Optimizing Training Efficiency through Conversational LLM

Final Project Report – Team deFacto Global

Members[1]: Boyuan (Daniel) Zhang, Tao Li, Vibhas Goel, Guang (Jacky) Yang

GitHub Repository[2]

Executive Summary

This comprehensive final report encapsulates the successful culmination of a Boston University Capstone Project aimed at developing a Conversational Large Language Model (LLM) for deFacto Global Inc, an industry-leading SaaS company headquartered in Troy, New York, renowned for its Extended Financial Planning & Analysis (xP&A) software[3]. The primary objective of this project is to harness the power of machine learning to enhance the operational efficiency of client-facing consultants by providing an innovative LLM solution. The proposed LLM leverages natural language processing techniques to automate the delivery and comprehension of training materials, thereby alleviating consultants from repetitive tasks, and empowering them to focus on value-added engagements. In collaboration with data scientists at deFacto Global, this project marks a significant stride towards transforming operational paradigms and ensuring a strategic edge in the realm of business modeling and planning.

The goal of this report is to provide a solution into the company's request and its objectives in optimizing its operations. We explore the approach that led to the successful extraction of complete text data from their official website. Furthermore, we delve into the findings obtained from an analysis of the explanatory text data. The report then documents our comprehensive research on three distinct models, detailing the rationale behind each exploration, along with the rationale for ultimately selecting Pinecone Model as our final choice. Additionally, this report aims to offer clear instructions for both technical and non-technical staff to seamlessly access and effectively utilize our developed model to enhance their business operations. Furthermore, we provide suggestions for future implications, including the deployment of the model on Google Cloud Platform and the creation of a user interface on WordPress, if needed in the future. By engaging in real-world usage of popular language models, this project serves as a catalyst for fostering discussions on leveraging machine learning to strategically position operations amidst competitors, ultimately enhancing stakeholder value.

Introduction

The deFacto Global Inc stands as an industry vanguard, distinguished by its pioneering prowess in the realm of business modeling and planning. These sophisticated capabilities

[1] Equal Contribution.

[2] Per agreement with the company, we are temporarily making the access to out GitHub Repository public for 24 hours for grading after masking sensitive information, **please do not share this link other than for grading purposes**.

[3] deFacto Global. (n.d.). "deFacto Global: Take Command of your Business Succuss." Retrieved from https://training.defactoglobal.com/course/defacto-planning-user-guide-planner/

seamlessly integrate within the fabric of Excel, Power BI, and web interfaces, seamlessly constituting a pivotal component of their cutting-edge xP&A platform[4]. Driven by an unwavering dedication to delivering solutions of enterprises, deFacto Global remains resolutely committed to furnishing organizations with the requisite tools and technologies essential for the optimization of their planning, analysis, and decision-making processes. The goal of our project is to conceptualize and construct an advanced Conversational LLM that would alleviate consultants from redundant tasks, elevate operational throughput, and underscore the company's commitment to technological innovation and customer satisfaction.

Methodology

Based on recent research of Burak Ceylan on AIMultiple, the development process of the conversational LLM was founded on established methodologies for training language models and utilized existing techniques to ensure precision, relevance, and comprehensiveness[5]. Our approach encompassed standard procedures, commencing with the extraction and pre-processing of text data from the official deFacto Planning Training Environment website[6]. Subsequently, we conducted a comprehensive study to identify potential LLM models suited for personalized text data, such as software user guides. After model selection, we engaged in training using pre-processed data and natural language processing techniques. Furthermore, the development process integrated methodologies for model evaluation and optimization. The performance evaluation of the LLM will be conducted using quantitative metrics, encompassing aspects such as querying speed and result accuracy with contextual understanding[7]. Notably, within the scope of deFacto Global, considerations pertaining to model cost are diligently considered. Qualitative evaluation, facilitated through reviews and feedback by business consultants at deFacto Global, is instrumental in affirming the LLM's efficacy in generating succinct and lucid user manuals.

