what is the total final product revenue at risk?" The final architecture will utilize Langchain’s text-2-sql agents and kuzuQAchain as seen on Figure 3-1.

Figure 3-1

Chatbot RAG Architecture with Langchain and Kuzu

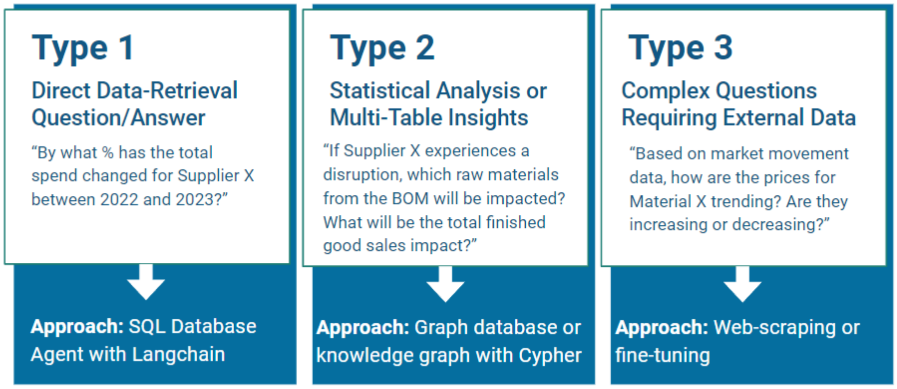

"Which category contributed to the highest number of sales and how much?"
Question
Text-2-SQL Agent and Text-2-Cipher
Azure OpenAI
Category Manager
Response
"Category A had the highest sales of $325,828"
SQL Orchestration:
SELECT c.category_name, SUM(s.sale_amount) as total_sales FROM categories c ....
Cipher Orchestration:
MATCH (supplier:Supplier {name: 'SupplierName'})-[:SUPPLIES]->(product: Product)-[:CONSISTS_OF]->(part:Part) RETURN product.name AS Product, COLLECT(part.name) AS Parts

3.2 Use Case Segmentation

Based on our discussions with the category managers, we compiled a comprehensive list of the types of questions the managers typically answered using data, either through dashboards and/or information passed through Excel sheets. The list was extensive (see appendix A-1 for full list of questions), so we divided them up into three categories of questions, each with a different technical approach to retrieving the answer. See figure 3-2.

Figure 3-2

Question Segmentation



Type 1
Direct Data-Retrieval Question/Answer
"By what % has the total spend changed for Supplier X between 2022 and 2023?"
Approach: SQL Database Agent with Langchain
Type 2
Statistical Analysis or Multi-Table Insights
"If Supplier X experiences a disruption, which raw materials from the BOM will be impacted? What will be the total finished good sales impact?"
Approach: Graph database or knowledge graph with Cypher
Type 3
Complex Questions Requiring External Data
"Based on market movement data, how are the prices for Material X trending? Are they increasing or decreasing?"
Approach: Web-scraping or fine-tuning

Type 1: Direct Data-Retrieval Question/Answer

- **Approach:** Use LangChain text-2-SQL agent; instruction tuning for column names.

Type 2: Statistical Analysis or Multi-Table Insights

- **Approach:** Graph database or knowledge graph with Cypher Type 3 Complex Questions Requiring External Data

Type 3: Complex Questions Requiring External Data (out of scope)

- **Approach:** Web-scraping or fine-tuning

We then narrowed down a select few use cases within Type 1 and Type 2 categories to test and identified the technical approach needed to achieve accuracy.

Use-Case 1: Extracting total spend volume by franchise, category, or supplier

- **Approach:** Use SQL-to-text Langchain agent; instruction tuning for column names

Use-Case 2: Identifying year-over-year spend/ price changes at the aggregate level

- **Approach:** Create an intermediate table that shows the prior year's spend for each row to reduce latency with SQL merge/join statements

Use-Case 3: Connecting BOM and spend data to analyze queries using both tables

- **Approach:** Use graph database and Langchain Cypher agent to query graph database

3.3 Measuring Accuracy

We evaluate the success of our implementation by running different prompts and comparing the results against the actual value from the underlying data tables to find the accuracy percentage. We then run a sensitivity analysis by changing the prompts slightly, varying in degree of how much explanation and direction was given in the prompt, to see if the results vary greatly and how sensitive the methodology is to changes in phrasing.

