Optimizing Procurement Analytics with Generative AI and Automated Data Visualization

by

Shen Yeong Loo
Bachelor of Engineering in Mechanical Engineering, Nanyang Technological University

and

Mariana Dias Pennone
Bachelor of Engineering in Industrial Engineering, State University of Campinas

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 2025

Signature of Author: _____
Department of Supply Chain Management
May 9, 2025

Signature of Author: _____
Department of Supply Chain Management
May 9, 2025

Certified by: _____
Dr. Thomas Koch
Postdoctoral Associate
Capstone Advisor

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

Optimizing Procurement Analytics with Generative AI and Automated Data Visualization

by

Shen Yeong Loo

and

Mariana Dias Pennone

Submitted to the Program in Supply Chain Management
on May 9, 2025 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

The growing complexity of procurement operations at a leading pharmaceutical company has led to an overload of dashboards, increasing reporting inefficiencies and limiting data-driven decision-making. To address these challenges, our project explores how Generative AI (GenAI) and automated data visualization can optimize procurement analytics. We developed a proof-of-concept (PoC) chatbot that allows procurement managers to use natural language queries to generate dynamic, accurate visual insights without relying on traditional dashboards or advanced technical skills. Leveraging open-source tools, notably LIDA (a grammar-agnostic library that generates visuals and infographics), the system connects to structured procurement data, processes user queries through Large Language Models (LLMs), and returns contextualized visual outputs. The framework was tested across four high impact use cases, including vendor spend summarization and forecasting, achieving a 96% success rate in producing accurate and contextually appropriate outputs. Key outcomes include reduced dependency on the BI team, significant time savings, and enhanced decision-making efficiency. The project also outlines a scalable deployment strategy, emphasizing data governance, user training, and prompt engineering to mitigate challenges like jargon misinterpretation and dataset scalability. This approach offers a sustainable, cost-effective alternative to commercial solutions, empowering procurement teams to focus on strategic value creation while maintaining flexibility for future AI advancements.

Capstone Advisor: Dr. Thomas Koch
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

We are deeply grateful to our capstone advisor, Dr. Thomas Koch, for his thoughtful guidance, steady mentorship, and unwavering support throughout the course of this project. His expertise and encouragement were essential in shaping both our approach and the final outcome. We also extend our sincere thanks to our writing advisor, Ms. Toby Gooley, whose detailed feedback and constant support greatly enhanced the clarity and quality of our report.

We are immensely thankful to our sponsor company for entrusting us with a meaningful and timely challenge, and for providing the data, insights, and ongoing engagement that enriched our project and made it possible to translate ideas into impact.

To our friends and peers in the Supply Chain Management cohort, thank you for the encouragement, collaboration, and community that have been a source of motivation and joy during this journey.

We also wish to thank our families for their support and patience, which sustained us through the demands of this endeavor. Finally, we would like to acknowledge each other, Shen Yeong Loo and Mariana Dias Pennone, for our mutual dedication, complementary strengths, and shared commitment to making this project a success. This journey has been a testament to the power of collaboration, and we are grateful to all who contributed to it.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 MOTIVATION

Our sponsor company is a leading multinational pharmaceutical corporation with multi-billion-dollar annual revenue. Operating across pharmaceuticals and medical devices, the company plays a critical role in developing and delivering healthcare solutions globally. With a strong focus on innovation, it invests over $16 billion annually in research and development (R&D), ranking among the top three pharmaceutical companies in R&D spending.

As a global leader in the pharmaceutical industry, our sponsor allocates billions of dollars annually to supply chain activities, particularly procurement. Strategic sourcing initiatives are integral to optimizing total costs, mitigating supply chain risks, and ensuring uninterrupted supply continuity, thereby safeguarding revenue streams. These procurement strategies not only fortify the company's market position but also drive top-line growth while maintaining bottom-line efficiency.

To improve financial performance, company leadership needs to extract actionable supply chain and procurement insights to support strategic decision-making. However, the current approach relies on traditional methods such as dashboards and data tables to derive procurement insights, including identifying high-spend areas and anticipating inflation trends. This process is inefficient and time-consuming, limiting the procurement team's ability to make data-driven decisions effectively.

In this context, Artificial Intelligence (AI) and Generative AI (GenAI) offer a transformative opportunity to optimize spend and demand management. When implemented effectively, these technologies can enhance productivity for procurement associates and category management leaders, enabling faster and more accurate insights. For example, according to a 2024 McKinsey & Company report, a global pharmaceutical company reduced time spent on procurement tasks by 90% by leveraging GenAI for spend intelligence. Another organization achieved an average 10% cost reduction through AI-driven cost modeling, demonstrating the significant potential of AI in procurement optimization.

By leveraging GenAI's natural language features, the procurement team can handle data analysis and visualization in a simpler, more intuitive way. Instead of using complex visual tools or coding, users can use daily language to ask questions like "Show me the total historical spending trend in indirect procurement for last year," and receive back the desired data in a proper chart type to visualize it, like a line chart or a bar chart to illustrate this time-series data. The decision to answer questions with charts instead of text is based on their ability to make it easier to quickly identify patterns, trends, and outliers at a glance. They enable decision-makers to quickly grasp relationships between variables, such as supplier performance or cost comparisons, without sifting through dense spreadsheets or reports (FactWise, 2022).

Building on this, GenAI not only facilitates intuitive data querying but also enhances the overall analytical process through capabilities like data inference and augmentation, automatic chart generation, and contextual explanations of visualizations. These features empower non-expert users to interact with data more meaningfully, making complex insights more accessible and actionable (Yilin Ye, 2024).

The motivation behind this project is to explore ways to reduce the time spent by category and procurement managers on data analysis and extraction, ultimately improving operational efficiency and effectiveness.

1.2 PROBLEM STATEMENT

The sponsor's organization faces significant challenges in generating actionable insights from procurement data. The proliferation of over 200 dashboards complicates the user experience and increases the risk of erroneous reporting, ultimately hindering data-driven decision-making.

Considering the current scenario, the objective of our sponsor company is to explore and incorporate generative AI to automate data visualization and graph plotting in order to "democratize" insight generation, enabling associates to derive actionable insights more efficiently while comprehensively communicating the story behind the data. There are numerous opportunities in this area, which give rise to the following research questions:

1. **Reducing operational costs associated with dashboard maintenance:** Which GenAI model is the most effective in identifying and eliminating overlapping functionalities among the existing dashboards?
2. **Alleviating capacity constraints faced by the BI team:** How can Gen AI be used to improve user experience, and to provide self-service tools, enabling associates to perform their own analyses?
3. **Promoting a more agile, responsive procurement environment:** How can GenAI be scaled effectively to handle growing volumes of procurement data while ensuring the viability and reliability of automated insight generation?

1.3 SCOPE AND EXPECTED OUTCOMES

The objective of this project is to address the significant challenges faced by the sponsor company's procurement team in generating actionable insights from data. By leveraging AI and visual storytelling, we seek to democratize data analysis, reduce operational costs, and improve decision-making efficiency. The expected deliverables of this project are:

- Proof-of-concept (PoC) AI model that utilizes the company's proprietary data to automate and enhance insight generation, with testing for accuracy and sensitivity to prompts.