Data Extraction & Exploratory Data Analysis (EDA)

a. Data Extraction

The process of extracting data involved direct collection from deFacto Global Inc's website, accomplished through the utilization of the Chrome Driver in conjunction with Selenium. This approach automated critical tasks such as login procedures, course selection, and the subsequent extraction of content from individual tabs. The selection of the Selenium Chrome Driver was a strategic choice, driven by its superior performance in comparison to alternative

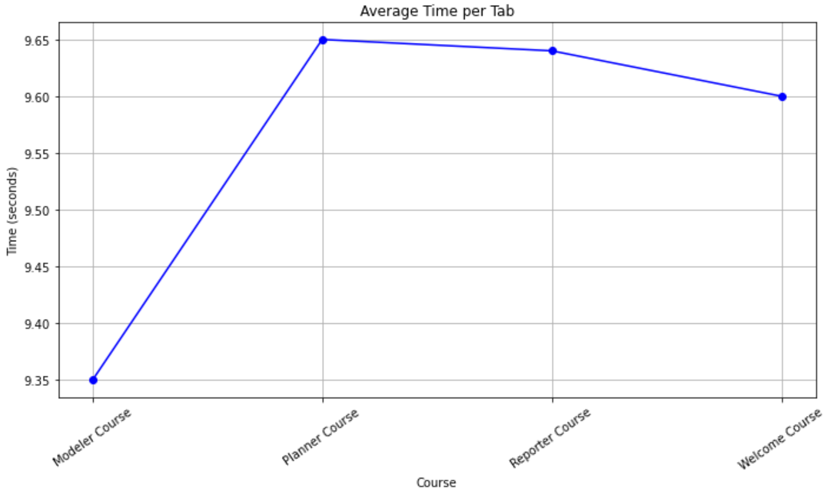
[4] deFacto Global. (n.d.). "deFacto Global: deFacto Power Planning (xP&A)." Retrieved from https://defactoglobal.com/defacto-power-planning/

[5] Ceylan, B. (2023, July 4). Large language model evaluation in 2023: 5 methods. AIMultiple. https://research.aimultiple.com/large-language-model-evaluation/

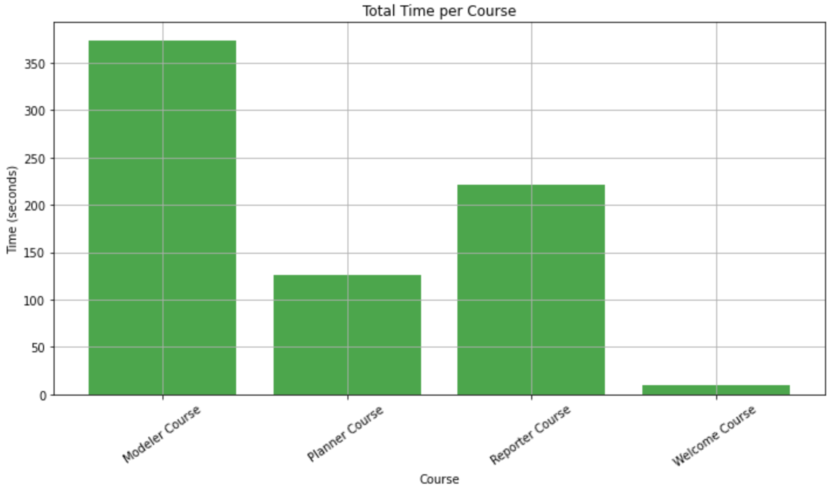
[6] Defacto Global. (n.d.). "Defacto Planning User Guide - Planner." Retrieved from https://training.defactoglobal.com/course/defacto-planning-user-guide-planner/

[7] Ceylan, B. (2023, July 4). Large language model evaluation in 2023: 5 methods. AIMultiple. https://research.aimultiple.com/large-language-model-evaluation/

data extraction tools, as researcher Ajay Sudhir Bale's article concluded in 2022[8]. This distinction was particularly evident in the chrome driver's capacity to extract a larger volume of data, reinforcing its appropriateness for this project's objectives. Through this meticulously orchestrated process, raw text data was successfully extracted from four distinct courses, each encompassing an array of 13 to 40 content-rich tabs. This amalgamated data was thoughtfully organized into a comprehensive table, enriched with corresponding topic names. Notably, the efficacy of the Chrome Driver is empirically evidenced in *Graphs 3-1* and *Graph 3-2*, presented below. These visual representations underscore the efficiency of the extraction process, with the average time required to extract data from a tab maintaining remarkable consistency at approximately 9.56 seconds. This noteworthy achievement signifies the expeditious acquisition of content from websites safeguarded by login and password protection mechanisms.



