Optimizing Training Efficiency through Conversational LLM

deFacto Global Inc

Final Presentation

Members:

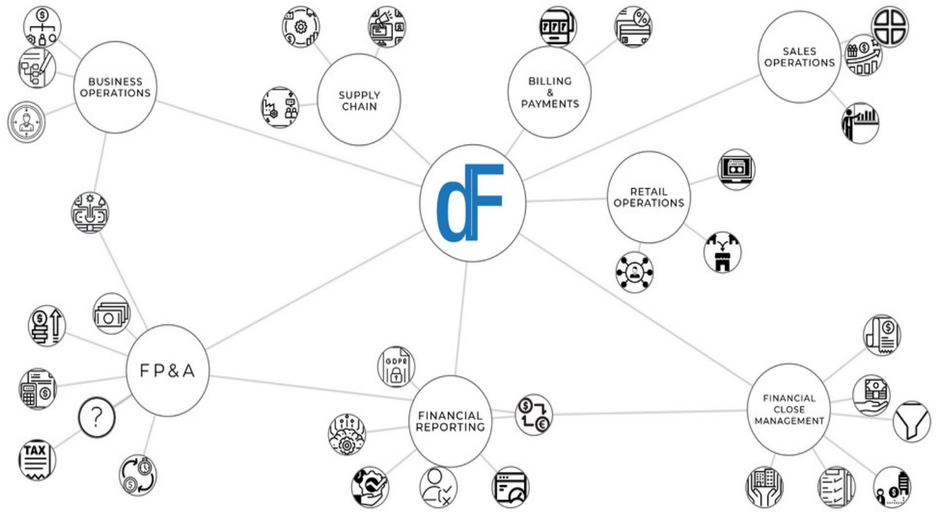
Boyuan (Daniel) Zhang, Tao Li, Vibhas Goel, Guang (Jacky) Yang

deFacto Global

BOSTON UNIVERSITY

Overview of the Corporation

- **deFacto Global** Inc specializes in **business modeling and financial planning** capabilities integrated into Excel, Power BI, and web interfaces as part of their extended planning analysis (xP&A) platform.
  - *Offers a user-friendly platform that optimize planning, analysis, and decision-making processes for financial and operational data.*