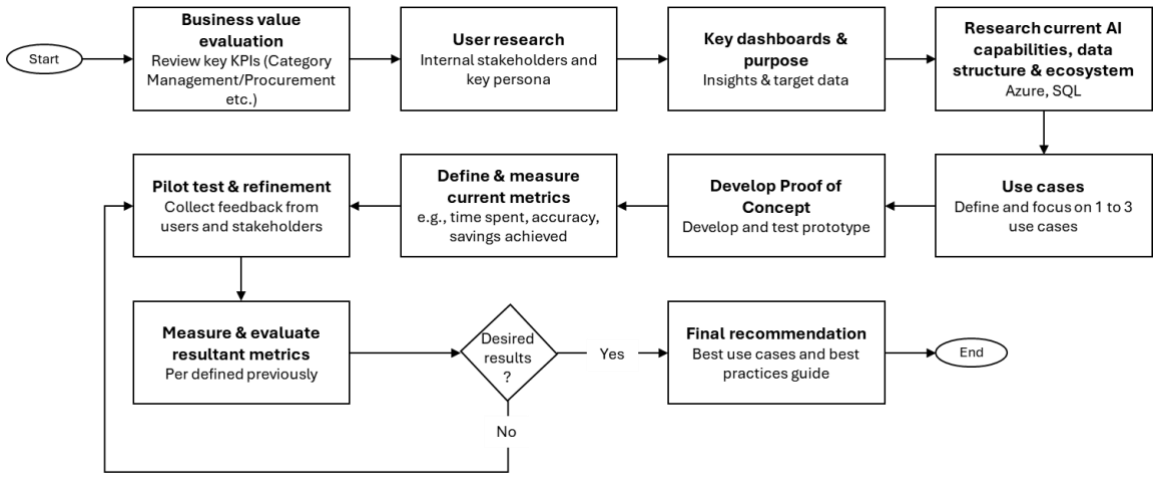
- Framework for automated chart plotting and data analysis that clearly communicates the insights derived from data.
- Measurement and evaluation of the impact of these capabilities on procurement operations, including time savings, cost reductions, and improved decision-making.
- User interface examples that enable associates to perform their own analyses with minimal training, reduced dependency on the BI team, and faster access to relevant insights.
- Clear definition of capabilities and limitations of the PoC proposed for the defined use cases.

1.4 PLAN OF WORK

Figure 1 represents the main steps carried out in this capstone. They include:

1. **Business Value Evaluation:** Conduct interviews with Sponsor Company's management to understand Key Performance Indicators for Procurement / Category. Scope the area of focus for utilization of GenAI to define use cases.
2. **User Research:** Conduct interviews and surveys to understand persona-based usage and data sources for existing dashboards. Gather insights on the types of inquiries directed to the BI team.
3. **Dashboard Utilization Analysis:** Analyze how users currently engage with existing dashboards to identify the conclusions they aim to draw.
4. **Selection of Pilot Scope:** Define four high-impact use cases to explore the applicability of Gen AI and storytelling capabilities.
5. **Proof-of-Concept Development:** Evaluate advanced algorithms and AI capabilities to dynamically generate actionable insights. Develop a prototype that incorporates visual storytelling elements to narrate the insights.
6. **Impact Measurement:** Design metrics to assess the effectiveness of the new system in addressing the outlined challenges. Collect feedback from users and stakeholders to refine the approach.
7. **Final Recommendation and Best Practices:** Deliver a final recommendation, outlining the best use cases and most impactful models. Create a best practice guide that explains how to extract insights effectively using Generative AI and integrate the system seamlessly with existing BI tools.

Figure 1: Plan of Work diagram



2 STATE OF THE PRACTICE

This section explores the technologies and tools currently used across industries to address our research questions on generating data visualization and business insights through Artificial Intelligence. An in-depth examination of AI, with a focus on generative AI for visualization, along with its strengths and limitations, provides a foundation for evaluating its applicability and potential benefits. These insights will guide the development of a proof of concept tailored to our sponsor company's needs.

2.1 ARTIFICIAL INTELLIGENCE AND BUSINESS IMPACT

The transformative potential of artificial intelligence (AI) has captured significant interest and secured strong support from leaders across industries. PwC's 2024 Global Annual Review highlights a collective investment of nearly $1 billion by its network of firms to enhance and scale AI capabilities, forge strategic partnerships, and deploy AI tools across service lines. Additionally, PwC's 2024 Global CEO Survey says that 89% of CEOs anticipate that AI will improve the quality of their products and services, reinforcing its critical role in driving business value (PwC, 2024).

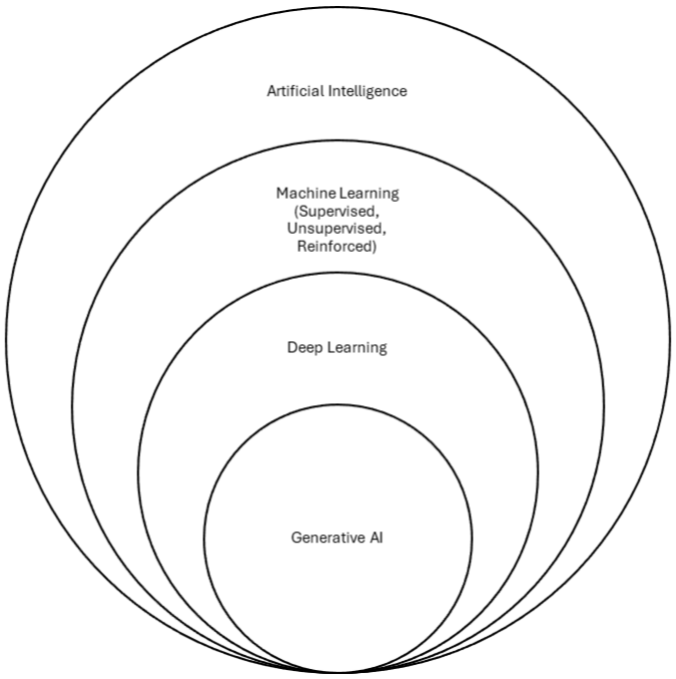
Within the broader field of AI, generative AI stands out as a game-changer. A 2023 McKinsey report estimates that generative AI alone could contribute between $2.6 trillion and $4.4 trillion annually to the global economy across 63 identified use cases. Among these, supply chain and operations emerge as high-potential areas, with an estimated value increase between $290 billion and $550 billion. As the fifth-largest area for AI-driven value creation, supply chain applications underscore how generative AI can transform business processes, improve efficiency, and boost revenue, with 70% of CEOs expecting generative AI to redefine value creation within their organizations (McKinsey, 2023).

2.2 ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) is the field focused on designing agents that can perceive their environment, make decisions, and take actions aimed at achieving specific objectives. AI systems are developed to mimic aspects of human intelligence or to operate rationally, employing techniques that range from rule-based algorithms to advanced machine learning models like neural networks (Russell & Norvig, 2020).

DeepLearning.AI (DeepLearning.AI, 2024) describes AI as a set of tools designed to assist humans in accomplishing tasks. At its core, general AI encompasses various approaches, including machine learning (ML), which involves training systems to recognize patterns and make decisions using supervised, unsupervised, or reinforcement learning techniques. Building on ML, deep learning employs neural networks modeled after the human brain to process vast amounts of data and uncover intricate patterns. Generative AI, a specialized branch of deep learning, focuses on creating new content—such as text, images, or data visualizations—by leveraging these advanced models. Together, these interconnected technologies form a continuum, with generative AI demonstrating the creative potential of broader AI systems. Figure 1 shows a simple representation of the relationship between Generative AI and AI.

Figure 2: GenAI as sub-discipline of AI



2.3 GENERATIVE ARTIFICIAL INTELLIGENCE (GenAI)

The technical backbone of Generative Artificial Intelligence (GenAI) lies in Large Language Models (LLMs), which leverage massive datasets and billions of parameters to generate human-like text, respond to

queries, and perform a variety of other tasks, including summarizing documents, translating languages, generating code, drafting emails, or even creating creative content like stories or poetry. A pivotal moment in the democratization of GenAI occurred with the release of ChatGPT in November 2022, which quickly made GenAI accessible to everyday users (Bengesi, et al., 2023). This conversational AI, capable of producing useful responses across a multitude of topics, has boosted efficiency and productivity, igniting widespread interest in its potential applications across education, science, and industry.

In parallel, OpenAI developed models like DALL-E (Ramesh, et al., 2021) and its successor, DALL-E 2 (Ramesh, et al., 2022) which integrate techniques from both natural language processing and image generation to create original images from text prompts.