Graph 3-1: Average Scraping Time per Tab (in seconds)
From GitHub Access in Appendix A-1



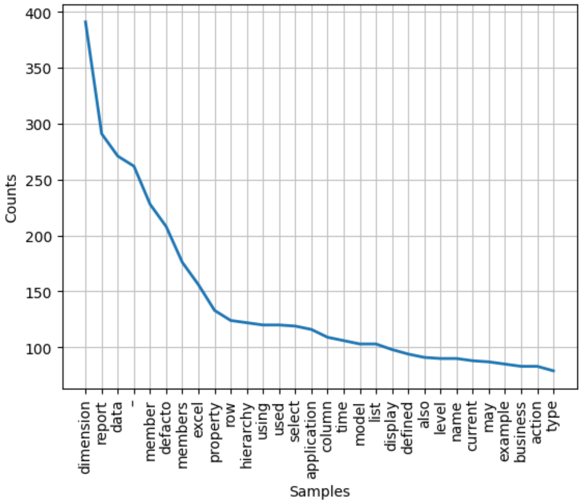
Graph 3-2: Total Scraping Time per Course (in seconds)
From GitHub Access in Appendix A-1

[8] Bale, A. S. (2022, September). "Web scraping approaches and their performance on modern websites." ResearchGate. Retrieved from https://www.researchgate.net/publication/363669276_Web_Scraping_Approaches_and_their_Performance_on_Modern_Websites

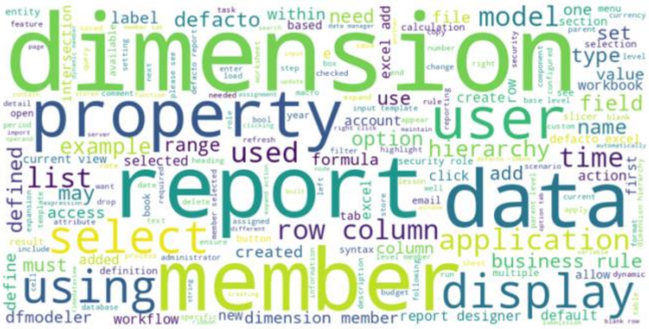
b. Data Pre-Processing & EDA

The primary preprocessing of the raw data entailed tokenization and lemmatization; a meticulous approach well-suited to the predominantly textual nature of the content. With terms such as "dimension," "will," "member," and "report" prominently featured in *Graph 3-3*, a visual depiction of the most frequently occurring words across all four courses, substantiates the verbal essence of the training materials tailored for clients of diverse backgrounds.

Furthermore, delving into the top 20 words exhibiting the highest lemmatization count (*Graph 3-4*) unveils intriguing differentiations from the most frequently occurring tokenized words. This observation underscores the nuanced nature of the materials, wherein the primary focus revolves around the intricate lexicon associated with in-software metrics. Notably, the recurrent token "dimension" (occurring 391 times) takes on a distinct contextual connotation, representing the "organized categories that define the structure of clients' business and the attributes of the tabular database."[9] This thread extends to similar terms such as "report," "property," and "hierarchy," all of which serve as descriptors for distinct features within the software framework. It is crucial to interpret these terms within the specific within-software features, rather than relying solely on their verbal meanings.



Graph 3-4: Lemmatized words with the highest count.
From GitHub Access in Appendix A-3