4 Results and Discussion

This section presents the results and discussion of the generative AI chatbot's pilot, focusing on its performance, accuracy, and user adoption. We analyze the effectiveness of the chatbot in delivering precise and relevant responses with various prompts and explore its impact on category management. Additionally, we discuss the potential barriers to adoption, and key learnings, both technical and operational. This analysis aims to provide insights into how the chatbot can be optimized and scaled across the organization, as well as strategies to mitigate risks to ensure a smooth integration and maximize the technology's benefits.

4.1 Accuracy

For each of the three use cases, we tested several different questions, and their variations, and compared the results to the results that came from the original data source. We have masked any proprietary information for confidentiality.

Our results are as seen on Table 4-1:

Table 4-1

Results from RAG Chatbot Chosen Use Cases

Use Case	**Questions**	**Number of Accurate Responses out of Total Number of Prompts Tested**
Extracting total spend volume by franchise, category,or supplier	Who are the top 10 suppliers for franchise group XYZ?	3/3 prompts
	Which supplier had highest spend per franchise group?	4/4 prompts
	Which materials had "1.5mm" in short description, and what suppliers were they purchased from?	token count too large for all suppliers.
Identifying year-over-year spend/ price changes at aggregate level.	Which 5 categories had the greatest percent change in spend between 2022 and 2023?	3/3 prompts
	What was the % change for each one?	4/4 prompts
	Which suppliers for the franchise XYZ had the greatest % price change?	4/4 prompts
	Which suppliers had the greatest YoY price change?	token count too large for all suppliers.
Connecting BOM data with supplier data to pull insights that leverage both.	If Vendor XYZ faces a disruption, what is the total spend of finished goods affected?	2/3 prompts
	Which raw material does supplier XYZ support and which material category do they belong to?	3/3 prompts
	Which FGs are impacted by changes to raw material XYZ?	2/3 prompts

The queries that result in "token count too large" occur because the large language model must process thousands of rows of data that the code returns and translate that into natural language. The GPT 3.5 Turbo was not able to handle this. However, simply narrowing the scope of the question, such as requesting the top 10 suppliers that meet the criteria instead of all suppliers, can alleviate this issue.

With Type 1 questions, such as those posed in use case 1, the text-2-SQL agent was able to perform well with different prompts. However, as the questions became more complex, to get accurate results a combination of instruction tuning, and prompt engineering had to be used. For example, the category manager colloquially uses the term "YoY" to indicate "year over year" changes in spend, but the large language model does not know how to interpret it unless explicitly indicated. Each use case that may rely on one of these terms needs to be programmed. Though this can seem counter-intuitive to the use of generative AI to perform the querying, the instruction is written only once in the code rather than each time the model runs, so it is still more efficient than writing each individual query from scratch. Without prompt engineering, the model wouldn't know how to query the "franchise" column instead of the "franchise group" column. These learnings stress the importance of defining the use-cases that the individuals intend to use with the model, as well as the importance of testing out each case before deploying the solution to the production environment. Over time, the model should be continuously improved and expanded to generate more capabilities.

Another important finding is that with every prompt we tested, the model did not return any false answers. Instead, it would return an error message if it wasn't able to identify an answer, which is preferred over a hallucination. This is key to maintaining confidence about the model's performance for the users.

4.2 Key Learnings

In this proof-of-concept, the idea of "generative" AI is solely based on the translation of human text into code. The model should not actually *generate* new information or data. Using a RAG approach can limit this issue because the data acts as a reference point for the model, which reduces its reliance on generalizations that could lead to hallucinations.

Originally, the questions posed by the category managers were huge, disparate, and far-reaching. Since different use cases can require different architecture and encoding of company knowledge or acronyms, it's important to select only a few use cases at a time to test for accuracy, and then continuously expand the number and complexity of questions over time.

Less is more. The data clean-up process is crucial, simple actions such as keeping column names consistent can help avoid joining errors and reduce latency. In this proof-of-concept, intermediate tables and views with reduced complexity had to be constructed to generate relevant and timely responses from the LLM and text-to-code agents.

Graph knowledge databases are preferred when querying vast networks because they eliminate the need for expensive join operations typical in relational databases. Graph knowledge databases offer a higher capability of contextualizing the dataset due to its architecture of nodes and directed edges. This methodology is very promising and state-of-the-practice research is currently being developed by top players such as Microsoft and DeeplearningAI.

4.3 Discussion

Using generative AI for a question-answer chatbot is advantageous over historical natural-language-processing (NLP) solutions because they can provide coherent and contextually relevant responses, even ones not explicitly programmed or written from existing documents.