These advancements underscore the versatility of GenAI, extending its reach beyond text-based interactions to visual content generation. To fully appreciate GenAI's impact and limitations, it is essential to understand its technical foundation and evolution. GenAI models have evolved over time, starting with simpler architectures and advancing to the sophisticated, parameter-rich models we see today. Key models in this space include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Transformers, and Diffusion Models, each contributing unique strengths to the field of generative modeling. For each model, we provide a summary along with typical use cases to illustrate its practical applications.

2.4 GENERATIVE ADVERSARIAL NETWORKS (GANs)

Introduced by Ian Goodfellow in 2014, Generative Adversarial Networks (GANs) represent a powerful and unique framework for generating synthetic data through the interplay of two neural networks: the generator and the discriminator. In this adversarial setup, the generator (G) creates artificial data samples (e.g., images, text) based on random noise, while the discriminator (D) evaluates these samples, determining whether each one is real (from the training set) or fake (produced by the generator). The generator's objective is to create increasingly realistic data to "fool" the discriminator, while the discriminator aims to correctly classify data as real or fake, thus creating a competition akin to a zero-sum game. This iterative process encourages the generator to improve its outputs until the discriminator can no longer reliably distinguish between real and synthetic data, effectively minimizing the discriminator's success and maximizing the generator's realism (Goodfellow, et al., 2014).

GANs are widely used across various applications, including image generation and editing, data augmentation, video and animation generation, text-to-image synthesis, super-resolution imaging, and music and audio generation.

2.5 VARIATIONAL AUTOENCODERS (VAEs)

Autoencoders (AEs) are unsupervised neural networks that consist of three main components: the encoder, latent representation, and decoder (Michelucci, 2022). The encoder transforms input data into a

lower-dimensional representation, also known as latent representation, or embedding, while the decoder reconstructs the data back to its original form, minimizing reconstruction error. The primary purpose of an autoencoder is to learn an efficient latent representation of the input data and use it to reconstruct the data, mimicking its original form as closely as possible (Bergmann & Stryker, 2024).

Variational Autoencoders (VAEs), a more advanced form of autoencoder, differ in how they represent latent variables. Instead of using a fixed, discrete representation, VAEs encode a continuous, probabilistic representation of the latent space. This approach enables VAEs not only to reconstruct the original input with high fidelity but also to generate new data samples that resemble the original data through variational inference (Bergmann & Stryker, 2024).

Autoencoders, including VAEs, are commonly applied in tasks such as dimensionality reduction, feature extraction, image denoising, image compression, image search, anomaly detection, and imputation of missing values.

2.6 DIFFUSION MODELS

Ho, Jain, & Abbeel (2020) introduce a class of generative models that create high-quality images by progressively denoising a noisy input. This approach has proved to be highly effective in tasks such as image synthesis and image-to-image translation, enabling the generation of remarkably realistic visuals.

The core principle of Denoising Diffusion Probabilistic Models (DDPMs) lies in reversing the process of adding noise to an image. Noise refers to random distortions introduced to an image, which progressively obscures its details until it becomes unrecognizable, resembling random static. During training, the model learns to iteratively denoise these noisy images, removing the noise step by step to restore them into detailed and realistic outputs. This process of denoising involves predicting and subtracting the noise added in each step, enabling the model to reconstruct intricate patterns and textures. This iterative refinement allows DDPMs to be highly versatile, excelling in tasks such as image synthesis and image-to-image translation (Ho, Jain, & Abbeel, 2020).

DDPMs have garnered significant attention for their ability to generate high-fidelity images and their adaptability across a wide range of applications. Notably, they have served as the foundation for numerous advancements in image generation, including models like DALL-E 2. By leveraging the iterative refinement of noisy inputs, DDPMs have pushed the boundaries of generative modeling, enabling the creation of highly detailed and lifelike images from textual descriptions.

2.7 TRANSFORMERS

The seminal paper “Attention Is All You Need” (Vaswani, Shazeer, & Parmar, 2017) introduced the Transformer architecture, a paradigm-shifting innovation in Natural Language Processing (NLP) and other sequence-based tasks. By eliminating the need for complex recurrent or convolutional neural networks, the

Transformer leverages a novel attention mechanism to efficiently process input and output sequences. A comprehensive understanding of its architecture is pivotal to appreciating its impact and utility in NLP research.

The Transformer architecture is specifically designed to process sequential data, such as sentences in natural language, and consists of two primary components: the encoder and the decoder. The encoder processes the input sequence, transforming it into a continuous representation through several stacked layers, each containing:

- **Self-Attention Mechanism:** This mechanism enables the model to examine relationships between all words in the input sequence. For instance, in the sentence “The cat sat on the mat,” the model identifies the relationship between “cat” and “sat.”
- **Feed-Forward Network (FFN):** Following the self-attention step, the feed-forward network applies computations to refine and enhance the representation of each word.

The decoder generates the output sequence by utilizing the representations created by the encoder. Like the encoder, it consists of several layers, incorporating:

- **Attention to Encoder Output:** This ensures that the decoder refers to the encoded input sequence while generating the output.
- **Self-Attention Mechanism:** Unlike the encoder, this mechanism is constrained to attend only to the words already generated, ensuring proper sequence in the output.

The Transformer's success is driven by key innovations, including Multi-Head Attention, which allows the model to process different aspects of the input simultaneously, capturing complex relationships within the sequence; and Positional Encoding, which compensates for the model’s lack of inherent order by embedding positional information, enabling it to understand sentence structure.

In short, the Transformer model’s attention-based architecture enables it to efficiently process and understand sequences of data, outperforming prior models in handling long sentences and complex relationships between words. Its versatility and effectiveness have made it a cornerstone in modern NLP, powering applications such as language translation and text generation.

The Transformer architecture has been the foundation for many subsequent advancements in natural language processing, including the development of large language models like the GPTs.

Table 1 shows a summary of the models, their use cases and applications. In our project, we would need to utilize a model which is capable of text-to-text, text-to-code, and text-to-image, and that leads us to the use of Transformers-based GenAI.

Table 1: GenAI models' comparison matrix

Feature	*GANs*	*VAEs*	*Diffusion Models*	*Transformers*
Architecture	Adversarial network (generator and discriminator)	Encoder-decoder architecture	Stochastic process of noising and denoising	Self-attention mechanism
Novelty/Features	Competitive learning, high-quality generation	Probabilistic framework, latent space representation	Novel training paradigm, gradual generation	Long-range dependencies, contextual understanding
Use Cases	Image generation, style transfer, data augmentation, super-resolution	Data generation, anomaly detection, dimensionality reduction, representation learning	Image and text generation, image-to-image translation, video generation	Natural language processing, machine translation, text generation
Application Mode	Image-to-image, text-to-image	Text-to-image, image-to-image	Text-to-image, image-to-image, text-to-text	Text-to-text, text-to-image, text-to-code
Applications	StyleGAN for realistic face generation, BigGAN for diverse image generation	VAE-GAN hybrids for improved image generation, VAEs for anomaly detection in medical images	DALL-E 2 for text-to-image generation, stable diffusion for high-quality image generation	GPT-4 for advanced language understanding and generation, Gemini for language modeling

2.8 GENERATIVE PRE-TRAINED TRANSFORMER

GPT-based language models can process and generate different types of content, handling tasks such as summarizing documents or translating languages (i.e. text-to-text capabilities), creating visuals based on written descriptions (i.e. text-to-image capabilities), and generation or debugging code (i.e. text-to-code capabilities). These models are applied in various areas, including simplifying complex information, automating customer support, and assisting with creative work like writing stories or designing prototypes.

A notable characteristic of these models is their ability to perform a wide range of tasks without needing extensive adjustments for each specific use. This flexibility has made them significant in ongoing AI research and practical applications (Radford et al., 2019; OpenAI, 2020).

