STIBO SYSTEMS
MASTER DATA MANAGEMENT

ESG Certification Recommendation System

Team 2: Jackson Duke, Sofia Laval, Nathan Ruiz

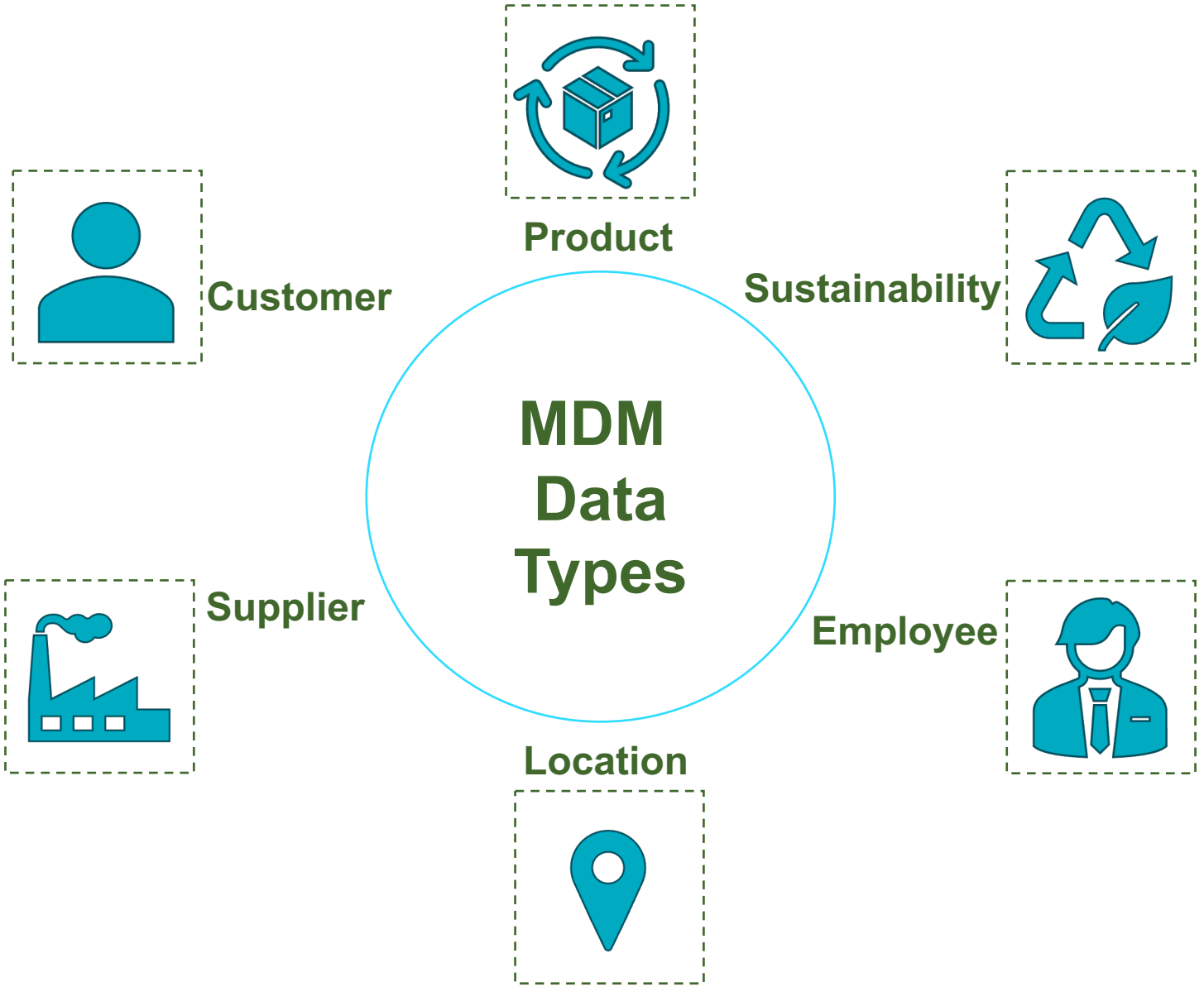
Agenda

1. Client Introduction
2. Problem
3. Exploratory Data Analysis
4. Methodology
5. Tool Demonstration
6. Model Evaluation
7. Conclusion

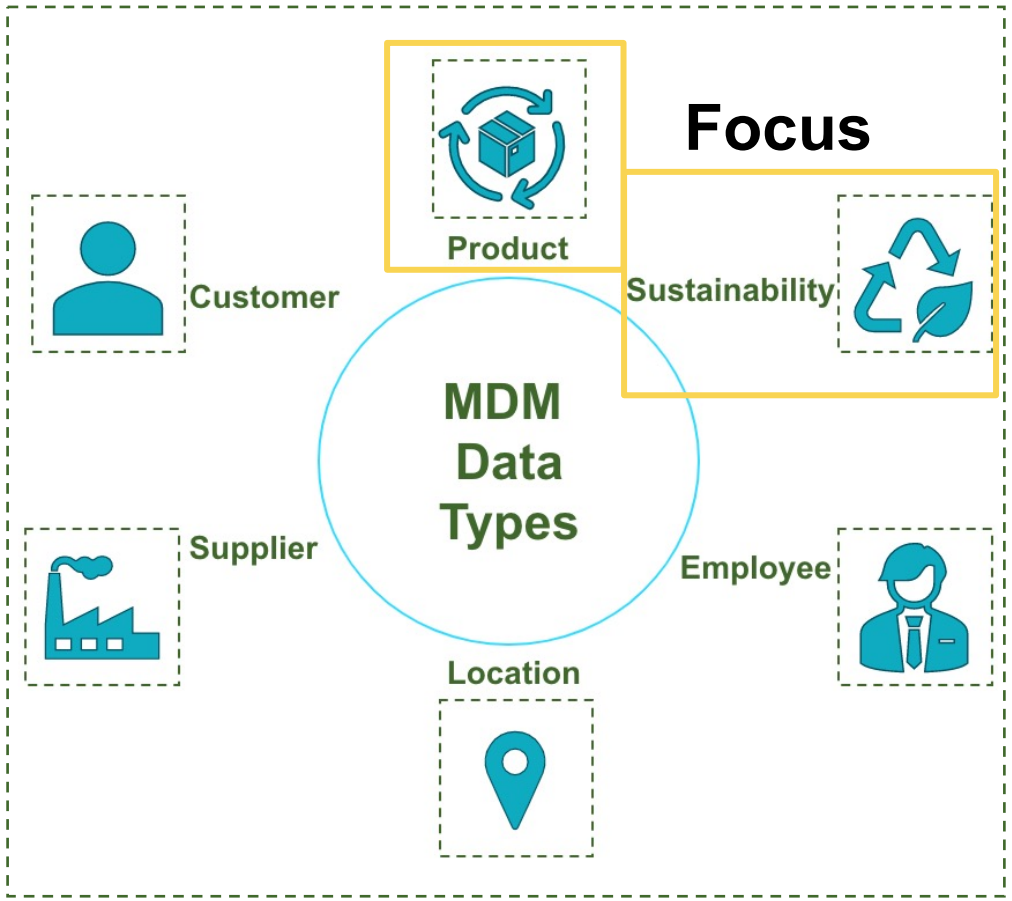
Client: Stibo Systems

Offers Master Data Management (MDM) solutions to centralize and manage critical data shared across core business departments, ensuring:

- Data consistency
- Data accuracy
- Data transparency



Problem and Motivation



How are products evaluated to ensure they meet the requirements to obtain certain sustainability certifications?