deFacto Global

BOSTON UNIVERSITY

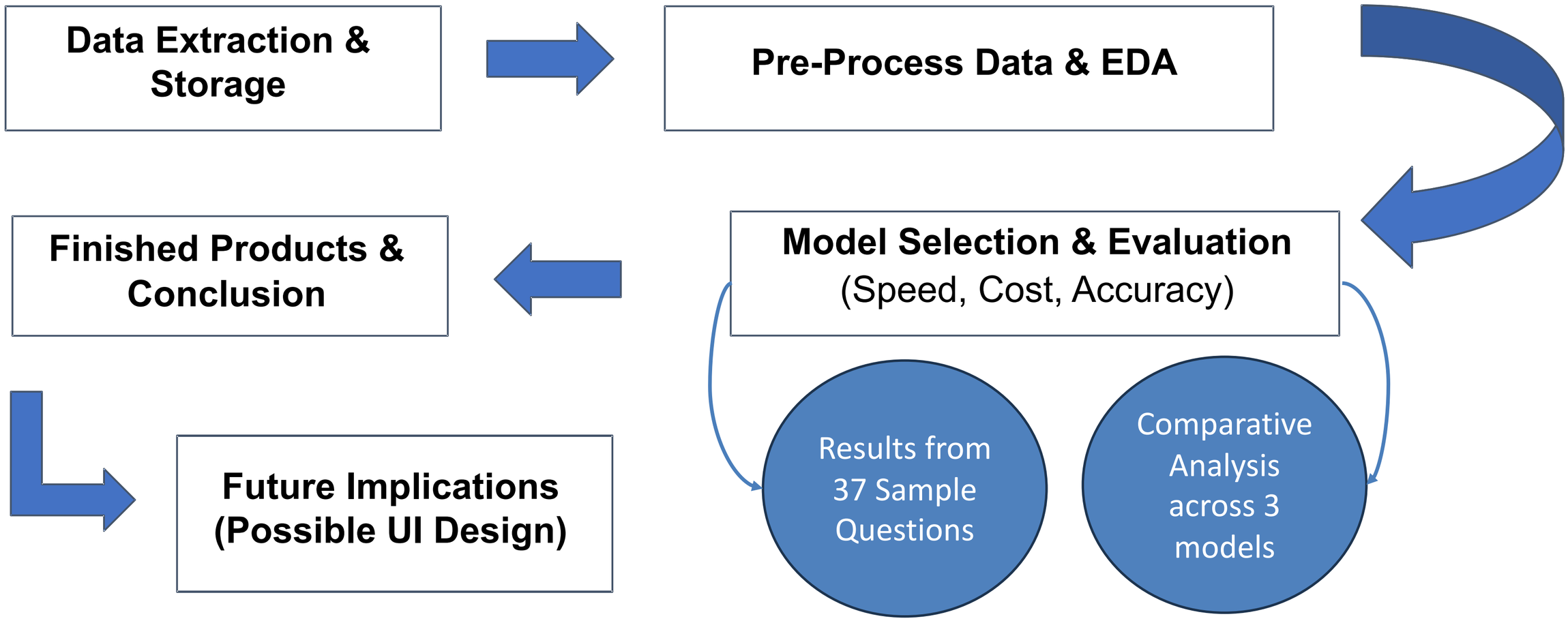
Project Objective

- Develop a **conversational large language model (LLM)** for its software training materials (user guide).
  - **Clients**
    - *Automated material delivery and comprehension process.*
    - *Faster and more straight-forward solution to answer questions regarding software.*
  - **Consultants**
    - *Reduced workload in repetitive explaining about technical difficulties from clients.*
    - *Empowered to focus on value-added job engagements.*
  - **Corporate**
    - *Transformed operational efficiency, strategically positioning in company operation.*
    - *Differentiating itself from competitors, enhance stakeholder value.*

