AI CLASSIFICATION MODEL FOR FINANCIAL INDUSTRY COMPLAINTS

1. Refined Problem Statement

The financial industry handles a multitude of customer complaints daily, ranging from billing issues to allegations of fraud. The efficiency of addressing these complaints is affected by existing complaint management systems, which require significant manual input and are constrained by rigid, rule-based classification systems. These limitations lead to increased handling times, customer dissatisfaction, improper initial contact with pertinent parties, and potential regulatory risks. This project proposes an AI-driven solution that utilizes a pre-trained Large Language Model (LLM) to automate the classification and routing of complaints, thus improving response times and accuracy.

Key Decisions:

API Model: We opted to use the OpenAI API instead of a pretrained model to minimize initial capital expenditures and to leverage the advanced capabilities of cutting-edge models like GPT-4o-mini. Additionally, with an estimated volume of 10,000 complaints per month, the API cost would be approximately $14.00 (**see Exhibit 1**).

Prompt Engineering: We have opted for prompt engineering over fine-tuning due to the additional effort and costs associated with managing multiple categories in fine-tuning using API. Instead, we are focusing on refining prompts through instructions, format, length, and persona, while also evaluating the inclusion of context.

User Interface: We have narrowed the prototype scope to a chat interface that allows users to enter their issues and receive an acknowledgment indicating the product team to which the issue has been assigned. Subsequently, we will generate a ticket number and aim to integrate an email notification for the assigned team**(1)**.

Literature Review

IJCSRR Study, Enhancing Customer Service with AI (2) : This study employs DistilBERT for intent classification in the banking sector, utilizing a training model to assess its performance against traditional natural language processing (NLP) methods, including Naive Bayes, Logistic Regression, and Support Vector Machines. By leveraging the Banking77 dataset, the research evaluates DistilBERT's effectiveness **(3)** using key metrics such as accuracy, precision, recall, and F1 score, demonstrating its superiority over conventional models. The findings highlight DistilBERT's capability to achieve high performance in classifying customer inquiries, underscoring its potential as a robust tool for enhancing customer service in the financial industry.

Health insurance, Smart claims management (4): McKinsey explores the application of artificial intelligence (AI) in health insurance, focusing on the use of self-learning software to enhance claims management. By implementing cognitive systems that emulate human decision-making, the research demonstrates how AI can systematically identify and correct errors in hospital claims, thereby reducing administrative costs and improving efficiency. The study emphasizes the potential of AI to transform traditional claims processes, offering a more accurate and cost-effective approach to managing health insurance claims.

Lowering operational costs (5). Aon investigates the current drivers of claims quality and explores the role of artificial intelligence in transforming claims management across various industries, with a particular emphasis on insurance. It highlights how AI technologies, such as machine learning and predictive analytics, can optimize claims processing by enhancing the accuracy and speed of the initial contact, investigation, and settlement phases. AI can reduce manual tasks, mitigate errors, and ultimately lower operational costs while improving customer satisfaction (see exhibit 2). The piece underscores the strategic importance of integrating AI to address emerging risks and align with future business needs

2. Initial Solution Design

High-level Solution Overview

Our solution architecture is designed to integrate a pre-trained large language model (LLM) via the OpenAI API with the existing infrastructure of financial institutions to classify customer complaints across various categories of Products (e.g. debt collection, credit card, checking accounts) and Issues (e.g. incorrect information on your report, improper use of your report, attempts to collect debt not owed, and account management). Reflecting on the user experience (UX) of the solution, we have defined a system that utilizes a chat interface, allowing users to input their issues. In addition to this, these issues will then be categorized and assigned a ticket for further processing by the resolution team. This integration will automate routing of complaints to the appropriate departments via email**(6)(7)**, significantly reducing processing time.

Model Selection and Justification

We selected GPT-4o-mini for its proven capability in understanding and generating human-like text. GPT-4's extensive training on diverse datasets makes it appropriate for understanding the language often found in customer complaints, which is crucial for accurate classification.