Graph 3-5: Tokens with the highest frequency.
From GitHub Access in Appendix A-3

Model Selection & Evaluation

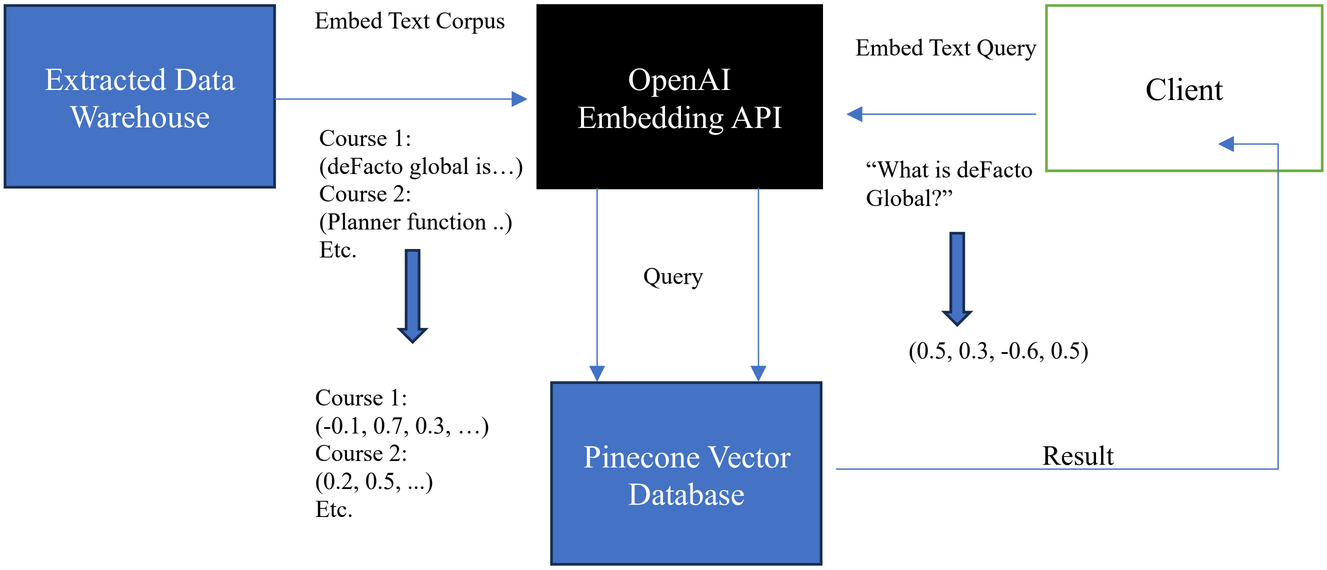
a. Pinecone Model

[9] deFacto Global. (n.d.). "deFacto Global: deFacto Power Planning (xP&A)." Retrieved from https://defactoglobal.com/defacto-power-planning/

It is noteworthy that while machine learning models may be trained on vast tracts of data, they lack inherent long-term memory or retention. Addressing this limitation, the Pinecone Model, as an external knowledge repository, comes to the fore. This fully managed vector database system imparts AI models with the essential long-term memory, enabling them to draw from contextual information and thereby ensure precise and timely responses.

The subsequent development of the model adhered to a systematic approach, succinctly depicted in *Graph 4-1*:

1. Utilization of the OpenAI Embedding API to generate vector embeddings of meticulously pre-processed data.
2. Subsequent upload of the vector embeddings into Pinecone, a platform adept at storing and indexing millions of vector embeddings, offering rapid searches at ultra-low latencies.
3. Passage of query text or questions through the OpenAI Embedding API once more.
4. Extraction of the ensuing vector embedding, subsequently serving as the query dispatched to Pinecone.
5. Receiving of semantically akin answers, even in instances where shared keywords may be absent.



Graph 4-1: Pinecone Model Structure
From GitHub Access in Appendix A-6

b. Dolly2 Model

Dolly2, an open-source Large Language Model (LLM) developed by Databricks, stands as a testament to advanced language processing. Notably, it circumvents the need for an external

API, ensuring the integrity of data and local execution of the framework[10]. Furthermore, Dolly2 provides a comprehensive package inclusive of training code, datasets, model weights, and an inference pipeline, all tailored for commercial utilization. With over 7 billion parameters, this fine-tuned model draws its proficiency from a human-generated instruction dataset. Dolly2 offers three variants, each characterized by varying parameter counts - 3 billion, 7 billion, and 12 billion[11]. The successful operation of this model, however, hinges significantly on computational resources and the desired precision level. Our exploration encompassed parameter variations, revealing that even the smallest configuration encountered operational challenges. Implementing this model at scale would necessitate a substantial capital infusion into computational resources, a proposition incongruent with the company's pragmatic expectations.

c. GPT4all + Langchain Model

GPT4all is a locally operated, privacy-conscious chatbot created by Nomic-AI, available for free use. This chatbot operates without the need for GPU or internet connectivity, ensuring that there is no risk of data leakage to external applications[12]. LangChain, on the other hand, is designed for tailoring any Large Language Model (LLM) with original data, enabling fine-tuning on custom datasets. This capability was harnessed to utilize Dolly2 for accessing and processing deFacto's specific training material. The LangChain process begins with document parsing, wherein the entire document corpus is segmented into chunks and their embeddings are stored in a VectorStore—an encompassing vector representation of the data[13]. Subsequently, the initial query is directed to the framework, which in turn interfaces with the selected language model. The vector representation is then matched against the vector store using similarity search, facilitating the retrieval of pertinent information chunks from the database. These chunks are then fed back into the language model.