deFacto Global

BOSTON UNIVERSITY

Methodology

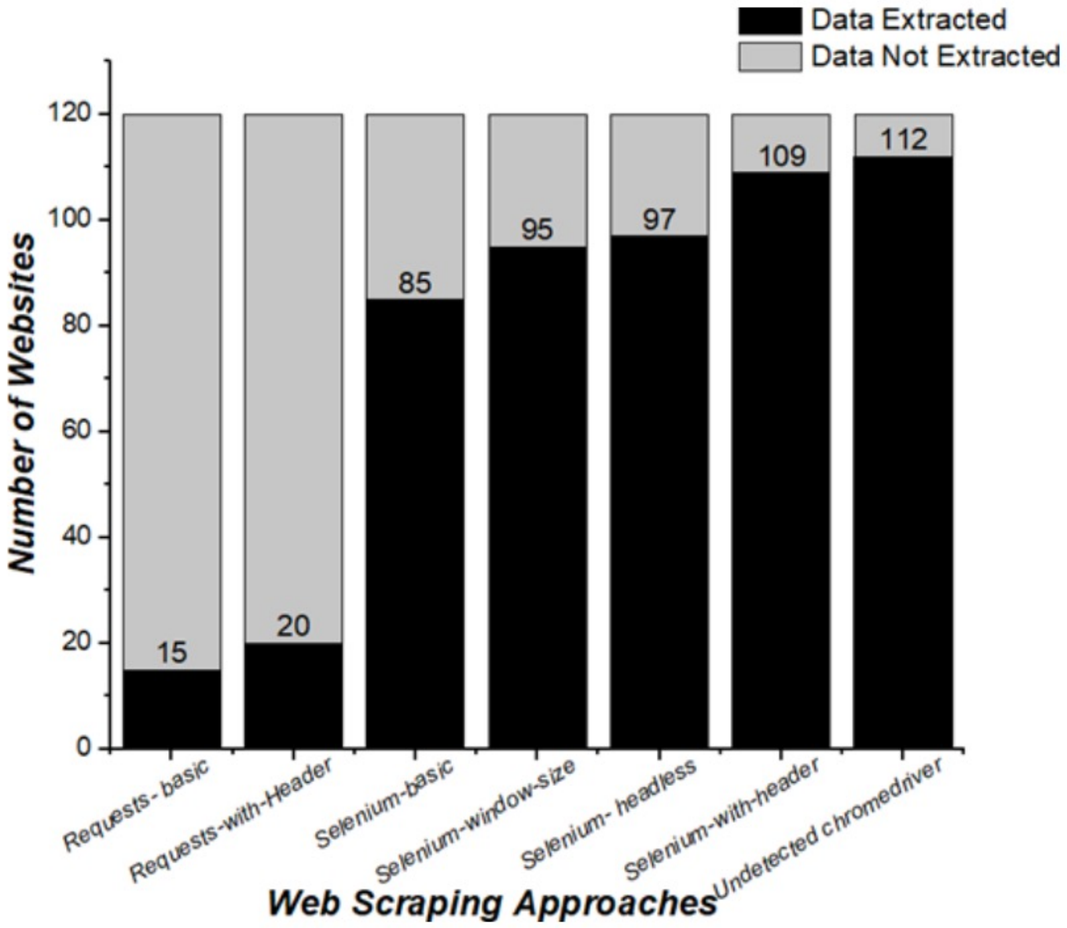


deFacto Global

BOSTON UNIVERSITY

Data Extraction

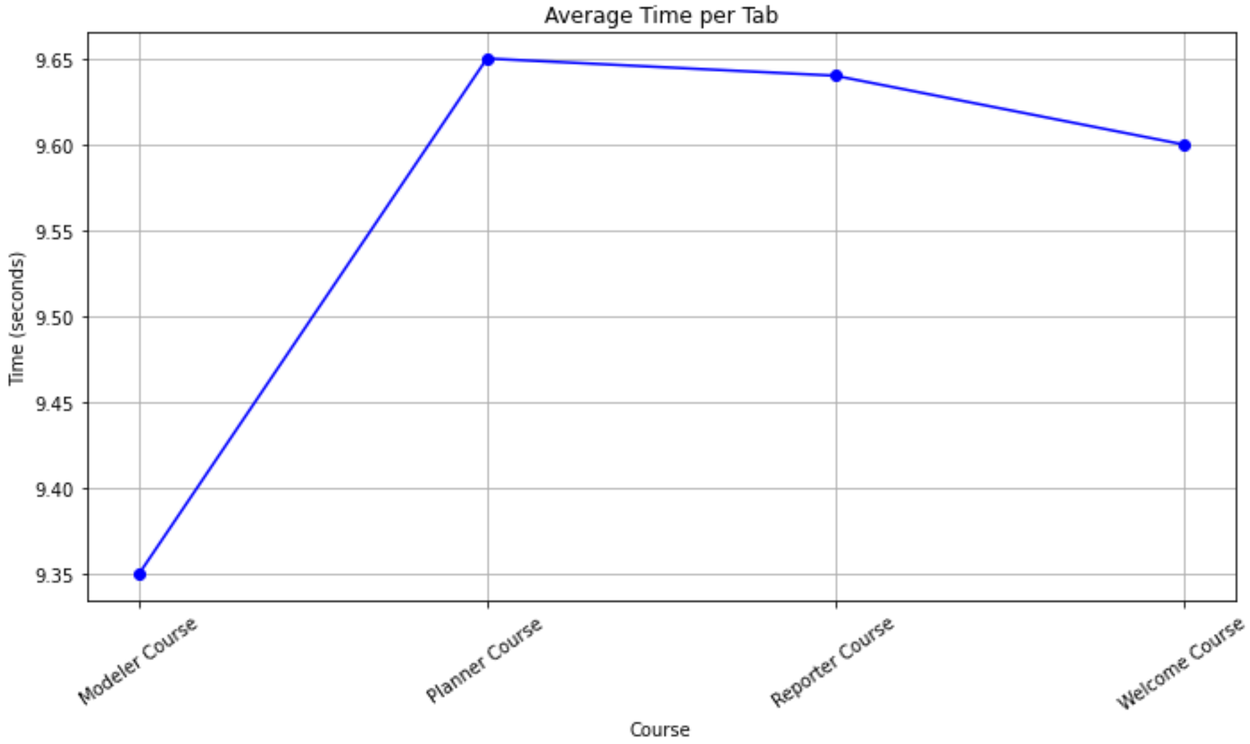
- **Most Efficient Method from Research:**
  - Selenium Chrome Driver
- **Automated Tasks:**
  - Login Procedures
  - Course Selection
  - Content Extraction from each Tab
- **Summary:**
  - 4 Courses
  - 96 Tabs
  - One Course: ~14,866 tokens



deFacto Global

BOSTON UNIVERSITY

Data Extraction (Results)