Problem and Motivation

Current Process

Sustainability ESG certification expert

Identify products that meet ESG certification(s) standards using various excel sheets

Map products to appropriate certifications

Problem

Error - prone

Time - consuming

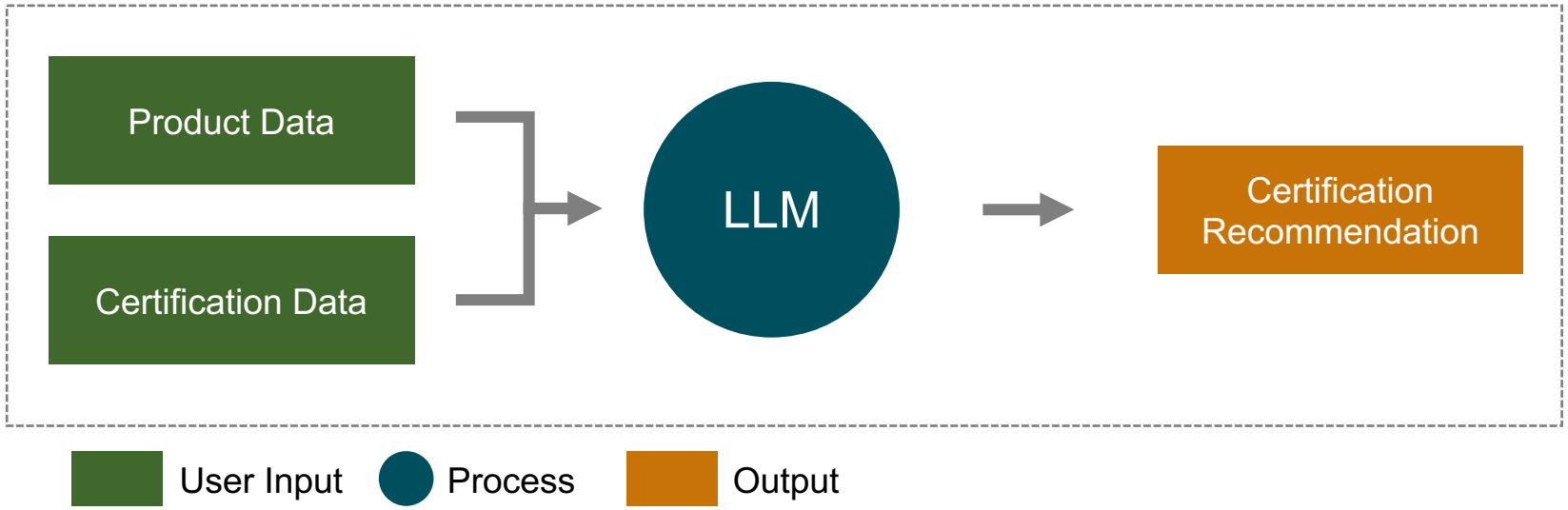
Manual

Incorrect mapping costly

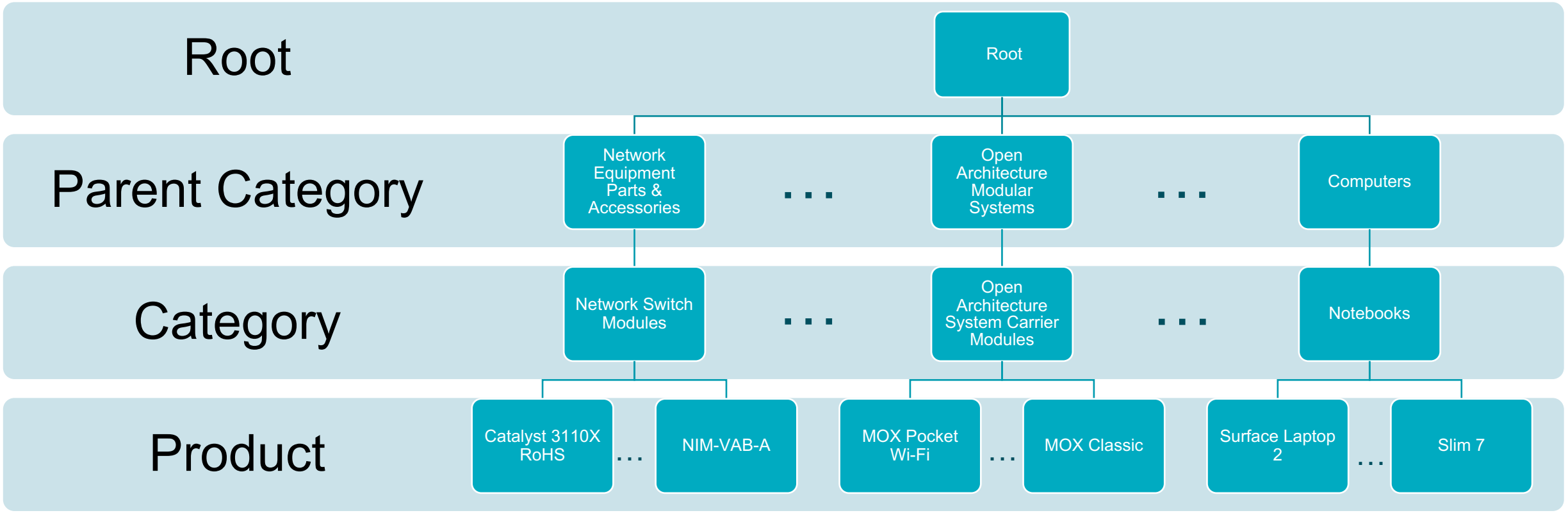
How can we improve the current process?