However, as we discover after building the GPT4all +Langchain model, it encounters challenges such as response truncation and hallucination problem. To mitigate these issues, a pre-structured prompt needs to be included in the querying phase. Nevertheless, this introduces a limitation for deFacto Global, as it could inadvertently shift the workload back to the consultants, necessitating them to educate clients on its utilization once again.

d. Model Evaluation

i) Results from Sample Questions:

[10] Databricks. (n.d.). " Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM." Retrieved from https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

[11] Kumar, A. (2023, April 24). "Dolly2 and Langchain: A game changer for text data analytics." Medium. Retrieved from https://ashukumar27.medium.com/dolly2-and-langchain-a-game-changer-for-text-data-analytics-7518d48d0ad7.

[12] MonicAI, gpt4all, (n.d.), GitHub repository, https://github.com/nomic-ai/gpt4all

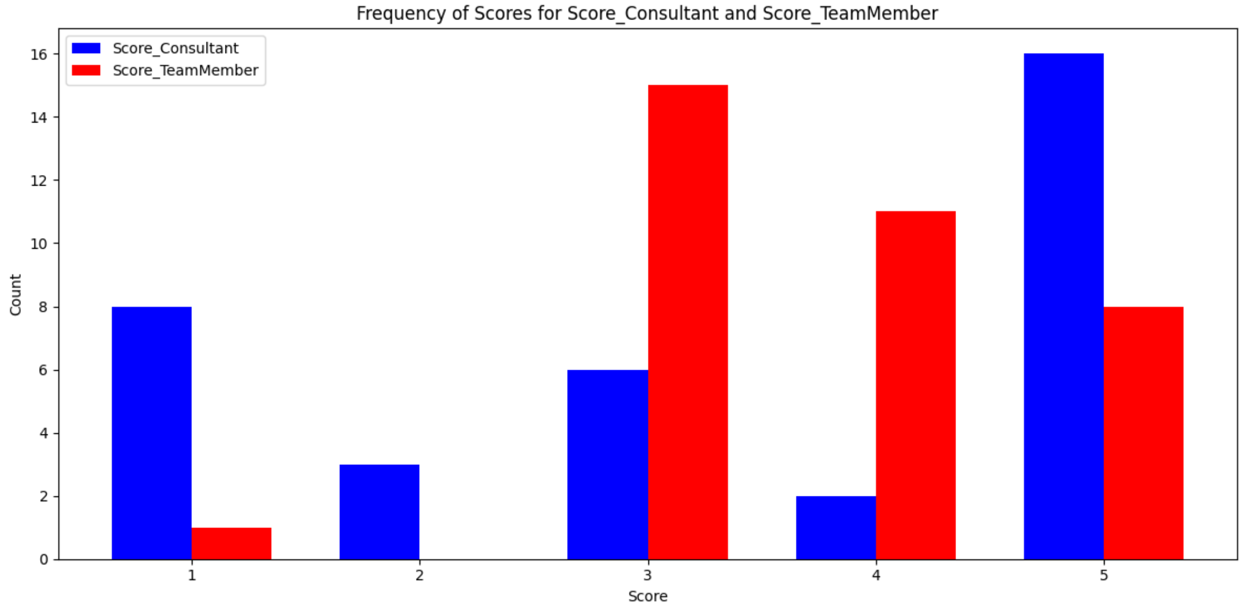
[13] Langchain. (n.d.). Power your applications with Large Language Models. https://www.langchain.com/

Given the extensive expertise of the consultants at the company, a comprehensive set of 37 frequently posed questions regarding software usage was meticulously compiled. Subsequently, each of these questions was subjected to rigorous testing utilizing the Pinecone model. Recognizing the intrinsic challenge of directly validating correct answers, we adroitly employed a user feedback approach, both from the adept consultants and our proficient team members.