Average Time per Tab: ~ 9.50 seconds

- **Relatively fast for websites with login procedures.**
- **Data stored in csv. Format.**

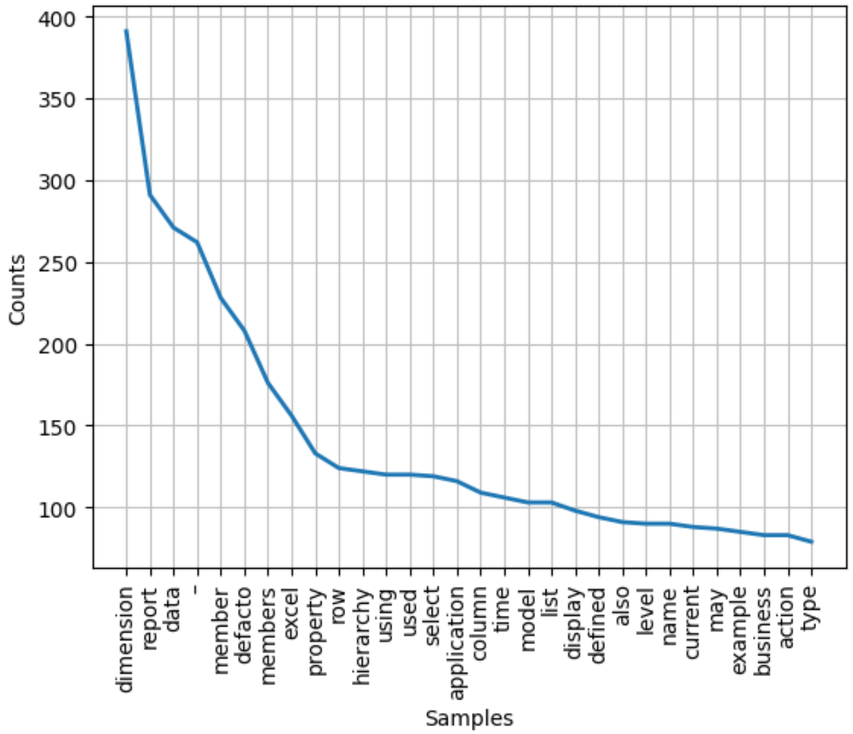
Total Time per Course : (Ranges)

- **Proves that the extraction method is efficient and stable.**

deFacto Global

BOSTON UNIVERSITY

Exploratory Data Analysis



- ***Lemmatized words with the highest count.***
  - **Ex. “Dimension”:**

    - “organized categories that define the structure of clients’ business and the attributes of the tabular database.”

- ***Tokens with the highest frequency***

deFacto Global

BOSTON UNIVERSITY

Models NOT Selected

- **Dolly2 Model.**
  - *An open-source Large Language Model (LLM) developed by Databricks.*

Key Features:

- No need for external API & local execution.
- Comprehensive package: training code, datasets, model weights, inference pipeline.

Parameters and Variants:

- 3 billion, 7 billion, and 12 billion parameters.

Operational Challenges:

- Computational resource dependency for operation.

- **GPT4all + Langchain Model**
  - *A locally operated, privacy-conscious chatbot created by Nomic-AI.*

Key Features of GPT4all:

- Locally operated, free for use.
- No GPU or internet needed, ensuring data privacy.

Key Features of Langchain:

- Document parsing: Corpus segmented & embeddings stored in VectorStore.
- Vector representation matched via similarity search for retrieval.

Challenges Encountered:

- Response truncation and hallucination problems.
- Inclusion of pre-structured prompt during querying.

Model Selected: Pinecone

- **Pinecone Model**
- *Pinecone: a platform adept at storing and indexing millions of vector embeddings, offering rapid searches at ultra-low latencies.*
  - Utilization of the OpenAI Embedding API to generate vector embeddings.
  - Subsequent upload of the vector embeddings into Pinecone,.
  - Passage of query text or questions through the OpenAI API once more.
  - Extraction of the vector embedding, subsequently serving as the query dispatched to Pinecone.
  - Receiving of semantically akin answers.