Improve upon the manual process by leveraging Large Language Models (LLMs). LLMs are gaining popularity for the ability to process and analyze large amounts of data. Given product and certification data, these models will assess whether a product meets the criteria for an ESG certification to:

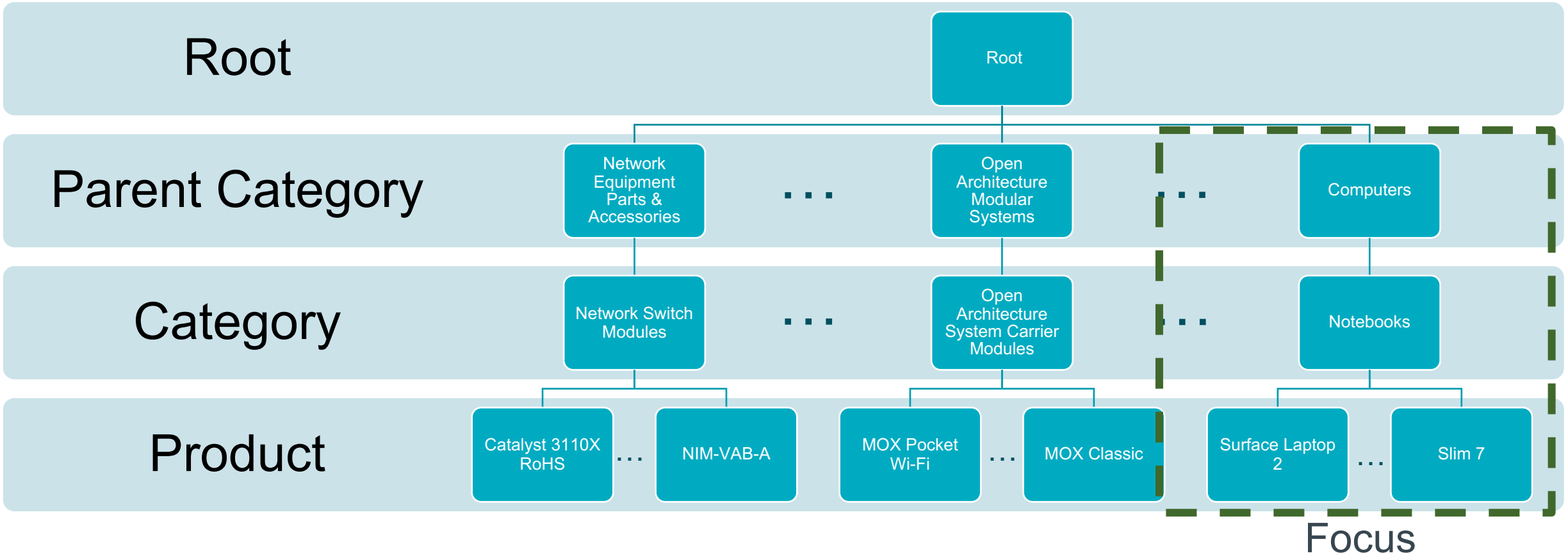
- Streamline the work for sustainability experts
- Improve accuracy of certification predictions
- Reduce costs associated with incorrect mappings



Exploratory Data Analysis - Hierarchical Structure of Product Data



Exploratory Data Analysis - Hierarchical Structure of Product Data



Exploratory Data Analysis – Notebook Product Data

Observations:

233,669

Notebook products

233,661

Unique notebook products

2,462

Notebook product features

Final Dataset dimensions

110,225

Or 47% ESG certified notebook products

346

Product features populated for >5% of rows

Example:

Attribute	Value
Product name	Slim 7
Processor manufacturer	Intel
Internal memory	16.0
Internal memory unit	GB
Battery Technology	Lithium Polymer
Sustainability Certificates	Energy Star
.	.
.	.
.	.

Exploratory Data Analysis - Certifications

- We are evaluating whether notebook products meet the certification requirements for TCO Certified and Energy Star. These certifications have specific criteria for notebooks, our chosen area of focus
- Each certification provides its requirements in a PDF format. The documents have varying lengths and formats to detail the various elements of a products that are necessary to receive certification
- With what's currently a manual process, we have scraped the most recent requirements PDFs for the relevant product mandates, resulting in a final list for the LLMs to assess:

Summary of certification mandates

Certification	Number of Mandates
Energy Star	9
TCO	17

Example certification mandate

TCO Mandate 5.2.1
Display Resolution

The display panel should have a pixel density of at least 100 PPI:

$$\text{PPI} = \frac{\sqrt{horizontal\ pixels^2 + vertical\ pixels^2}}{diagonal\ of\ the\ panel\ in\ inches}$$

Methodology

- Methods used for assessing ESG certification eligibility:

Data Enrichment

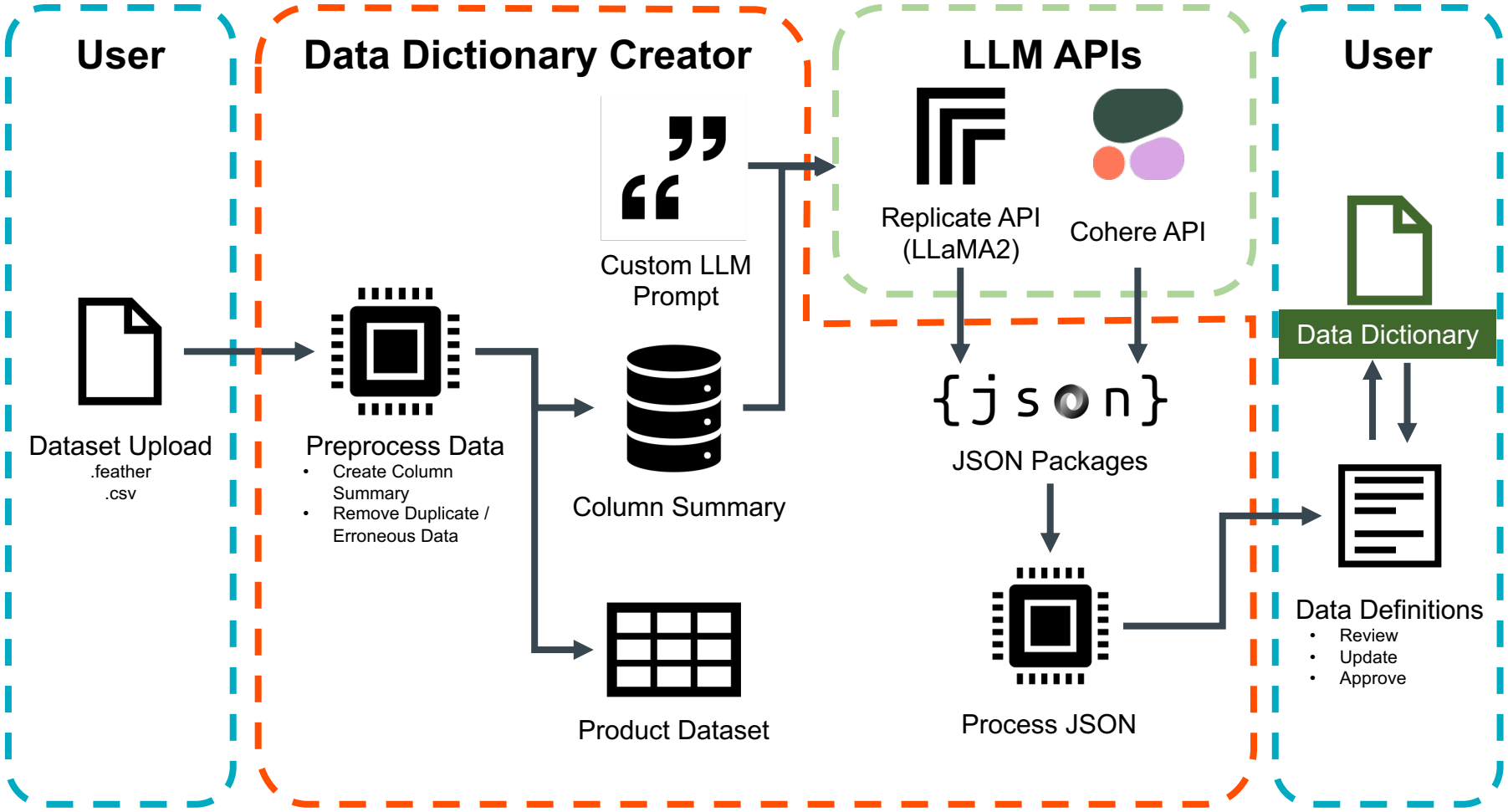
LLM Querying

Machine Learning Classification*

* Used as a baseline for comparison

Data Enrichment: Data Dictionary Creator

Architecture



Methodology

Goal: Using Large Language Model (LLM) APIs, enrich the product dataset with column definitions to create a more comprehensive summary of each product.

Steps:

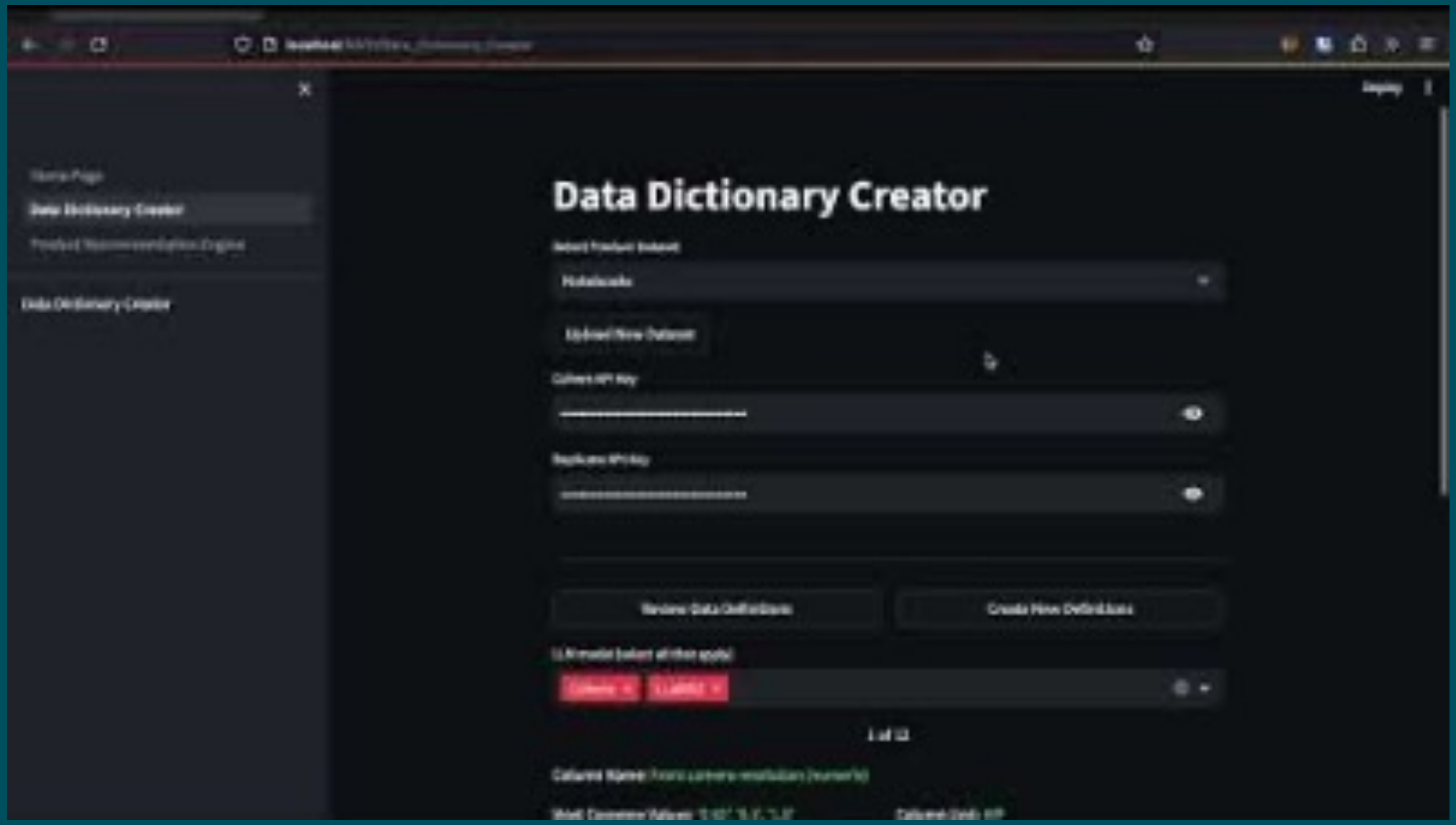
1. Upload product dataset
2. Preprocess data into column summary:

NAME	VALUES	UNIT	MIN	MAX
...	...	...	...	...

3. Query LLMs using a custom prompt to generate a column definition.
4. Review, update, and approve the generated definitions.
5. Export a complete data dictionary for the dataset to be used in the certification recommendation process.