The evaluation process involved a judiciously designed metric, enabling the assignment of ratings on a discerning scale ranging from 1 to 5, wherein 5 signifies the utmost accuracy. *Graph 4-2* vividly depicts the outcomes of this assessment, encapsulated in a histogram. Notably, the histogram showcases a pronounced trend, indicating a predominant clustering of answers within the accuracy levels of 3 to 5. This resounding consensus reverberates uniformly among both the esteemed consultants and dedicated team members, reaffirming the model's efficacy.

ii) Result from Comparative Analysis

During the crucial phase of model selection, a rigorous evaluation process was executed to identify the most optimal model. Our selection criteria revolved primarily around two key factors: response speed and result accuracy. After meticulous analysis, the Pinecone Model emerged as the clear choice, owing to its impressive processing speed, commendable accuracy in responses, and cost-effectiveness. Refer to *Table 4-3* for details. The Pinecone Model demonstrated the least capital investment at $0.004 per thousand tokens, outperformed others with an average query response time of 10 to 15 seconds, and consistently provided highly accurate answers, aligning well with the correct responses. Compared to the GPT4all model and Dolly2 model where significant monthly charges could be incurred, the Pinecone Model is the most cost-efficient option.



Graph 4-2: Consultant/Teammate Evaluation of Sample Results
From GitHub Access in Appendix A-6

Model	**GPT4all +Langchain**	**Dolly2**	**Pinecone**
Parameters	7 billion	3 billion	1.76 trillion (Pre-trained)
Speed	Crushed	10-15 minutes / query	Avg. 15 seconds/query
Cost	$2000+ per month	$1800+ per month	$0.004 per 1,000 tokens
Accuracy	High accurate if successful	High accurate if successful	Accurate
Limitations	Hallucination, Structured prompt	Significant computational resources	Relatively less accurate than others

Table 4-3: Comparative Analysis Result
From the process of 3 model building, the Cost for GPT4all and Dolly2 are based on the VertexAI's computational need as shown in Google Cloud Platform.

Finished Product and Conclusion

The culmination of this research endeavor yields two significant outcomes. Firstly, the development of a tailored LLM designed specifically for deFacto Global Inc's training materials achieves the objective of creating a potent and user-friendly tool. This tool enriches comprehension and accessibility of the xP&A platform by furnishing users with a comprehensive knowledge base. It offers lucid explanations, step-by-step guidance, and comprehensive examples for harnessing the platform's business modeling and planning capabilities. This empowerment equips users to fully exploit deFacto Global Inc's xP&A platform, resulting in heightened operational efficiency, precise decision-making, and overall business performance enhancement.

The deliverables encompass the following components:

1. Provision of the complete GitHub Repository inclusive of Google Collaboratory Notebooks. These notebooks encapsulate the entire process, including data extraction and storage, data pre-processing and analysis, algorithm development (inclusive of unsuccessful models), and model evaluation. (*Appendix A-1)*

2. Availability of a Demo Google Collaboratory Notebooks designed for querying sample questions and showcasing the retrieval of results, facilitating illustrative purposes for company executives and data scientists. (*Appendix A-7)*

3. A dedicated GitHub document elucidating the model's access and deployment process. This document provides precise file locations, meticulous instructions, maintenance approach and comprehensive solutions to previous encountered challenges. (*Appendix A-9)*

These deliverables collectively manifest our commitment to ensuring seamless knowledge transfer, accessible implementation, and tangible utilization of the developed model for the benefits of deFacto Global Inc.

In conclusion, this project marks a noticeable achievement in revolutionizing the way deFacto Global Inc. empowers its consultants and clients within the realm of LLM utilization. Through the meticulous development of a tailored model integrated with the training materials, we have successfully created a powerful, user-friendly tool that distinctly enhances the comprehension and accessibility of the xP&A platform. This achievement directly addresses the need to relieve the workload of consultants by providing a comprehensive knowledge base, streamlining client interactions, and facilitating efficient problem-solving.