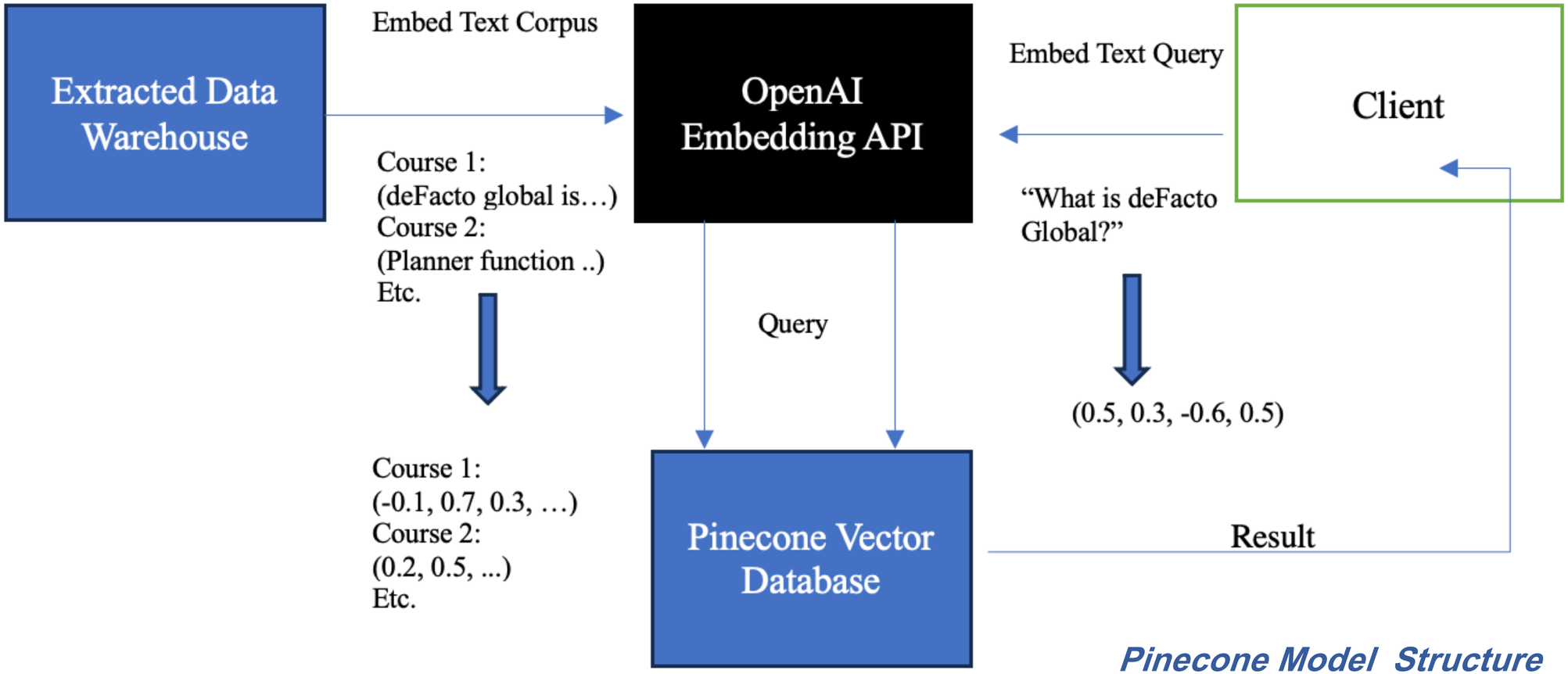
Advantages:

- Effective even when shared keywords are absent.
- Enables fast and accurate retrieval of relevant information.

deFacto Global

BOSTON UNIVERSITY

Model Selected: Pinecone (Continued)