An important question for the project team is whether it is better to use an off-the-shelf AI solution to meet these various demands rather than developing a solution in-house. An off-the-shelf chatbot can avoid the time and cost of developing a tailored solution. However, an in-house solution is advantageous for several reasons. Generative AI is not protected against the well-established "garbage in, garbage out," or GIGO, effect, which underscores the importance of data quality and modeling. The complexity of the models and data architecture significantly impacts the resulting output. Hence, the benefit of designing a tailored model while still using a proprietary LLM is the increased flexibility this affords data scientists to swiftly iterate, test various models, and deploy solutions in secure, private environments. Another advantage of the in-house option is that it can be tailored to specific organizational needs while keeping costs and the need for computationally skilled persons manageable. Furthermore, given the experimental nature of this field, piloting a proof-of-concept first can help foster trust among relevant stakeholders and make it easier to secure internal project funding. The project's "low-hanging fruit" was to address Type 1 questions first. Much higher accuracy was achieved when using careful prompt engineering. This approach poses two critical questions: how can the likelihood of user error be accounted for, and how can category managers be protected from potentially inaccurate information when questions are posed ambiguously?

One strategy is to encourage category managers to learn and embrace pseudocode—a mix of plain language and coding syntax that explains how a program should work—without using actual programming language. Even without a technical background, using words like "filter" and "aggregate" or breaking down complex queries into smaller sentences first can greatly increase the tool's accuracy.

Another approach focuses on refining the user interface. Implementing a drop-down menu can guide users away from entering free-form text where this type of input is not ideal for the model. A help area that clarifies users' objectives before inputting free-form text can be included. In addition, instruction tuning in the design of the application can guide the agent in answering a specific category of questions in a purposely directed way.

Graph knowledge bases are not easy to construct. Corporate products such as Neo4J are available in the market, which offer more capabilities than open-source solutions. We strongly recommend this methodology when evaluating the deployment of LLM models, as they offer an enriching contextualization of complex datasets, allowing the company to easily connect direct and indirect procurement data and perform data mining of Type 2 and Type 3 questions.

Apart from technical constraints, our research and firsthand conversations with the teams highlight several barriers to adoption within the company.

1. Accuracy: There is a risk of under-utilization if users notice inaccuracies at the very beginning of testing. It's important to only release pilots when confident of accuracy for given use cases.

2. Explainability: Negotiations necessitate the ability to back up or justify any recommendations with solid data and reasoning. A solution is to embed the chatbot within the current dashboards to see the original data next to the queries. Instruction tuning can also be used to make the chatbot's responses document any complex SQL language or assumptions made.

3. Complacency: There is a concern of over-reliance for adopters who use it solely

as a substitute for the current solution. It’s important to encourage validation from users during the testing phase and release new use cases one by one only when significant accuracy has been reached under various prompts. 4. Contextualization: It’s not always clear how to measure the baseline accuracy needed for each context. A strategy to address this is to shadow users, follow their process, and document the risk level of inaccuracies for each type of use case. This can also help to deploy iterations of the model more frequently depending on the risk level.

Other than these barriers, additional challenges must be overcome before the chatbot becomes integral to the company’s procurement operations. These include data quality and access issues, API permissions for security, difficulties of latency and accuracy when using relational databases for the LLM, engineering a front-end application with a user-interface, and the reliability of deployed applications. However, the chatbot has significant potential to deliver substantial benefits. This project sets the stage for steady, transformative progress in advanced AI and may establish a new benchmark for efficiency and innovation in procurement.

5 Conclusion

This project tackled the ambitious goal of developing a proof-of-concept generative AI model that could bring significant value to our sponsor's global supply chain procurement. In a company that manages multibillion dollars of procurement spend with diverse raw materials, multiple business sectors, a global presence of suppliers, and several scattered ERP systems, we delivered a functional model to improve efficiency for the category management team.

Our work not only highlights the benefits of applying a generative AI chatbot for procurement managers but also summarizes the challenges to be addressed when implementing this technology. Overcoming hallucinations is a consistent hurdle for this model. Hallucinations carry the threat of returning wrong or unreliable information to answer a question, creating problematic outcomes, and breaking the trust of the user. We dove into prompt-engineering techniques and highlighted the RAG approach as a way of mitigating hallucination issues. Additionally, deep data cleaning and data preparation are required to obtain reliable results.

Our research also calls attention to the need to build a proof-of-concept by focusing on specific questions. When building LLM models, it is not possible to answer all types of questions at once from the beginning. Our work offered a framework that breaks questions into three different types. This helped the construction of a solid working LLM model that minimizes hallucinations.