Leveraging these features, Generative AI, using models like GPT, changes how we visualize data by offering several useful features. These models can generate charts and visuals based on what the user asks

for, pull out different insights from the data, and make exploring the data straightforward. This makes analyzing data simpler and more efficient, especially for people who aren't experts in data analysis. Overall, this flexibility helps turn data into useful information, making it easier to make decisions based on the analysis. However, models like GPT are prone to hallucinations—generating content that seems convincing but is factually inaccurate—due to their reliance on patterns learned during training, which may not always reflect current or complete knowledge.

To address this limitation, Retrieval-Augmented Generation (RAG) offers a solution by blending a pre-trained language model, which generates text based on learned patterns, with an external knowledge source, like a searchable Wikipedia database. Introduced by Lewis et al. (2021), RAG enhances text generation by retrieving relevant, up-to-date information based on a user's question or prompt and incorporating it to produce more accurate and detailed responses, effectively merging the strengths of text creation and information lookup. By using external information to improve its outputs, RAG minimizes hallucinations, addressing the flaws of GPT models that rely only on their potentially incomplete or outdated internal knowledge (IT Convergence, 2024).

Continuing from above, we will dive into related works and explore the current solutions available for data visualization using Generative AI in both commercial domains (Section 2.9) and open-source domains (Section 2.10). This analysis will provide insights into existing frameworks and tools, highlighting their capabilities, limitations, and relevance to the objectives of this project.

2.9 COMMERCIAL TOOLS FOR DATA VISUALIZATION

To enable data visualization from natural language prompts, researchers have identified and addressed several critical challenges. These challenges have driven advancements in this field, leading to the development of tools, intelligent agents, and integrated solutions that bridge the gap between natural language inputs and effective data visualization.

In the commercial space, a range of tools has emerged to facilitate data exploration and visualization for deriving actionable insights. For example, our sponsor company currently uses Tableau to visualize indirect procurement spend. Tableau AI (Tableau Resource, n.d.) offers a GenAI-powered data analysis feature as a premium upgrade for its users. Other tools, such as ThoughtSpot and AutoInsights, provide similar capabilities for interactive data exploration and automated insights (ThoughtSpot Product, n.d.; Alteryx Products, n.d.).

However, our project emphasizes non-commercial, open-source solutions. While commercial tools are referenced for context, this document will not delve into their specifics but will instead focus on exploring and evaluating open-source alternatives.

2.10 OPEN-SOURCE TOOLS FOR DATA VISUALIZATION

In this section we will discuss notable open-source toolkits that provide the capabilities required to achieve our objectives.

2.10.1 NL4DV

Mitra et al. (2020) introduced a Python-based toolkit named Natural Language for Data visualization (NL4DV) designed to bridge the capability gap in natural language interfaces for data visualization. Before NL4DV, developers faced numerous challenges when building such interfaces, including:

1. **Complexity of NLP Implementation:** Developing low-level natural language processing techniques from scratch was both time-intensive and technically demanding.
2. **Ambiguity in User Queries:** Natural language queries often suffer from ambiguity and under-specification, making it difficult to extract clear intents.
3. **Need for Domain Expertise:** Effective visualization design and visual analytic tasks required deep domain knowledge, often limiting accessibility for non-experts.
4. **Balancing Query Flexibility with Usability:** Systems had to reconcile the flexibility of natural language queries with the need for clear, intuitive system responses.
5. **Integration Challenges:** Incorporating natural language inputs into existing visualization pipelines posed significant technical hurdles.

NL4DV addresses these issues by translating natural language queries into recommended specifications for data visualization. This enables developers to generate visualizations based on datasets by using libraries such as D3 or Vega-Embed to render and display the outputs across different platforms. By automating key processes, NL4DV significantly reduces the complexity of creating natural language-driven visualizations.

While NL4DV is a notable advancement, its implementation remains technical, requiring skilled developers with expertise in software engineering to integrate it into a fully packaged solution. Despite this limitation, its capabilities make it a promising tool for projects like ours, where generating visuals from natural language input is a core objective.

2.10.2 CHAT2VIS

In their paper on Chat2VIS, Maddigan and Susnjak (2023) introduce an innovative end-to-end system designed to translate free-form natural language (NL) queries into data visualizations. Chat2VIS harnesses state-of-the-art large language models (LLMs), including ChatGPT, GPT-3, and Codex, to address the challenges of accurately converting NL inputs into meaningful visual outputs. The authors emphasize the importance of prompt engineering in enhancing the performance of LLMs, particularly in managing ambiguous

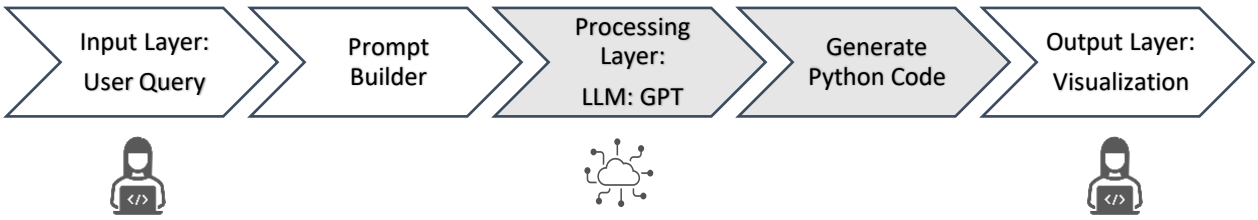
or imprecise user queries. By leveraging these advancements, Chat2VIS provides a reliable, efficient, and secure solution for generating visualizations while maintaining data privacy.

As a packaged solution, Chat2VIS integrates NLP to interpret user queries and generate customized visual outputs. The use of prompt engineering ensures that the graphs and charts produced are both accurate and relevant to the user's intent. Additionally, Maddigan and Susnjak (2023) evaluate Chat2VIS across various LLM models, demonstrating its adaptability and effectiveness in generating appropriate visualizations across different scenarios.

As illustrated in Figure 3, Chat2VIS architecture is designed to streamline the process of generating visualizations from natural language queries and is structured into the following layers:

1. **Input Layer:** Accepts free-form natural language queries from users, accommodating a wide range of complexities and clarity levels.
2. **Processing Layer:** Employs LLMs to interpret the queries, extract user intent, and generate the necessary code for visualizations. This layer is key to handling ambiguities and ensuring that user needs are accurately addressed.
3. **Output Layer:** Produces data visualizations tailored to the processed queries. The system automatically determines the appropriate chart types based on the data and the nature of the query, minimizing the need for manual adjustments.
4. **User Interaction:** Enables users to upload datasets and interact with the system for visualization generation. This functionality makes Chat2VIS accessible for experimentation and allows users to test various query scenarios.

Figure 3: Chat2VIS Process Flow, adapted from (Maddigan & Susnjak, 2023)



Overall, Chat2VIS represents a significant step forward in bridging the gap between natural language and data visualization. Its architecture prioritizes efficiency, user accessibility, and data security, making it a versatile tool for NL-driven visualization tasks.

In our trial run, we observed that Chat2VIS delivers an end-to-end solution for converting queries into visualizations. However, the system lacks conversational interactivity with the generated charts. This limitation poses a challenge for our project, as our objective extends beyond visualization to include interactive

exploration and insight extraction from procurement data. Consequently, while Chat2VIS offers a strong foundational framework, additional capabilities would be required to fully align with the scope of our project.

2.10.3 LANGCHAIN

Another approach to our project involves developing a task-specific tool to transform natural language queries into data visualizations. This can be achieved by creating an AI agent tailored to our objectives—capable of interpreting user queries, performing subtasks such as data retrieval and transformation, and generating visual outputs.

LangChain, a framework for building LLM-powered applications, offers modular tools to construct such agents efficiently (LangChain, 2024). According to LangChain's 2024 survey, the top use case for AI agents was research and summarization (58.2%). Other relevant use cases include code generation (35.5%) and data transformation and enrichment (33.8%), which align closely with our project goals.

While this approach provides greater flexibility and customization, it demands significant time and technical expertise in software engineering to implement effectively.