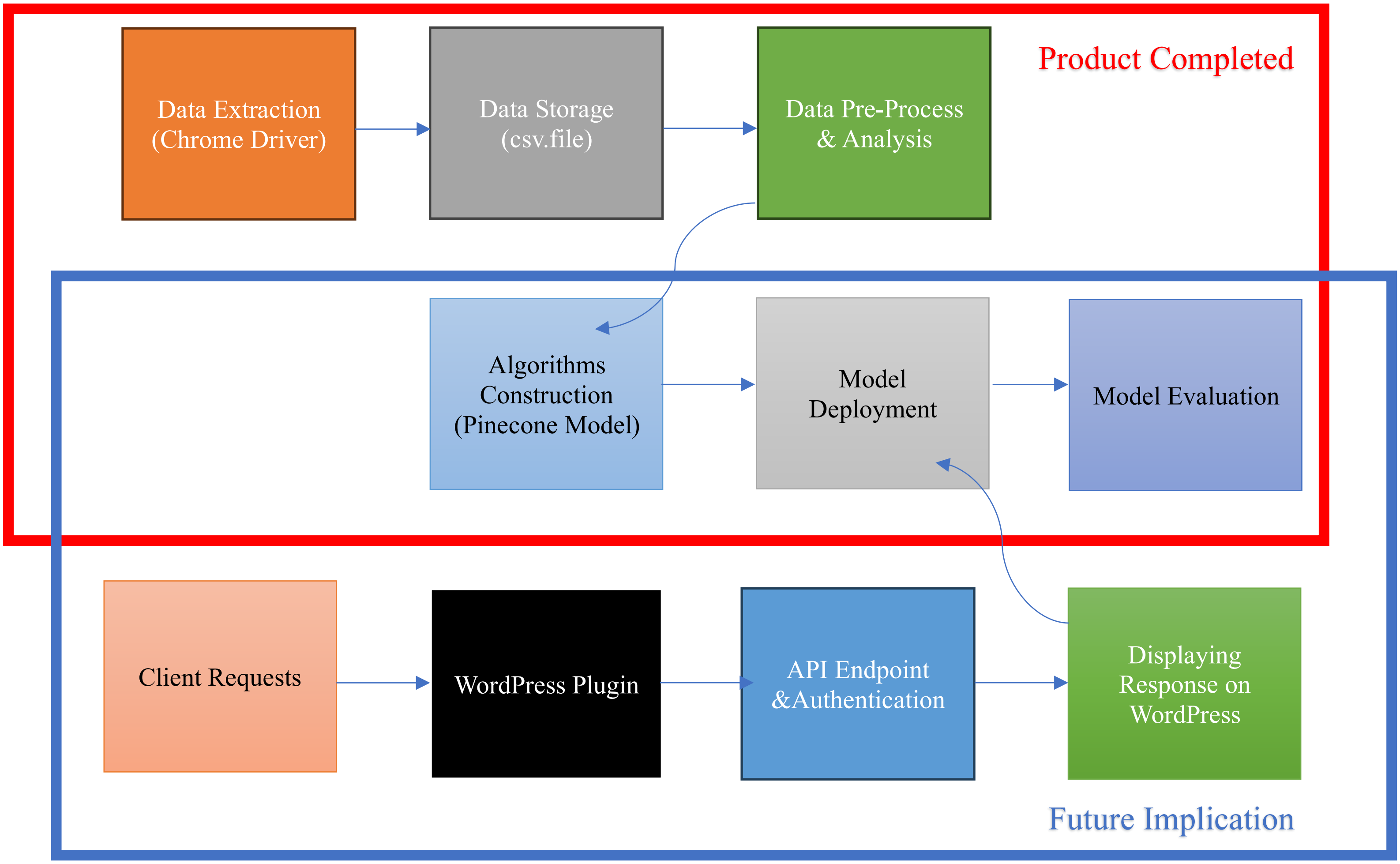
What truly sets our solution apart is its unparalleled focus on practicality, user-centricity, and cost-effectiveness. By leveraging the Pinecone Model, we have harnessed rapid processing capabilities, exceptional response accuracy, and a remarkably low capital investment. This strategic choice positions our solution as a standout competitor, not only providing accurate and insightful responses but doing so at an unmatched speed and cost-efficiency. Moreover, this project has provided us with invaluable insights and experiences. Our deep dive into the realm of conversational LLM has broadened our understanding of their intricate workings, enabling us to navigate the complexities of model selection, development, and evaluation. The incorporation of a cost-based approach in our evaluation process has further fortified our decision-making process, ensuring that the solutions we deliver not only excel in performance but also align seamlessly with budgetary considerations.