Additionally, we created a graph knowledge database to answer Type 2 questions. This is currently the most modern approach to implementing LLM models given the greater capability of contextualizing data that a graph database offers. In future work, we recommend comparing relational databases and graph databases to evaluate which strategy offers more benefits in terms of speed, accuracy, and tokens required to answer specific questions.

Generative AI is a transformative technology with the potential to impact virtually every industry, including procurement. The influence of these models on the workforce is profound, comparable to the Industrial Revolution. By pioneering the implementation of generative AI in the challenging environment of procurement, organizations can gain significant competitive advantages, particularly in enhancing workforce efficiency. Our project demonstrates the potential of these models to democratize data mining capabilities, making them accessible to non-technical managers and broadening the scope of data-driven decision-making.

References

AI, S. (2023, November). *Fine-Tuning Open AI's GPT-3.5 to Unlock Enterprise Use Cases - Video.* Retrieved 2023-11-09, from https://exchange.scale.com/home/videos/fine-tuning-open-ais-gpt-35-to-unlock-enterprise-use-cases-2023-11-08

Amini, A. (2023, March). *MIT Introduction to Deep Learning | 6.S191 - YouTube.* Retrieved 2023-10-17, from https://www.youtube.com/watch?v= QDX-1M5Nj7s&list=PLIZzfdW6faP3FviFUY1GtAGVON5YHuBeU&index=113

Ashoori, M., Goehring, B., Humphrey, T., Naghshineh, M., & Rodenbeck Reese, C. (2023). Generating ROI with AI | IBM.

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023, July). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159* , 107197. Retrieved 2023-11-27, from https://linkinghub.elsevier.com/retrieve/pii/S0950584923000514 doi: 10.1016/j.infsof.2023.107197

Bi, Q. (2023). Analysis of the Application of Generative AI in Business Management. *Advances in Economics and Management Research*, *6*(1), 36–36. Retrieved 2023-11-24, from https://madison-proceedings.com/index.php/aemr/article/view/1192

Borges, P. (2018). *Deep Learning: Recurrent Neural Networks.* Retrieved 2024-05-01, from https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010

Borji, A. (2023, April). *A Categorical Archive of ChatGPT Failures.* arXiv. Retrieved 2023-11-24, from http://arxiv.org/abs/2302.03494 (arXiv:2302.03494 [cs])

Castellanos, A. (2022). *"Introduction to Natural Language Processing - Lecture 5, IE Business School, Master of Big Data and Business Analytics 2022".*

Chandrasekaran, A., Miclaus, R., & Goodness, E. (2023, September). *Gartner: A CTO's Guide to the Generative AI Technology Landscape* (Tech. Rep.). Retrieved 2023-11-18, from https://www.gartner.com/en

Chui, M., Hazan, E., Roberts, R., Singla, A., & Smaje, K. (2023). The Economic Potential of Generative AI. Retrieved 2023-11-23, from http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/14313/1/The-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf (Publisher: McKinsey & Company)

CIPS. (2022, January). *Harnessing AI and machine learning to optimise performance in procurement.* Retrieved 2023-10-14, from https://www.youtube.com/watch ?v=8AUPDOaqCp4

Coveo. (2023). *5 Steps From GenAI Use Case to Value Realization | Coveo.* Retrieved 2023-10-14, from https://get.coveo.com/lp/blog/gen-ai/wsj-value -realization/?utm_source=wsj&utm_medium=email&utm_channel=&utm

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., . . . Lakhani, K. R. (2023, September). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality* [SSRN Scholarly Paper]. Rochester, NY. Retrieved 2023-11-23, from https://papers.ssrn.com/abstract=4573321 doi: 10.2139/ssrn.4573321

Dhamani, N., & Engler, M. (2024). *Introduction to Generative AI.* Simon and Schuster.