2.10.4 LIDA

LIDA (a novel tool for generating grammar-agnostic visua**LI**zations an**D** infogr**A**phics) is an innovative text-to-visual tool, implemented as a Python library, that is designed to simplify the creation of visualizations and infographics from natural language queries. Developed by Victor Dibia and his team, LIDA leverages advanced NLP techniques to accurately interpret user intent and generate customized, high-quality visual representations (Dibia, 2023). The tool has four key functional features:

1. **Summarizer:** Extracts and highlights key insights and trends by analyzing input data and generating concise summaries.
2. **Goal Explorer:** Assists users in refining their visualization objectives by suggesting relevant visualizations and data transformations.
3. **Viz Generator:** Produces a variety of visualizations, such as bar charts, line charts, and scatter plots, tailored to the user's query.
4. **Infographer:** Creates visually engaging infographics that present complex information in a clear and impactful manner.

By automating the visualization process, LIDA enables users to extract meaningful insights from their data with greater efficiency and ease, making it a valuable tool for data-driven decision-making.

In summary, LIDA's purpose and functionality align closely with our objectives of generating visualized insights from natural language inputs. Recognizing its potential, we decided to further evaluate LIDA by accessing its source code to test its capabilities and identify any limitations. This hands-on exploration will

allow us to better understand how effectively LIDA can support our goals and contribute to advancing user-centric data visualization. As part of our project solution, we emphasize key functionalities such as the Summarizer and Viz Generator to enhance data insight generation.

2.11 SUMMARY OF TOOLS

Table 2 summarizes the tools discussed in the previous sections. Based on our evaluation, our team has chosen to proceed with building our tool using the LIDA library in Python.

Table 2: Open source and commercial tools comparison

Feature	***NL4DV***	***Chat2Vis***	***LangChain***	***LIDA***	***Tableau +AI***	***ThoughtSpot/ AutoInsights***
Core functionality	Natural language to visualization specifications	Direct code generation for visualization	Framework for building LLM applications	Language-driven interactive data analysis	Business intelligence and data visualization platform with AI capabilities	Business intelligence and data visualization platform with AI capabilities
Customization	Medium	Medium	High	Medium to high	Medium to high	High
Ease of use (for end-user)	Medium	Medium	Medium to high	Medium to high	High	Low to medium
Technical expertise required	High	Medium	High	Medium	Medium	Medium
Visualization capabilities	Depends on integration	High	Depends on integration	High	High	High
Collaboration and sharing	Depends on integration	Depends on integration	Depends on integration	Depends on integration	Strong collaboration features	Strong collaboration features
Cost	Open source	Open source	Open source	Open source	Commercial	Commercial
Additional features	Focus on visualization specifications	Simple, direct interaction	Versatile framework for various LLM applications	Interactive exploration , real time updates	AI-powered insights, automated data preparation, natural language queries	AI-powered insights, automated data preparation, natural language queries, advanced analytics

3 METHODOLOGY

This section outlines the technical approach used to design and evaluate the generative AI solution for procurement analytics. The methodology is designed to demonstrate how Large Language Models (LLMs), when integrated with open-source visualization libraries, can transform natural language queries into actionable and visually intuitive outputs. Our goal is to operationalize the project's objectives—reducing reliance on static dashboards, enabling self-service analytics for procurement managers, and enhancing the speed and accuracy of insight generation.

Drawing from the tools and frameworks reviewed in the State of the Practice, we selected OpenAI's GPT models and the LIDA library as the foundation of our implementation. LIDA stood out for its ability to generate diverse chart types and infographics based on natural language prompts, while also allowing prompt customization and workflow integration. This makes it a strong fit for business environments where users require flexible, accurate visualizations.

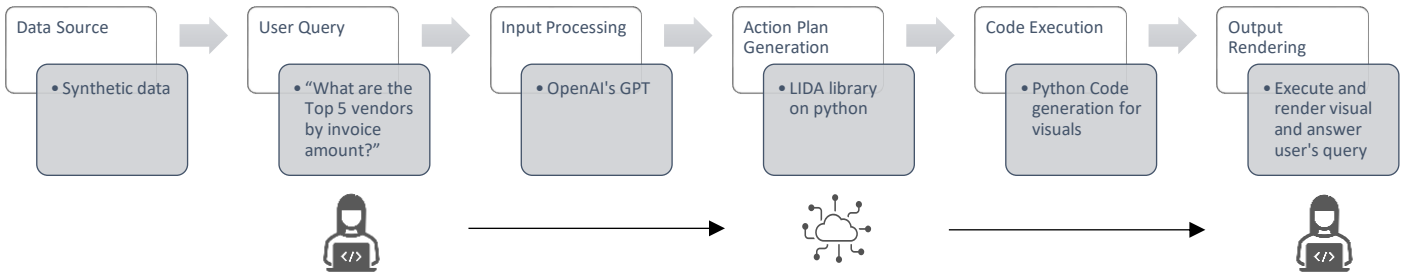
We also evaluated other tools such as NL4DV, Chat2VIS, and LangChain. While these offer valuable features, they presented either technical limitations (e.g., lack of conversational interactivity in Chat2VIS) or higher implementation complexity (as with LangChain). In contrast, LIDA's modular design, combined with the reasoning capabilities of LLMs, offered the best balance of performance, usability, and adaptability to the sponsor company's procurement use cases.

To operationalize this integration, our framework chains together the LLM and the LIDA library into a seamless pipeline that executes the following steps:

1. **Data Source:** We connect to synthetic offline data that was designed to preserve the structure, magnitude, and behavioral characteristics of actual procurement data provided by our sponsor company. This approach allows us to validate our framework while upholding data privacy, ethics, and compliance standards, while reinforcing the applicability of our results to real-world scenarios. To process this data, we utilize the Pandas library in Python to read the data file. The use of offline data ensures efficiency and safety, minimizing the risk of operational interruptions during the proof-of-concept development.
2. **User Query:** The system accepts natural language inputs from the user, such as the question: "What are the Top 5 vendors by invoice amount?" The Python script captures the user's query for further processing.
3. **Input Processing:** OpenAI's GPT is used to process the query and interpret the user's intent. The LLM is prompted to assume the role of a supply chain manager with expertise in data analysis who is tasked with providing procurement insights. To assist with query interpretation, we utilize LIDA's library.

4. **Action Plan Generation:** Upon understanding the query, the LLM generates a step-by-step plan to perform the necessary data analysis and determine the appropriate visual representation. For example: “Step 1: Sort the aggregate invoice amount by supplier name (LLM understands the interchangeability of ‘vendor’ and ‘supplier’). Step 2: Identify the top 5 vendors by invoice amount. Step 3: Generate Python code to produce the visualization.”
5. **Code Execution:** The Python script runs the LIDA tool to execute the generated code and create the corresponding visualization.
6. **Output Rendering:** The resulting visualizations, such as charts or tables, are displayed to the user in an interpretable format. To enhance the user experience, the system always includes the user’s original question as the title of the visualization, simulating a conversational Q&A interaction with the GenAI-powered tool.

Figure 4: Pipeline / Workflow from User Query to Output Generation



This pipeline, illustrated in Figure 4, bridges the gap between the LLM’s language-generation capabilities and the need for actionable, user-friendly outputs, aligning with the project’s broader goal of democratizing procurement insights. By automating this process, we aim to reduce the technical expertise required to analyze and interpret complex datasets, making data-driven decision-making more accessible across the organization.

3.1 DEFINING USE CASES

To guide the development and application of our framework, we conducted interviews with key users of procurement dashboards, specifically two procurement directors representing different product categories. These directors, frequent users of the dashboards that are relevant to our project, shared valuable insights and unique pain points encountered in their data analysis processes.