Demo: Data Dictionary Creator

https://youtu.be/hYOZidi2hwY



LLM Querying: Recommendation Engine

Methodology

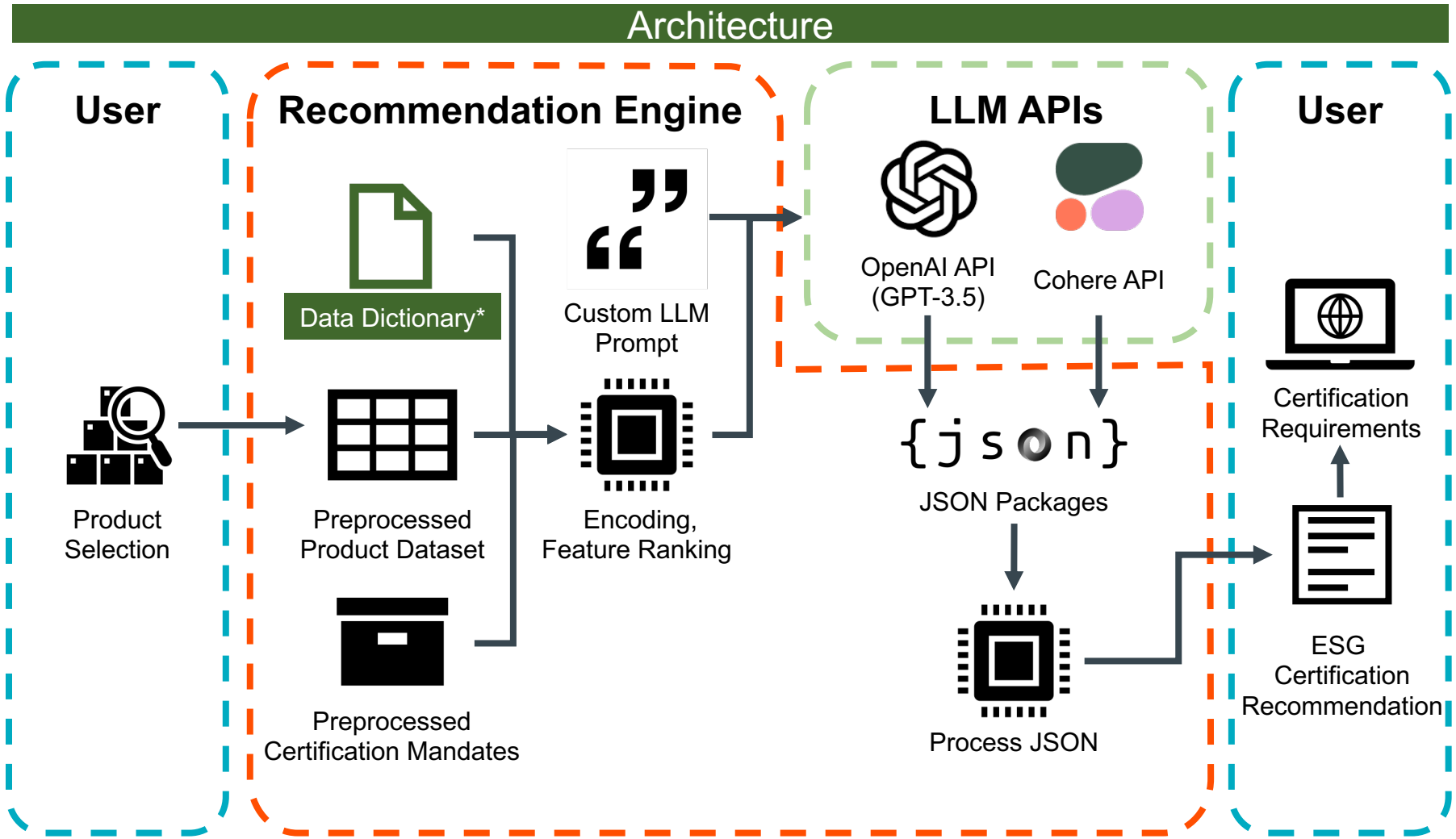
Goal: Using Large Language Model (LLM) APIs, summarize, encode and embed products and certification mandates to provide an assessment for certification.

Steps:

1. Preprocess: remove duplicate and erroneous data, create dataset columns definitions.
2. For each mandate, rank columns by relevance.
3. Summarize, encode, and embed relevant product data for each mandate.
4. Query LLMs via API to receive recommendation on mandates

Evaluation:

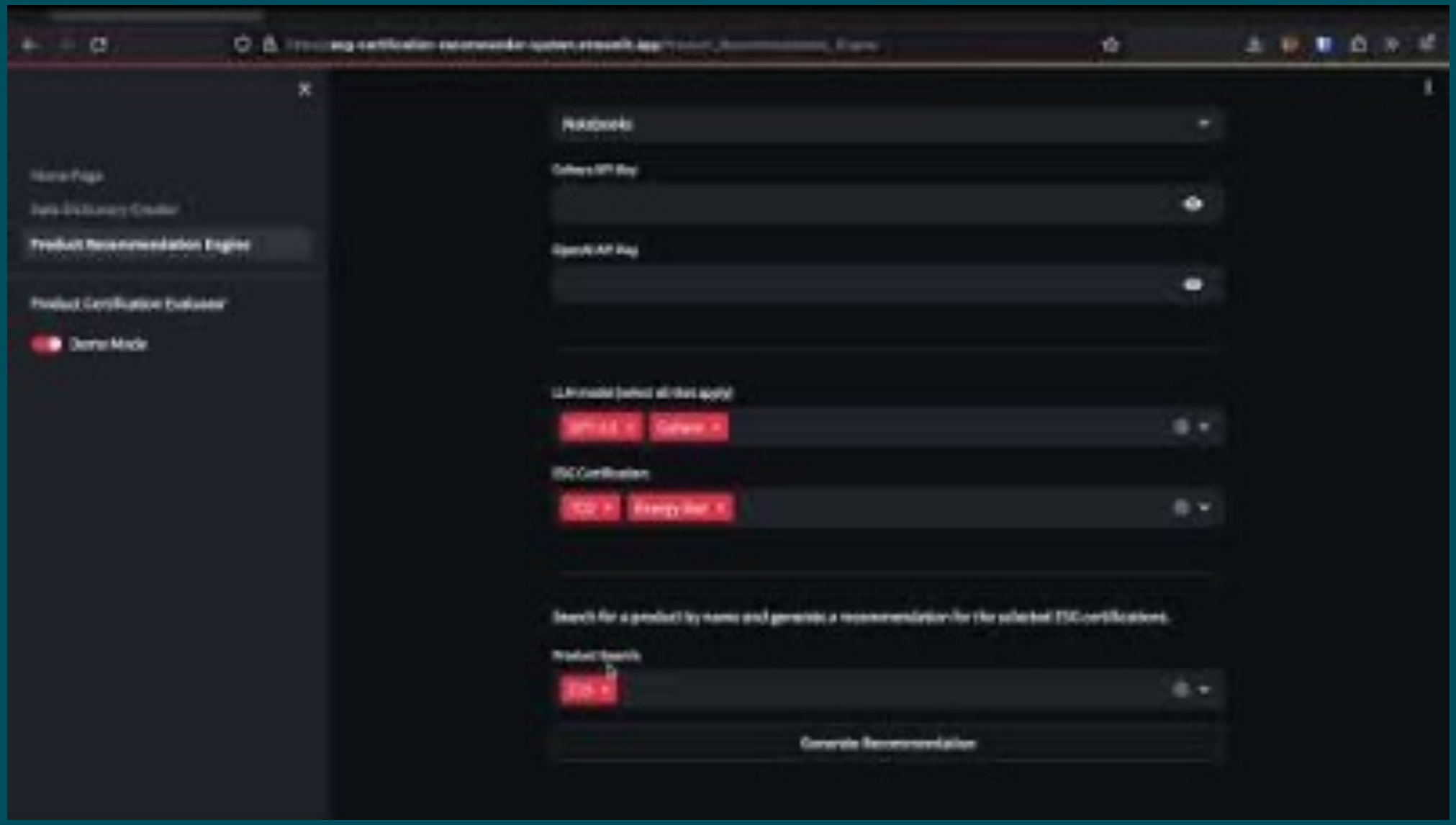
- Compare model output with list of already certified products.
- Compare performance between different LLMs.