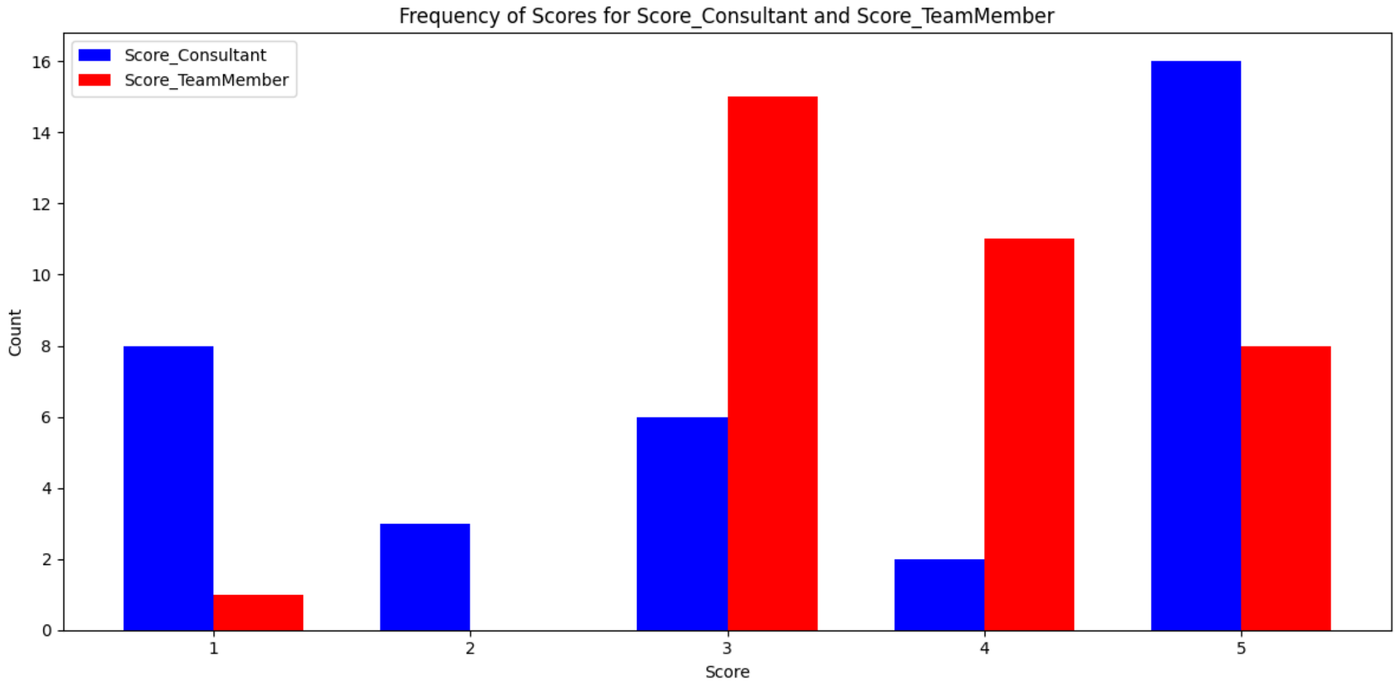
Pinecone Model Structure

deFacto Global

BOSTON UNIVERSITY

Model Evaluation

- **First Approach:**
  Sample Questions
  - 37 frequently posed software usage questions gathered.
  - Each question subjected to thorough Pinecone model testing.
  - Expert consultants and team members provided feedback.
  - Designed metric for rating answers on a 1 to 5 scale. (Guideline Provided)



Pinecone Model Evaluation

deFacto Global

BOSTON UNIVERSITY

Model Evaluation

- **Second Approach: Comparative Analysis**

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

deFacto Global

BOSTON UNIVERSITY

Demo Display

- **Double-click the below for a demo display of the Pinecone Model:**

deFacto Global

BOSTON UNIVERSITY

Conclusion

- **Development of a tailored LLM designed specifically for deFacto Global Inc's training materials**

1. Heightened operational efficiency, 2. Precise decision-making, 3. Overall business performance enhancement.

  - Document provided to the company regrading the model's access and deployment process.
  - Demo model for querying sample questions and showcasing the retrieval of results.
  - Achievements directly addresses the need to relieve the workload of consultants by providing a **comprehensive knowledge base, streamlining client interactions, and facilitating efficient problem-solving**.
  - Meaningful: incorporation of a cost-based approach in our evaluation process.