From these discussions, three primary use cases were identified, directly targeting the sponsoring company’s most pressing procurement challenges. These use cases are designed to deliver significant benefits by improving efficiency and enabling more strategic decision-making. Additionally, a fourth use case

was added to incorporate greater interactivity with GenAI, further enriching the user experience and supporting more informed decision-making. The use cases are as follows:

1. **Summarize procurement spend by vendor type and commodity.** In this use case, the framework enables users to generate visualizations that break down procurement spend data by vendor type and commodity. Through natural language prompts, users can request summaries that highlight key trends and patterns, such as which commodities drive the highest costs or which vendor categories dominate spending. The resulting visualizations provide actionable insights, enabling procurement teams to identify cost-saving opportunities and optimize supplier relationships.
2. **Summarize procurement spend with visualizations to support strategy development and spend forecasting.** This use case focuses on leveraging spend data visualizations to support long-term strategic planning and forecasting. By summarizing historical and current spend patterns, the framework helps users uncover trends and anomalies, informing budget allocations and procurement strategies. Visual outputs such as time-series charts, comparative tables, or heatmaps facilitate a deeper understanding of spend dynamics, allowing teams to forecast future expenditures with greater precision and confidence.
3. **Analyze procured items by interpreting keywords in purchase order headers using GenAI.** This use case explores how GenAI can be used to interpret and analyze Purchase Order (PO) headers by extracting relevant keywords. GenAI processes the unstructured text in PO headers and classifies them into meaningful categories, such as product types, suppliers, or procurement categories. This approach enables automated tagging and categorization, facilitating better spend analysis and decision-making.
4. **Interact with visual charts.** This use case enables users to interact with dynamic visualizations to gain deeper insights into procurement data. By allowing users to query, filter, or drill down into visual charts (e.g., bar charts, pie charts, scatter plots), this functionality enhances the interactivity of the dashboard. Users can tailor the data views to their specific needs, enabling more personalized analysis and facilitating faster, more informed decision-making.

Table 3 summarizes each use case and provides examples of relevant questions and instructions. By addressing these use cases, the framework not only demonstrates its utility in simplifying complex data analysis tasks but also automates repetitive analyses that analysts currently perform manually every month. This automation reduces the time and effort required for routine tasks, allowing procurement teams to focus on higher-value activities. Through intuitive, AI-driven visualization tools, our project aspires to empower stakeholders with insights that are both accessible and actionable.

Table 3: Defined use cases

No.	*Use Case*	*AI's Capability*	*Examples of Questions/Instructions*
1	Summarize procurement spend by vendor type and commodity	Generate visual in answering question	Top category by invoice amount?
2	Summarize procurement spend with visualizations to support strategy development and spend forecasting.	Generate simple projection and answer question with visual	What are the top 5 vendors by invoice amount? What is the spend split by priority flag? Average monthly PO amount for 2024, use average to forecast and fill up to month 12.
3	Analyze procured items by interpreting keywords in PO headers	Generate text summary	What items are procured through vendor X? What items are supplied most by vendor Y?
4	Interact with visual charts	Edit chart based on natural language input	Change chart type from bar to line chart.

4 RESULTS AND DISCUSSION

This section presents the outcomes of the GenAI chatbot and discusses its applicability across procurement use cases. We first evaluate the chatbot's performance by comparing the visual insights it generates—such as charts and summaries—with the expected results calculated directly from the source database or the synthetic data used during development. This comparison serves as the basis for assessing the chatbot's accuracy and reliability.

Next, we explore how our sponsor organization could extend this prototype into a production-ready solution, including potential scalability, integration pathways, and support requirements. Finally, we discuss the current limitations of using LLM-powered chatbots for insight extraction, along with recommended strategies to mitigate or navigate these challenges in future deployments.

To evaluate the chatbot's performance in realistic use cases, we conducted interviews with two category managers from the procurement group. Based on their feedback, we designed a set of prompts that simulate typical day-to-day operational queries. To assess our chatbot performance, we evaluated its outputs in 5 criteria, listed and explained in Table 4.

Table 4: Criteria used to assess chatbot performance

No.	*Example Prompt*	*Expected Chart Type*	*Data Accuracy Requirement*	*Required Chart Elements*
1	Top category by invoice amount?	Bar chart (vertical)	Aggregate PO invoice amounts per category	Title, legend (if multiple categories), optional average line
2	Top 5 vendors by invoice amount for 2024?	Bar chart (vertical or horizontal)	Aggregate invoice amounts per supplier, filtered for 2024	Axis labels, vendor names, ranked order
3	What is the spend split by preferred supplier flag?	Bar chart (vertical)	Aggregate PO spend by preferred_supplier_flag	Labels for binary flags, clear title
4	What is the spend split by diversity flag?	Bar chart (vertical)	Aggregate PO spend by diversity_flag	Distinct labels per group, title, legend
5	Average monthly PO amount for 2024, use average to forecast up to month 12	Line or dot chart with trendline	Monthly PO spend aggregation; accurate averaging	X-axis = months, trendline + forecast annotation

To evaluate the chatbot's outputs, we compared the generated visualizations against expected results, which we define in terms of (1) chart format suitability—using bar charts for categorical comparisons, and line charts for time series and trends; (2) data accuracy—ensuring that the figures shown reflect correct aggregations from the underlying dataset; and (3) visual clarity—including accurate labels, legends, and titles that help users interpret the chart without ambiguity. Table 5 summarizes the evaluation across five independent runs per prompt, assessing the consistency of these dimensions.

Table 5: Evaluation of model output accuracy across 5 runs per prompt

No.	*Example Prompt*	*Chart Type Accuracy*	*Data Accuracy*	*Labels & Legend*	*Comments*
1	Top category by invoice amount?	Yes (5/5)	Yes (5/5)	Yes (5/5)	—
2	Top 5 vendors by invoice amount for 2024?	Yes (5/5)	Yes (5/5)	Yes (5/5)	—
3	What is the spend split by preferred supplier flag?	Yes (5/5)	Yes (5/5)	Yes (5/5)	—
4	What is the spend split by diversity flag?	Yes (5/5)	Yes (5/5)	Yes (5/5)	—
5	Average monthly PO amount for 2024 with forecast	Yes (4/5)	Yes (5/5)	Yes (5/5)	One instance failed: 'PO amount' was misinterpreted, causing incorrect field use and no chart output.

Based on the testing, the chatbot demonstrated strong performance in both numerical accuracy and chart selection. It successfully produced accurate charts and correct data aggregations in nearly all tested cases, achieving a 96% success rate across five representative prompts. The only observed issue involved one instance of data misinterpretation during a forecasting task, where the LLM incorrectly processed the field used for PO amount aggregation, resulting in a missing chart. Despite this isolated case, all other outputs consistently met expectations in terms of chart type, data accuracy, and visual clarity.

This is a promising result for corporate users who may still have reservations about entrusting data analysis tasks to AI — in this case, a large language model (LLM)-powered chatbot.

This level of accuracy is expected, as our solution incorporates a structured dataset that serves as a reference for the LLM's responses. This approach aligns with the Retrieval-Augmented Generation (RAG) technique, a best practice in GenAI development that mitigates hallucination by grounding the model's answers in contextually relevant data.

The use of an advanced LLM such as OpenAI's ChatGPT — known for its extensive training on Python code and data-related tasks — further enhances the reliability of code generation, particularly for creating accurate visual outputs such as charts. Comparable LLMs from other providers, such as Google's Gemini and Microsoft's Copilot, offer similar capabilities and represent potential alternatives for enterprise deployment.

To improve contextual understanding and output consistency, we also applied prompt engineering. This includes injecting predefined instructions into the chatbot's prompt to guide its responses in alignment with procurement use cases. For instance, when asked about spend trends over time, the chatbot is explicitly instructed to return a line chart rather than a bar chart — as line charts better reflect temporal patterns or anomalies.

The snapshots in Figures 5, 6, and 7 illustrate several prompt engineering strategies embedded in our implementation. These prompts are coded into one of LIDA's core components, the vizgenerator, which leverages the LLM's ability to generate Python scripts in response to user queries. To ensure consistency in chart format, interpretability of results, and contextual relevance, the prompts were carefully designed to guide the LLM's behavior.