Final output from the Data Dictionary Creator module

Demo – Recommendation Engine

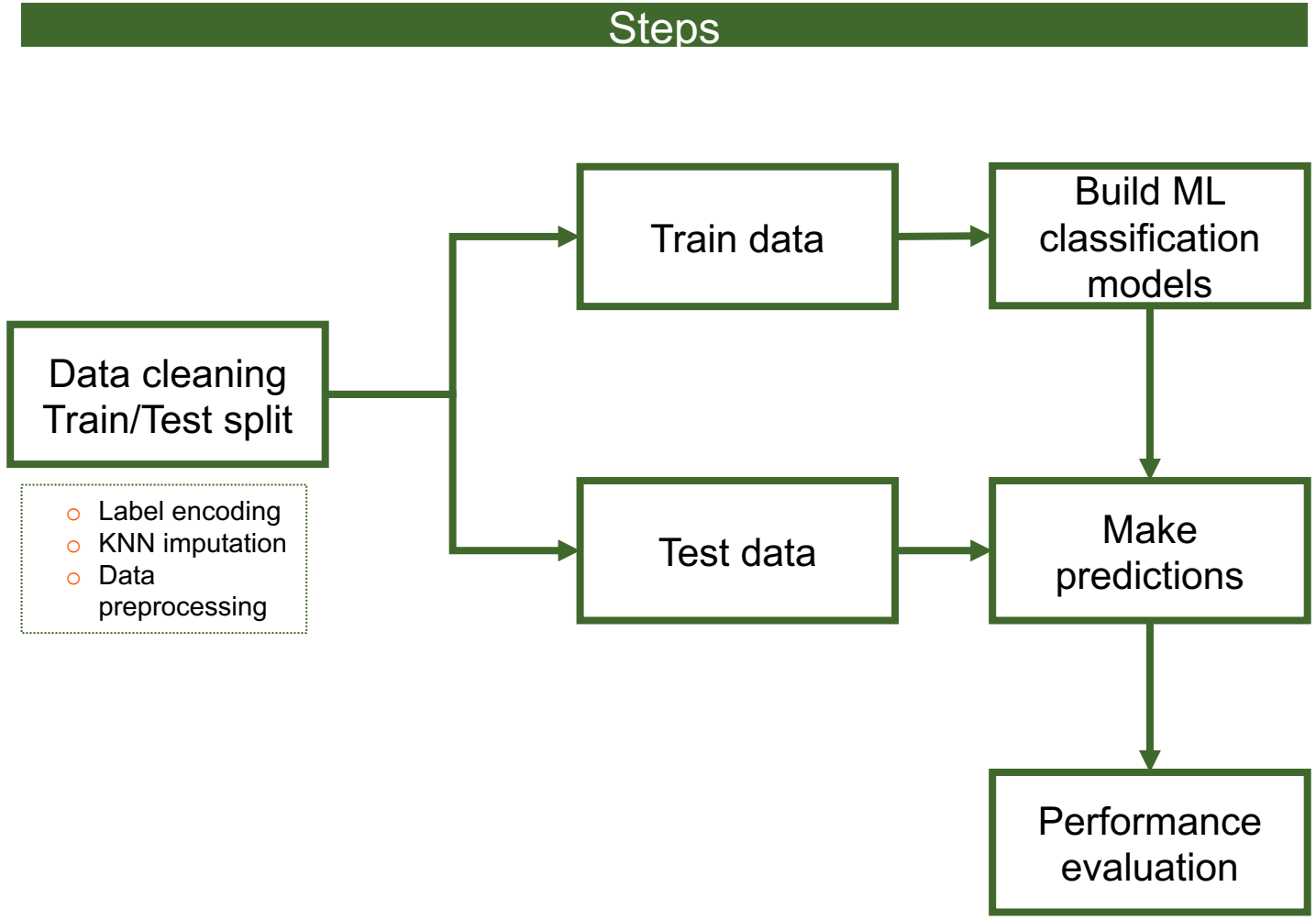
https://youtu.be/hYOZidi2hwY



Machine Learning (ML) Classification

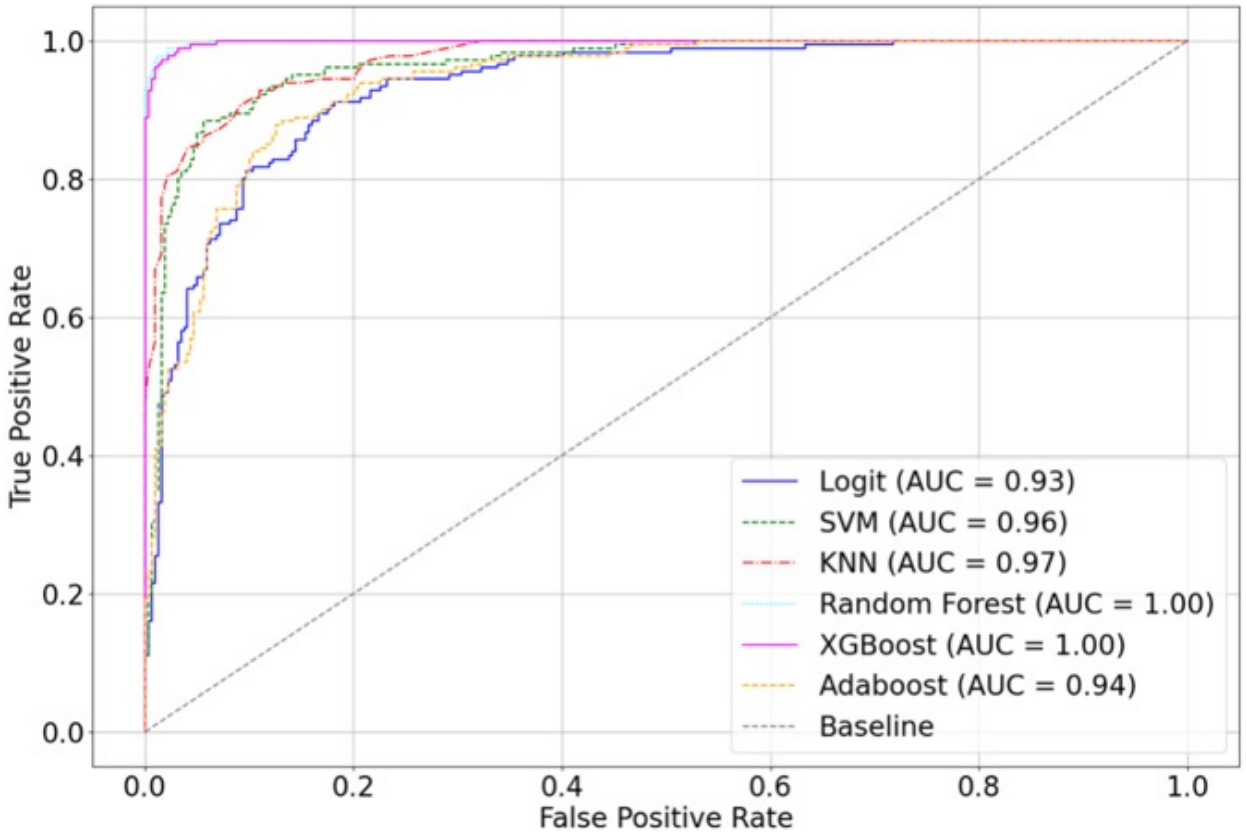
Traditional ML classification algorithms used to predict whether a product adheres to the ESG certification requirements for TCO and Energy Star:

- Logistic Regression (Log. Reg.)
- Support Vector Machine (SVM)
- K-Nearest Neighbor (KNN)
- Random Forest (Ran. For.)
- XGBoost
- AdaBoost



Model Evaluation: Traditional ML Energy Star

Receiver operating statistic (ROC) for six Energy Star classification methods



Performance metrics for six Energy Star classification methods

Method	Accuracy	Precision	Recall	Specificity
Log. Reg.	85.2%	75.8%	86.7%	84.3%
SVM	90.2%	83.0%	91.7%	89.3%
KNN	91.2%	90.1%	85.1%	94.7%
Ran. For.	**98.0%**	**98.3%**	**96.1%**	**99.1%**
XGBoost	97.6%	**98.3%**	95.0%	**99.1%**
AdaBoost	86.6%	83.1%	79.0%	90.9%

* Refer to Appendix for detailed explanations of these metrics

Discussion

Random Forest performed best across all performance metrics, misclassifying only 10 out of the 500 test products (98% accuracy). XGBoost also performed well (97.6% accuracy), while Logistic Regression and AdaBoost scored significantly worse (85.2% and 86.6% accuracy respectively.

Model Evaluation: Traditional ML TCO

Performance metrics for six TCO classification methods

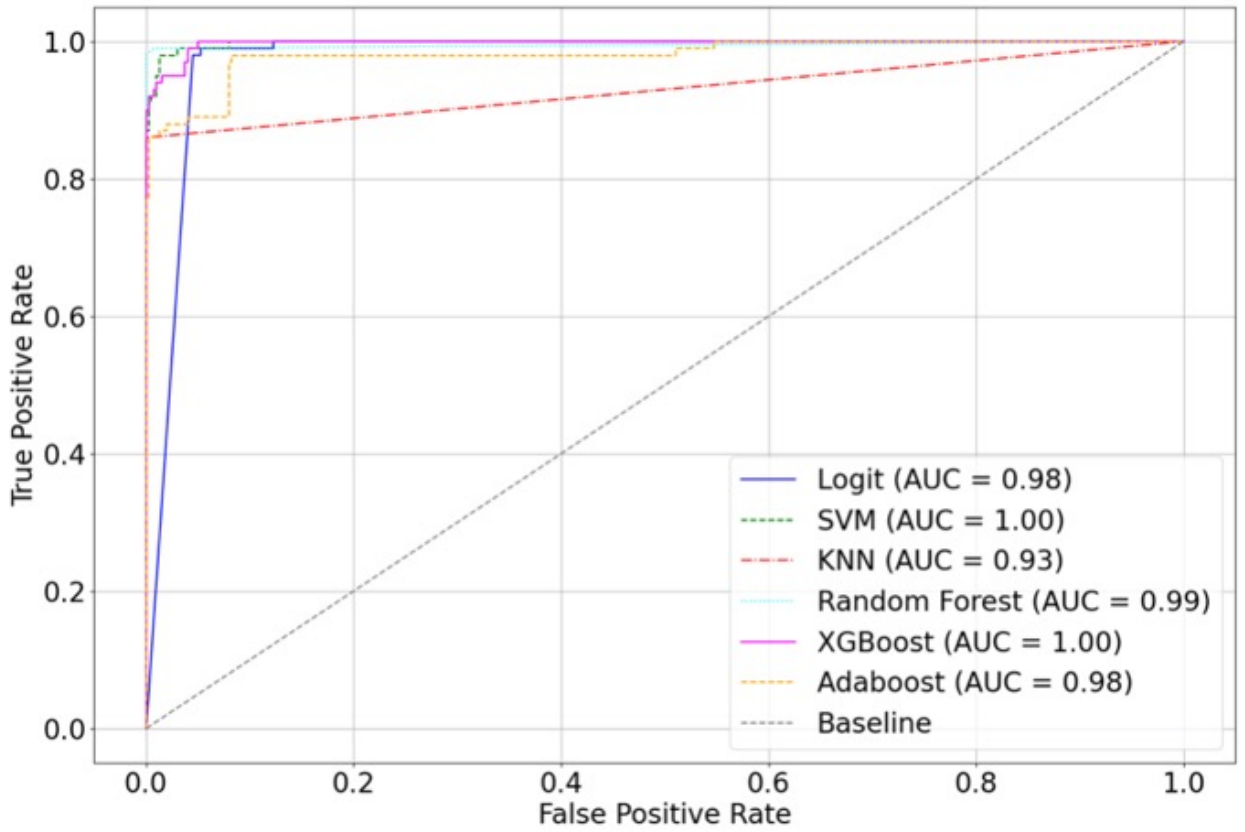
Method	Accuracy	Precision	Recall	Specificity
Log. Reg.	95.6%	83.1%	**98.0%**	95.0%
SVM	**97.8%**	97.8%	91.0%	99.5%
KNN	91.4%	**100.0%**	57.0%	**100.0%**
Ran. For.	92.0%	**100.0%**	60.0%	**100.0%**
XGBoost	93.4%	**100.0%**	67.0%	**100.0%**
AdaBoost	92.8%	**100.0%**	64.0%	**100.0%**

* Refer to Appendix for detailed explanations of these metrics

Discussion

SVM had the highest accuracy of around 97.8%, along with a relatively high precision, recall, and specificity scores compared to the other models. KNN, Random Forest, XGBoost, and AdaBoost displayed lower recall scores, indicating difficulty in accurately predicting positive classes due to the imbalanced nature of the dataset. Despite efforts to adjust the models to account for the imbalance, the issue remained challenging since only 0.07% of the training data had a TCO certification.

Receiver operating statistic (ROC) for six TCO classification methods



Model Evaluation: LLM Approach

GPT-3.5 ES

True \ Predicted	1	0
1	**86**	28
0	69	**45**

Accuracy	**57.5%**
Precision	55.5%
Recall	75.4%
Specificity	61.6%

Cohere ES

True \ Predicted	1	0
1	**2**	112
0	1	**113**

Accuracy	**50.4%**
Precision	66.7%
Recall	01.8%
Specificity	50.2%

Discussion

Cohere appears to be overly cautious at recommending a product for certification, predicting only 3 of the 228 products to be compliant. While not always correct, GPT-3.5 has a better mix of predictions (155 positive, 126 negative). The recall for GPT-3.5 is strong at 75.4%, meaning that for Energy Star certified products, GPT-3.5 is correct in its assessment 75.4% of the time.