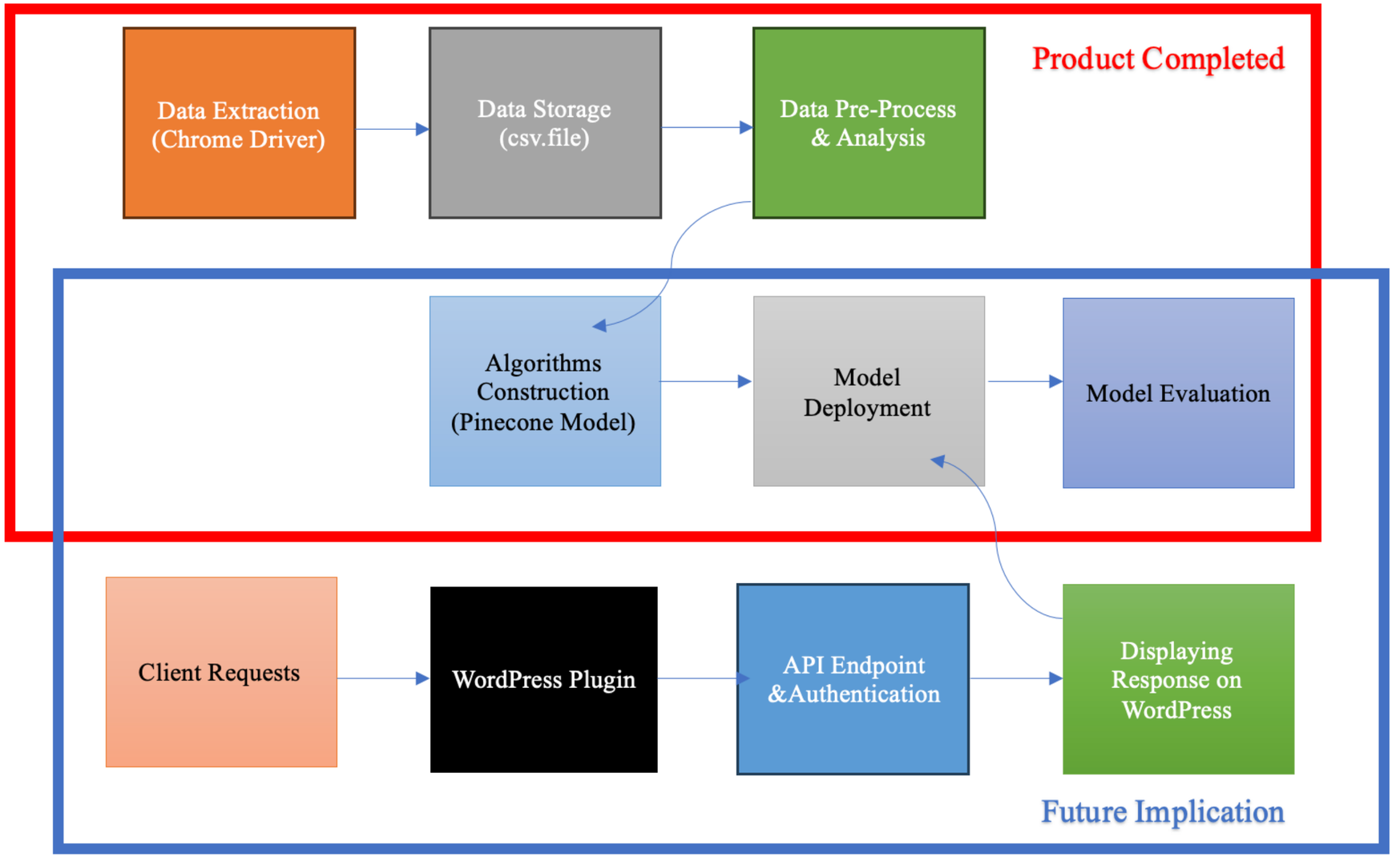
deFacto Global

BOSTON UNIVERSITY

Future Implications

- **User Interface (UI)**

1. Create an API Endpoint
2. API Authentication
3. Deploy the Pinecone Model
4. Create a WordPress Plugin
5. Displaying Responses
6. Usage Limitations
7. Testing and Optimization



Works Cited:

1. 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

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

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

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

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

deFacto Global

BOSTON UNIVERSITY

Selected GitHub Links :

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

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

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