For instance, Figure 5 presents a sample prompt that encodes visualization best practices, such as prioritizing line charts for time-series questions. This design decision ensures that users receive visuals aligned with the nature of their query — such as trend analysis over a given period (e.g., "What is the purchasing volume and amount trend in 2023?"). The prompt not only improves the relevance of the output but also enhances its impact by encouraging the most suitable chart type for temporal insights.

Figure 5: Prompt engineering for visualization best practices

```
For visualization best practices:
1. Axis Orientation:
   - For bar charts: Categories/names on x-axis, values on y-axis
   - For time series: Time on x-axis, values on y-axis
   - For scatter plots: Independent variable on x-axis, dependent on y-axis
   - Always rotate long x-axis labels (plt.xticks(rotation=45, ha='right'))
   - if action plan includes date data, please consider it as a time series chart and use line chart.
```

Similarly, Figure 6 showcases a prompt designed to enforce proper data preprocessing before any code is generated. This is particularly critical in handling date-type fields, which often contain inconsistent formats such as Year-Month, Month-Day-Year, or Day-Month-Year. Without normalization, these inconsistencies can lead to code failure due to incompatible datatypes. The prompt instructs the model to standardize all date fields using 'pd.to_datetime()' — a generic yet robust directive that ensures smoother execution and improves data handling reliability. Additionally, the prompt implicitly guides the LLM to identify which date columns are most relevant for time-based filtering or aggregation, thereby improving the contextual quality of the insights produced.

Figure 6: Prompt engineering for robust data preprocessing

```
2. Data Preprocessing:
   - Handle missing values appropriately
   - Convert date strings to datetime using pd.to_datetime()
   - Group data by appropriate categories or time periods
   - Sort data in a meaningful order (e.g., by value for rankings)
   - If data is in dollar or header suggests so (e.g. PO amount), please convert data in appropriate
   units.
   - For PO data:
     * Convert dates: data['Day_of_PO_create_date'] = pd.to_datetime(data['Day_of_PO_create_date'])
     * Remove invalid dates: data = data[pd.notna(data['Day_of_PO_create_date'])]
     * Filter by year: data_2024 = data[data['Day_of_PO_create_date'].dt.year == 2024]
     * Group by month: data.groupby('Period_spend_month')['PO_amount__USD_'].sum()
     * Add mean line: plt.axhline(mean, linestyle='--', label=f'Mean: {mean:,.2f}')
```

In Figure 7, we highlight modifications made to another component of LIDA called the scaffold, which serves as a Python code template for structuring the visualization output. Unlike the vizgenerator, which focuses on dynamic content generation, the scaffold enforces consistency in code formatting and output structure, particularly to accommodate syntax-specific requirements of the visualization library used—in our case, Matplotlib. During a review session with the sponsor's management team, a key feedback point emerged: the need for data traceability in GenAI's visual outputs. Management emphasized that including source attribution in the generated charts would increase user confidence in the accuracy and validity of the insights. Responding to this, we embedded a data source "watermark" within the scaffold template. As shown in Figure 7, this includes code that explicitly sets the watermark text, its positioning, and color. The watermark references "PES," a term used internally by the sponsor to denote a primary data source during the proof of

concept. This enhancement ensures that every chart generated by the GenAI chatbot is not only visually informative but also transparently traceable to its underlying dataset.

Figure 7: Scaffold template with source watermarking for data traceability in visual outputs.