GPT-3.5 TCO

True \ Predicted	1	0
1	**86**	5
0	49	**21**

Accuracy	**66.5%**
Precision	63.7%
Recall	94.5%
Specificity	80.8%

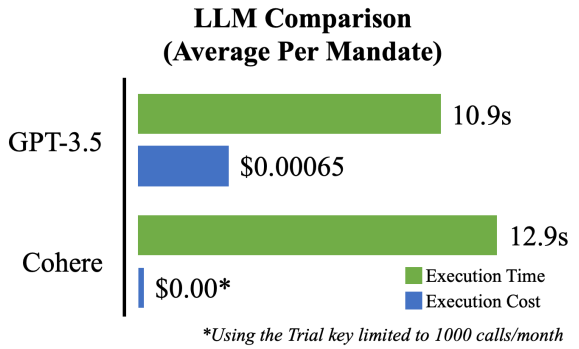
Cohere TCO

True \ Predicted	1	0
1	**39**	52
0	7	**63**

Accuracy	**63.4%**
Precision	84.8%
Recall	42.9%
Specificity	54.8%

Discussion

GPT-3.5 and Cohere both perform well, with GPT-3.5 performing slightly better. Cohere has a better precision score, so is more cautious to recommend certification once again. The recall score for GPT-3.5 is outstanding, misclassifying only 5 TCO certified products.



Time and cost for Cohere and GPT-3.5 per mandate.

Conclusion

- **Using the provided dataset, traditional ML classification algorithms are better at predicting product certifications than LLMs.** The worst performing ML algorithm, KNN, still scored a higher accuracy on identifying Energy Star products than GPT-3.5, the best performing LLM for TCO (**85.2%** vs **66.5%**)

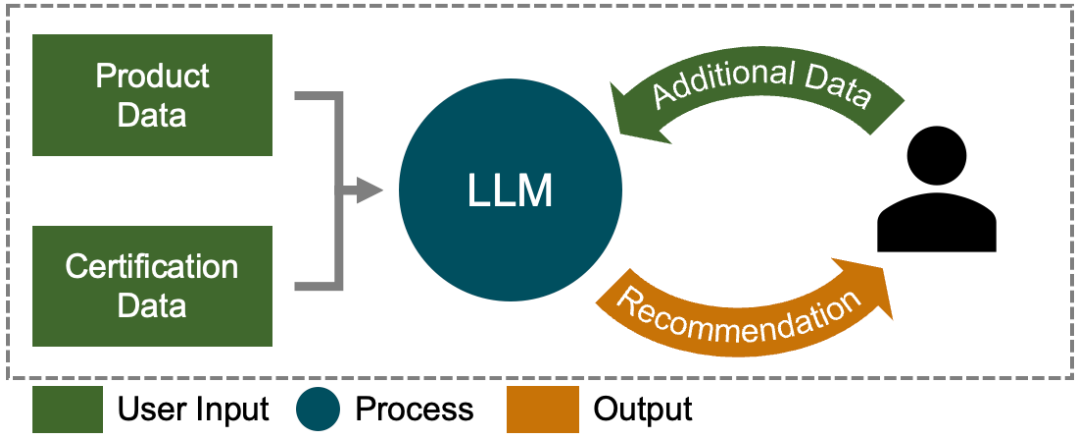
- Further analysis needed to understand whether the observed exceptional performance of the ML models can be generalized to other products, or if it is a result of our specific methodology and dataset. Continued approaches include:
  - Alternative forms of imputation
  - Cross-validation
  - More balanced data
  - Integration of additional data resources
  - Other machine learning algorithms

Conclusion Cont. and Next Steps

- ML algorithms lack clarity around the reason for classification and often require extensive data cleaning, LLMs address both drawbacks

- For clear and concise mandates where relevant data was available, the LLM approach allows for detailed and correct explanations around compliance

- For vague mandates where no data is available, the LLMs struggled
  - To improve the effectiveness of the LLM assessments, the user should be able to provide additional data and information about the products as necessary:



Updated recommendation engine architecture

Appendix

Confusion Matrix:
Table which compares the predicted labels (classes) to the actual labels (classes). By default, the confusion matrix assigns label 1 to the probabilities that are greater than or equal to 0.5 and label 0 otherwise:

		Predicted	
		1	0
True	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

Accuracy:
Overall accuracy of the model:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision:
Percentage of the predicted positive class which are actual positives:

$$Precision = \frac{TP}{TP + FP}$$

Recall:
Percentage of the actual positive class which were predicted positive by the model:

$$Recall = \frac{TP}{TP + FN}$$

Specificity:
Percentage of the actual negative class which were predicted negative by the model:

$$Specificity = \frac{TN}{TN + FP}$$

Receiver operating characteristic (ROC):
A graphical representation that demonstrates the performance of classification models across various thresholds, with false positive rate (1 – specificity) on the x axis and true positive rate (recall) on the y axis. Each point on the curve corresponds to a specific threshold's false positive and true positive rate.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html

Area under curve (AUC):
A performance metric that summarizes the overall performance of an ROC curve by calculating the area under the curve, representing the integral of the curve from 0 to 1 across all thresholds.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html

Appendix: Literature review

- Tutorial on Large Language Models for Recommendation
  - This paper examines the benefits of large language models over other recommendation techniques, including an LLM's ability to understand natural language.
- Leveraging Large Language Models in Conversational Recommender Systems
  - This paper examines the process of creating a functional large language model, showcasing how an LLM can take the context from available natural language and use external sources to generate an appropriate response.
- Training language models to follow instructions with human feedback
  - This paper examines the importance of human feedback in the creation and usage of large language models to ensure their output is representative of the desired outcome.

Appendix: Workload Distribution

Task	Description	Team Member Contributions
EDA: Product Data	Analyze and understand given data to find any meaningful insights, select product category to proceed with.	Jackson, Sofia
EDA: ESG Certifications	Analyze, understand, and consolidate technical requirements across ESG certifications.	Nathan
Data Cleaning / Transformation	Preparing ESG requirements, cleaning product dataset, summarize and encode product summary and certification requirements.	Nathan – Certifications Jackson, Sofia - Products
Methodology: Data Dictionary Creator	Execute proposed methodologies for the Data Dictionary Creator.	Jackson
Methodology: Recommendation Engine	Execute proposed methodologies for the Recommendation Engine.	Jackson, Sofia, Nathan
Analysis and Results	Evaluate performance of models using labeled data.	Jackson, Sofia, Nathan
Midterm Report	PowerPoint Presentation Slides	Jackson, Sofia, Nathan
Final Report	Final report document	Jackson, Sofia, Nathan

Key

Complete