(Google-Books-ID: LYrxEAAAQBAJ)

Dieruf, D. (2023, November). *What Is LangChain?* Retrieved 2024-03-30, from https://datastax.medium.com/what-is-langchain-b5583de2989a

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., . . . Wright, R. (2023, August). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. Retrieved 2023-11-24, from https://www.sciencedirect.com/science/article/pii/S0268401223000233 doi: 10.1016/j.ijinfomgt.2023.102642

Edell, G. (2023, October). *Insights for CISOs—Preparing for the Dual Nature of Generative AI.* Retrieved 2023-11-23, from https://research-frost-com.libproxy.mit.edu/assets/1/f0a502e4-4463-11e8-b626-1aa9f74f20ad/

Fernando, J. (2023). *Return on Investment (ROI): How to Calculate It and What It Means.* Retrieved 2023-11-23, from https://www.investopedia.com/terms/r/returnoninvestment.asp

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023, July). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25* (3), 277–304. Retrieved 2023-11-24, from https://www.tandfonline.com/doi/full/10.1080/15228053.2023.2233814 doi: 10.1080/15228053.2023.2233814

Gartner. (2023a). *Generative AI: What Is It, Tools, Models, Applications and Use Cases.* Retrieved 2023-11-23, from https://www.gartner.com/en/topics/generative-ai

Gartner. (2023b). *Learn to Build an AI Strategy for Your Business.* Retrieved 2023-11-23, from https://www.gartner.com/en/information-technology/ topics/ai-strategy-for-business

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, *11*, 80218–80245. Retrieved 2023-11-25, from https://ieeexplore.ieee.org/abstract/document/10198233 (Conference Name: IEEE Access) doi: 10.1109/ACCESS.2023.3300381

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (3rd edition ed.). Beijing Boston Farnham Sebastopol Tokyo: O'Reilly Media.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., . . . Mirjalili, S. (2023). A survey on large language models: Applications, chal- lenges, limitations, and practical usage. *TechRxiv*. Retrieved 2023-11-27, from https://www.techrxiv.org/ndownloader/files/41501037

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., . . . Mirjalili, S. (2023, November). *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects* (preprint). Retrieved 2024-03-03, from https://www.techrxiv.org/doi/full/10.36227/techrxiv.23589741.v4 doi: 10.36227/techrxiv.23589741.v4

Hardesty, L. (2017, April). *Explained: Neural networks.* Retrieved 2023- 11-25, from https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Hatem, R., Simmons, B., & Thornton, J. E. (2023). Chatbot confabulations are not hallucinations. *JAMA Internal Medicine*, *183* (10), 1177–1177. Retrieved 2023-11-24, from https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2808091 (Publisher: American Medical Association)

Khorana, N. (2023, May). *Generative Artificial Intelligence Emerges as a Globally Disruptive Technology.* Retrieved 2023-11-23, from https://research-frost-com.libproxy.mit.edu/assets/1/f0a502e4-4463-11e8-b626-1aa9f74f20ad/ 8e86de94-00fa-11ee-b85c-66e28fc9bfff/research?eui=9e9bb4e2-a6b2

Koech, K. (2022). *The Basics of Neural Networks (Neural Network Series) — Part 1.* Retrieved 2024-04- 03, from https://towardsdatascience.com/the-basics-of-neural-networks-neural-network-series-part-1-4419e343b2b

Langchain. (2024). *Q&A with RAG | Langchain.* Retrieved 2024-03-03, from https://python.langchain.com/docs/use_cases/question_answering/

Larson, J., & Truitt, S. (2024, February). *GraphRAG: Unlocking LLM discovery on narrative private data.* Retrieved 2024-03-30, from https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/

Ognjanovski, G. (2020). *Everything you need to know about Neural Networks and Backpropagation — Machine Learning Made Easy and Fun.* Retrieved 2024-02- 01, from https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a

OpenAI. (2023, March). *GPT-4 Technical Report* (Tech. Rep.). Retrieved from https://cdn.openai.com/papers/gpt-4.pdf

OpenAI. (2023, November). *OpenAI DevDay, Opening Keynote.* Retrieved 2023-11-25, from https://www.youtube.com/watch?v=U9mJuUkhUzk

Porter, J. (2023, March). *ChatGPT bug temporarily exposes AI chat histories to other users.* Retrieved 2023-11-25, from https://www.theverge.com/2023/3/21/23649806/chatgpt-chat-histories-bug-exposed-disabled-outage

Rajani, V., & Deng, M. (2023). *Generative AI in Sourcing and Procurement Operations.* Retrieved 2023-11-24, from https://www2.deloitte.com/us/en/blog/business-operations-room-blog/2023/generative-ai-in-procurement.html

Ramakrishnan, R. (2024, February). *Lecture 7: Transformers.* MIT Campus.