Future Implications

In the future, if feasible, integrating data scraping functionality into the user interface (UI) of the LLM will significantly enhance its capabilities and usability. As we are unavailable to construct a UI due to the time constraint of this project, we propose the following steps for the company with hypothetical WordPress plugin on Google Cloud Platform (*Graph 5-1*):

1. Create an API Endpoint
2. API Authentication
3. Deploy the Pinecone Model
4. Create a WordPress Plugin: Develop a custom WordPress plugin that sends requests to your API endpoint with user inputs (prompts) and receives responses from the language model.
5. Displaying Responses: Decide how to display the responses from the language model on your WordPress site.

6. Usage Limitations
7. Testing and Optimization


Product Completed
Data Extraction (Chrome Driver)
Data Storage (csv.file)
Data Pre-Process & Analysis
Algorithms Construction (Pinecone Model)
Model Deployment
Model Evaluation
Client Requests
WordPress Plugin
API Endpoint &Authentication
Displaying Response on WordPress
Future Implication

Graph 5-1: Finished Product vs. Future Implications

Appendix A:

Below we list links to our GitHub Repository based on categories:

GitHub Repository
https://github.com/K-3-LT/defacto_global_bu/tree/main

Data Analytics:
1. Data Extraction
https://github.com/K-3-LT/defacto_global_bu/blob/main/Data_Extraction.ipynb

2. Scraped Data
(This is sensitive information of the company, please contact us if you want to download)

3. Data Pre-process & EDA
https://github.com/K-3-LT/defacto_global_bu/blob/main/Training_material_visualization.ipynb

Machine Learning Models:
4. GPT4all + Langchain Model:
https://github.com/K-3-LT/defacto_global_bu/blob/main/GPT4_Model.ipynb

5. Dolly2 Model
https://github.com/K-3-LT/defacto_global_bu/blob/main/dolly2_3billion.ipynb

6. Pinecone Model with Model Evaluation
https://github.com/K-3-LT/defacto_global_bu/blob/main/Pinecone_Model.ipynb

7. Pinecone Model Demo Notebook
https://github.com/K-3-LT/defacto_global_bu/blob/main/Query.ipynb

Instructions for deFacto Global on Model Usage
8.Model Deployment and Instructions
https://github.com/K-3-LT/defacto_global_bu/blob/main/README.md

9. Evaluation Metrics for Consultants
https://github.com/K-3-LT/defacto_global_bu/blob/main/Consultant%20Rating%20Guideline.docx

10. Future Implications in Detail

https://github.com/K-3-LT/defacto_global_bu/blob/main/Future%20Implications.docx

Project Management

11. Scrum Master – GitHub Project Management
https://github.com/users/bdanielzhang/projects/2

12. Sprint Reports
https://drive.google.com/drive/folders/1w0Ib4tovbNb6o4TIfROS1VXEpww-QN16?usp=sharing

Appendix B:

LinkedIn Promotional Message –

Over the past three months, the collective efforts of Boyuan (Daniel) Zhang, Tao Li, Vibhas Goel, and Guang (Jacky) Yang have yielded a remarkable achievement – the successful development of a Conversational Large Language Model at deFacto Global Inc as their capstone project at Boston University Questrom School of Business. This endeavor, in collaboration with deFacto Global Inc, underscores our commitment to innovative learning and impactful industry partnerships.

At its core, their project sought to harness the power of machine learning models to elevate the operational efficiency of client-facing consultants. The driving force behind this initiative was the creation of an ingenious LLM solution, strategically designed to employ natural language processing techniques. This approach automates the dissemination and comprehension of crucial training materials, liberating consultants from routine tasks and enabling them to focus on value-driven engagements.

The project represents an intricate journey of discovery, beginning with an exploration of data extraction methodologies, culminating in a comprehensive analysis of model selection and evaluation. Students will find a meticulous breakdown of the decision-making process that ultimately led them to handle a real-world problem. This capstone experience has fortified their understanding of complex systems and fortified their ability to navigate the nuanced landscape of model development.

We extend our sincere gratitude to deFacto Global Inc for their unwavering support and guidance. This project stands as a testament to the possibilities that emerge when academic excellence converges with real-world innovation. It is our hope that this achievement inspires future cohorts to reach new heights at the nexus of education and industry.