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
<imports>
# plan -
def plot(data: pd.DataFrame):
    <stub> # only modify this section
    plt.style.use('default')  # Reset to default style
    plt.rcParams['axes.prop_cycle'] = plt.cycler(color=['#000080', '#00008B', '#000000'])  # Set default
color cycle
    plt.title('{goal.question}', wrap=True)
    plt.text(
    0.01, 0.98,
    'datasource: PES',
    fontsize=8,
    color='gray',
    ha='left',
    va='top',
    transform=plt.gca().transAxes
)
    return plt;
```

4.1 KEY LEARNINGS

Throughout the development of the chatbot for the sponsor company, we engaged in several discussions and meetings with key stakeholders to understand pain points and explore potential solutions. In the early phase, much of our conversation focused on previous digital initiatives—particularly why some solutions failed to reach full-scale deployment.

One of the first things we learned was that most users are subject-matter experts in procurement but possess limited experience in data analysis or programming. This insight led us to pivot toward a solution that improves the efficiency of procurement managers by delivering insights directly to them, augmented by AI. Generative AI is well-suited for this purpose, as it can translate natural-language queries into data visualizations by orchestrating the entire pipeline—aggregating relevant data, writing Python code, and rendering charts.

A critical decision point in the project was choosing between building on commercial AI platforms or leveraging open-source tools to create a custom solution tailored to the sponsor company's needs. While commercial platforms offer pre-integrated features and faster onboarding, they come with long-term drawbacks, such as vendor lock-in, unpredictable upgrade costs, and migration complexity. Major cloud providers—such as Microsoft (Copilot), Google (Gemini), and Amazon (Bedrock)—are actively rolling out GenAI-enabled services integrated with their databases and cloud ecosystems, which are attractive for enterprise adoption.

However, through this project, we learned that by utilizing open-source tools like Language Interface for Data Analysis (LIDA), the complexity of building and maintaining a generative AI chatbot remains well within the capabilities of the sponsor company's current data team. This approach reduces dependency on external providers while allowing full customization and control over the solution.

In short, the chatbot developed in our proof of concept (POC) bridges the gap between the data team and business users. It enables more effective collaboration by allowing end users—who are most familiar with the data and its business context—to directly interact with insights. This enhances both the relevance and reliability of outputs and positions the company for a smoother transition into wider AI adoption.

4.2 LIMITATIONS AND MITIGATION STRATEGIES

While the proof-of-concept (PoC) demonstrated the feasibility and value of integrating Generative AI into procurement analytics, it was intentionally developed and tested using synthetic data in a controlled local environment. This approach ensured data privacy and enabled rapid prototyping but also require a comprehensive evaluation of system behavior at scale in real-world conditions moving forward. As a result, further testing with actual procurement datasets and deployment in a production environment will be essential to assess performance under realistic operational conditions, including integration with live databases, user authentication, token limit constraints, and enterprise network constraints. The limitations identified during the PoC phase, along with proposed mitigation strategies for future development, are outlined in Sections 4.2.1 to 4.2.4.

4.2.1 USER PROMPTS WITH JARGON OR INTERNAL TERMINOLOGY

Users often rely on internal jargon or abbreviations (e.g., "CCSL", "div PO amt") that the LLM may not interpret correctly. This can lead to misaligned or incomplete responses.

Suggested Mitigation: Maintain a catalog of standardized prompts for end users to reference. In parallel, create and continuously update a prompt engineering library for the maintenance team, enabling targeted refinements to improve accuracy and reliability. Encourage validation of generated insights before use in decision-making.

4.2.2 INCONSISTENT CHART TYPE SELECTION

The chatbot may return inconsistent or suboptimal visualizations, such as using a bar chart for a time-series trend.

Suggested Mitigation: Use prompt engineering rules to guide chart selection and periodically review visualization logic with the procurement team. This ensures alignment with business expectations and improves the interpretability of results.

4.2.3 SCALABILITY AND PERFORMANCE BOTTLENECKS

LIDA leverages Pandas DataFrames for data handling, which may lead to performance issues when processing datasets exceeding 1 million rows. This can result in slower response times or system instability.

Suggested Mitigation: Scope the dataset appropriately before loading it into the chatbot interface. Pre-aggregate or filter data based on user queries to reduce memory load and improve response speed.

4.2.4 TOKEN LIMIT CONSTRAINTS OF LLMS

Large datasets or wide tables (with many columns) may exceed the token limits of the LLM, leading to truncated prompts or incomplete responses.

Suggested Mitigation: Restrict use cases to well-defined tables within operational token limits. For reference, most LLM APIs (e.g., OpenAI's GPT-4) have token limits ranging from 8,000 to 32,000 tokens per interaction, depending on the model variant. When dealing with large databases, consider using data summaries or multi-turn query decomposition to stay within processing limits.

5 CONCLUSION

This project addresses the sponsor company's challenge of managing the growing number of procurement dashboards, each developed to meet specific user needs. This proliferation creates maintenance overhead and strains the Business Intelligence (BI) team's capacity.

To tackle this, we developed a chatbot powered by Generative AI (LLM-based), enabling users to query procurement data using natural language and receive insights in the form of dynamic visualizations. This solution reduces the need for customized dashboards and enhances self-service analytics, empowering category managers and procurement analysts to obtain relevant insights more efficiently.

Our model is built to work with structured data and can connect flexibly to different data sources. This adaptability not only minimizes reliance on the BI team for routine reporting but also improves user experience by delivering instant, visual responses to business questions.

Using open-source tools such as LIDA for chart generation and Streamlit for the chatbot interface, we demonstrated that an in-house GenAI solution is both feasible and sustainable. The system's complexity is well within the capabilities of the sponsor company's existing BI and data teams. Compared to commercial solutions, our approach offers significant cost savings and avoids vendor lock-in — offering greater flexibility if the company chooses to change its data platforms in the future.

As we progressed with the proof-of-concept (POC), we also discussed scale-up considerations with the sponsor company to support broader deployment in a production environment. A pilot testing phase is recommended as a transitional step to validate real-world performance, assess user adoption, and fine-tune prompts or workflows prior to full-scale rollout. Key considerations include:

- **Generative AI governance:** Establishing policies and safeguards to ensure responsible use of LLMs, including privacy controls, prompt filtering, and oversight of model behavior. This is critical to maintain ethical standards, protect user data, and prevent misuse of the technology.
- **Data governance:** Ensuring that only approved datasets are exposed to the model. This prevents unauthorized or sensitive information from influencing outputs, safeguarding data integrity and compliance.
- **Access control and auditability:** Integrating chatbot usage with corporate authentication systems. This ensures only authorized users can access the system and provides a traceable record of interactions for security and accountability.
- **Model selection and hosting:** Choosing between public APIs (e.g., OpenAI) or private LLMs hosted on-premises. This decision impacts cost, control, and data security, tailoring the solution to the organization's specific needs.
- **Monitoring and feedback loops:** Capturing usage patterns and user feedback (particularly from procurement managers) to improve performance and trust. This enables continuous improvement, ensuring the system meets users' needs and builds confidence in its reliability.
- **Pilot testing and phased rollout:** Deploying the chatbot in a controlled environment to gather operational insights before scaling. This step validates functionality and identifies issues early, reducing risks during full deployment.
- **Change management and user training:** Supporting adoption among procurement analysts and decision-makers. This promotes user acceptance and maximizes the tool's utility by equipping staff with the skills to leverage it fully.

As AI continues to evolve, driven by leaders such as OpenAI, Google, Microsoft, and Meta, businesses have growing opportunities to embed GenAI into day-to-day operations. Our project illustrates that even with basic Python and prompt engineering skills, companies can harness GenAI to streamline insight generation, reduce operational burden, and enable procurement professionals to shift their focus toward strategic decision-making and value creation.

This project not only addresses a specific need within the sponsor company's procurement function but also sets a foundation for broader adoption. Other teams within the company, such as finance, operations, or supply chain planning, can leverage this framework to democratize data insights, tailoring it to their specific datasets and queries. Beyond the sponsor organization, other companies in industries like manufacturing, retail, or healthcare—where complex procurement data drives strategic decisions—can adopt this approach to optimize analytics, cut costs, and foster agility. Our contribution is a step forward to empower organizations to shift from static, resource-intensive reporting to dynamic, AI-driven insights, unlocking strategic value creation and paving the way for widespread AI integration in business operations.

6 REFERENCES

5 Procurement Unlocks Achieved via Data Visualizations. (2022). *FactWise.* Retrieved April 4, 2025 from https://factwise.io/blog/post/data-visualization-procurement

Mittal, A., Cocoual, C., Erriquez, M., & Liakopoulou, T. (2024). Revolutionizing procurement: Leveraging data and AI for strategic advantage. *McKinsey & Company*. Retrieved March 27, 2025 from https://www.mckinsey.com/capabilities/operations/our-insights/revolutionizing-procurement-leveraging-data-and-ai-for-strategic-advantage#/

Alteryx (2024). Alteryx Auto insights. Retrieved March 27, 2025 from https://www.alteryx.com/products/auto-insights

Bengesi, S., Hoda, E.-S., Sarker, M., Houkpati, Y., Irungu, J., & Oladunni, T. (2023). *Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers.* Retrieved March 27, 2025 from https://arxiv.org/abs/2311.10242

Bergmann, D., & Stryker, C. (2024). *What is a variational autoencoder?* Retrieved from https://www.ibm.com/think/topics/variational-autoencoder

Ng, A. O. (n.d.). Generative AI for Everyone. Coursera. https://www.coursera.org/learn/generative-ai-for-everyone/

Dibia, V. (2023). *LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models*. Retrieved March 27, 2025 from https://arxiv.org/abs/2303.02927

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). *Generative Adversarial Nets*. Retrieved March 27, 2025 from https://arxiv.org/abs/1406.2661

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. Retrieved March 27, 2024 from https://arxiv.org/abs/2006.11239

How to Overcome AI Hallucinations Using Retrieval-Augmented Generation. (2024). *IT Convergence.* Retrieved April 4, 2025 https://www.itconvergence.com/blog/how-to-overcome-ai-hallucinations-using-retrieval-augmented-generation/

LangChain (2024). *LangChain State of AI Agents*. Retrieved March 27, 2025 from https://www.langchain.com/stateofaiagents

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Retrieved April 4, 2025 from https://arxiv.org/abs/2005.11401

Maddigan, P., & Susnjak, T. (2023). Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models. *IEEE*. Retrieved March 27, 2025 from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10121440

McKinsey. (2023). *The economic potential of generative AI: The next productivity frontier.* McKinsey & Company. Retrieved March 27, 2025 from https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

Michelucci, U. (2022). *An Introduction to Autoencoders*. Retrieved March 27, 2025 from https://arxiv.org/abs/2201.03898

Mitra, R., Endert, A., Narechania, A., & Stasko, J. (2020). *NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries*. Retrieved march 27, 2025 from https://arxiv.org/abs/2008.10723

OpenAI. (2020). Language Models are Few-Shot Learners. Retrieved March 27, 2025 from https://arxiv.org/abs/2005.14165

PwC. (2024). *Global Annual Review 2024: Artificial Intelligence.* Retrieved March 27, 2025 from https://www.pwc.com/gx/en/about/global-annual-review/artificial-intelligence.html

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *OpenAI blog.* Retrieved March 27, 2025 from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*. Retrieved March 27, 2025 from https://arxiv.org/abs/2204.06125

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., . . . Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. Retrieved March 27, 2025 from https://arxiv.org/abs/2102.12092

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach.* Pearson.

Tableau (n.d.). *How Tableau Pulse powered by Tableau AI is Reimagining the Data Experience*. Retrieved March 27, 2025 from https://www.tableau.com/blog/tableau-pulse-and-tableau-ai

ThoughtSpot (n.d.). *The AI analyst for everyone*. Retrieved March 27, 2025 from https://www.thoughtspot.com/product/ai-analyst

Vaswani, A., Shazeer, N., & Parmar, N. (2017). *Attention Is All You Need*. Retrieved March 27, 2025 from https://arxiv.org/abs/1706.03762

Yilin Ye, J. H. (2024). Generative AI for Visualization: State of the Art and Future Directions. Retrieved March 27, 2025 from https://arxiv.org/abs/2404.18144