Ramos, L., & Chandrasekaran, A. (2023, August). Quick Answer: What Are the Pros and Cons of Open-Source Generative AI Models? *Gartner*, *ID G00799591* . Retrieved 2023-12-24, from https://www.gartner.com/en

Robuck, M. (2023, December). *Feature: How GenAI is transforming AT&T.* Retrieved 2023-12-22, from https://www.mobileworldlive.com/att/feature-how-genai-is-transforming-att/

Salihoglu, S. (2024). *RAG Using Structured Data: Overview & Important Questions.* Retrieved 2024-03-03, from https://kuzudb.com/docusaurus/blog/ llms-graphs-part-1/

Sallam, M. (2023, January). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, *11* (6), 887. Retrieved 2023-11-24, from https://www.mdpi.com/2227-9032/11/6/887 (Number: 6 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/healthcare11060887

Shahriar, S., & Hayawi, K. (2023). Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *Artificial Intelligence and Applications*. Retrieved 2023-11-24, from http://arxiv.org/abs/2302.13817 (arXiv:2302.13817 [cs]) doi: 10.47852/bonviewAIA3202939

Sirtori-Cortina, D., & Case, B. (2023, April). Walmart Is Using AI to Ne- gotiate the Best Price With Some Vendors. *Bloomberg.com*. Retrieved 2023-11-03, from https://www.bloomberg.com/news/articles/2023-04-26/ walmart-uses-pactum-ai-tools-to-handle-vendor-negotiations

Srivastava, N. (2023). *Procurement Made Easy with Open AI | LinkedIn.* Retrieved 2023-10-14, from https://www.linkedin.com/pulse/ procurement-made-easy-open-ai-nishant-srivastava/?utm_source= share&utm_medium=member_android&utm_campaign=share_via

Sullivan, F. . (2021, December). *Global Artificial Intelligence in Supply Chain Management Growth Opportunities.* Retrieved 2023-11-23, from https://research-frost-com.libproxy.mit.edu/assets/1/f0a502e4 -4463-11e8-b626-1aa9f74f20ad/7bd92c96-57de-11ec-b464-7ee464b354e3/

Sullivan, F. . (2023, May). *Beyond ChatGPT—Understanding the Impact of Large Language Model Driven Generative AI.* Retrieved 2023-11-18, from https://research-frost-com.libproxy.mit.edu

Tang, H., & Koleva, I. (2023, October). *Disrupt your industry with generative AI.* Retrieved 2023-12-22, from https://www.databricks.com/resources/ webinar/disrupt-your-industry-generative-ai

Topsakal, O., & Akinci, T. C. (2023, July). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *Interna- tional Conference on Applied Engineering and Natural Sciences*, *1* , 1050–1056. doi: 10.59287/icaens.1127

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved 2023-10-05, from https://proceedings.neurips.cc/paper_files/ paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vu, T. (2023). *Use case of Deep Learning in Procurement.*

Wiles, J., & Sallam, R. (2023). *Measuring the ROI of GenAI: Assessing Value and Cost.* Retrieved 2023-11-16, from https://www.gartner.com/en/articles/ take-this-view-to-assess-roi-for-generative-ai

Yse, D. L. (2023). *Your Guide to Knowledge Graphs.* Retrieved 2024-03- 03, from https://lopezyse.medium.com/your-guide-to-knowledge-graphs -509c847f3d69

Appendix A

Table A-1

Types of Questions

Type 1	• What is the spend for X subcategory in 2023 by quarter? Show the data for the top 3 vendors. • For categories X, Y, and Z, what are the good receipts spend by category? Show in a graph by category. Compare 2022 to 2023. • Who are the top vendors for each category? Are they different from 2022 to 2023? How much has the spend increased or reduced? • Between October 2022 and October 2023, what is the spend by vendor for Category X by subcategory? • What materials have we purchased in 2023 from a supplier with a manufacturing location in Hong Kong? How many in total per material? • Which FGs have the highest cost, and which have the lowest? Which RMs are the largest drivers of FG unit cost?
Type 2	• Which material had the largest increase in unit price over the last 5 years? • Which vendor offers the best prices for Material X? Assume price is the only metric. • We are expecting all shipments by Supplier A to be delayed. Based on their raw materials and the BOM, which FGs will be impacted by a delay in RMs, and how much of the total spend will be impacted? • Which materials are only manufactured in a single country or close geographic area? Which are manufactured in diverse locations? • What is the cost of purchasing the raw materials to make a set quantity of a finished good?
Type 3	• Based on market movement data, how are the prices for Material X trending? Are they increasing or decreasing? Is there seasonality in the price? • Can I reduce a factory's manufacturing capacity by 5% and still meet the demand? • Which materials are good candidates for a second supply source, due to supply risks? • Which vendors are relied upon for supply but have a track record of poor OTIF (on time in full)?