Data Alignment and Usage

The project will utilize data from the US Consumer Financial Protection Bureau, which includes a wide variety of complaint types. This dataset serves as a reference to identify the industry-standard categories, which are used in the prompt instructions for the model. The model will classify complaints by type of product and type of issue. This ensures the model is equipped to recognize and categorize the diverse complaints it will encounter

Manage Flow of Classification

To enhance the accuracy of issue classification, we will use LangChain's sequential chains. The first chain will categorize the reported product and issue, while the second chain will evaluate and direct the issue to the appropriate resolution team. This structured approach allows for efficient data flow and ensures that each step informs the next, streamlining the overall classification process.

Technology Stack

The user interface will be developed using Streamlit to allow users to submit complaints easily. Backend operations, including classification and routing tasks, will be managed through LangChain while leveraging the OpenAI API "gpt-4." Additionally, we will utilize either the

SMTP library or the Gmail API to connect to an email server, enabling us to create notifications and inform the responsible team when a new issue has been assigned to them.

Prototyping and Initial Tests

A preliminary prototype will be presented on Google Colab to demonstrate the complaint submission process and the subsequent classification. Migrating it to Streamlit and the routing will be addressed for the next milestone. Initial tests will focus on evaluating the model's accuracy using a subset of the dataset.

Complaint Datasets

The project will utilize a dataset **(8) (9)** of customer complaints from the financial sector, containing labeled data across various complaint types. This dataset will help to evaluate the model accuracy classifying complaints based on natural language. We will use a pre-existing dataset from the Consumer Financial Protection Bureau's (CFPB) complaint database, ensuring coverage of diverse complaint types such as fraud, regulatory issues, billing disputes, and more.

3. Prototype or Development Progress

Core Components the model Setup. These are the main components of the setup:

Data sourcing and preprocessing. The implemented functionality begins with the acquisition of data, followed by cleaning and analysis. The data is downloaded in a compressed format, extracted, and read into a DataFrame using Python's pandas library. The preprocessing focuses on cleaning the dataset, removing unnecessary columns and missing values. Visual tools like missingno were used to examine and handle missing values, ensuring a dataset that's consistent and ready for modeling.

Exploratory analysis generated insights into complaint distributions, allowing us to understand the different classifications. A summary of "Product" and "Sub-product" categories helped visualize which issues were most prevalent, providing valuable information for the subsequent tasks.

Defining LLM model and doing initial classifications. Once the dataset cleaning and exploratory phases were completed, the next phase involved integrating OpenAI's pre-trained LLMs to classify complaints. Initially, the gpt-4o-mini model is integrated to classify complaints at a single level, such as "Sub-product." Then, a multi-step classification system is developed to classify complaints hierarchically: first by "Product," followed by "Sub-product," and finally by "Issue."

Refining Outcomes with Model Parameters and Prompt Definition. We refined our approach by adjusting the prompt instructions, format, length, and tone to clearly define the specific categories the model should use and how to present the results. This added context significantly improved the accuracy of the outcomes. We tested the classification in batches of 100 complaints at a time and visually tracked the improvements. Additionally, we experimented with the model's temperature settings, settling on a lower value to ensure consistency, which resulted in a 5% improvement in output compared to higher temperature settings.

Evaluating model performance using metrics such as accuracy from scikit-learn offers insights into the effectiveness of the model in classifying complaints. We have implemented various versions of the classification model, incorporating prompt enhancements and filtered options to boost accuracy at each level. This functionality establishes a foundation for further refinement and more precise complaint categorization. Additionally, we conducted multiple iterations of the output to ensure result consistency and confirm similar outcomes.

Technical Challenges:

One initial challenge encountered while identifying a suitable model was the attempt to use and download a pretrained model of the ones available in HuggingFace for further fine-tuning. Although there are multiple models available, given the complexity of managing multiple classification outputs, such as issue, sub-issue, product, and sub-product, and the significant amount of processing required for diverse inputs, this task was more challenging than expected. Unlike training a model for text generation, the intricacies of multiple categories added complexity.

Another challenge was the relatively low accuracy of the model when categorizing issues at the "Issue" level. This low accuracy can be attributed to the high similarity between different issues, which made it difficult for the LLM to distinguish between them effectively. Moreover, the LLM sometimes struggled to align with exact labels required for classification, as these labels were nuanced, and the prompt engineering needed continuous refinement. This complexity in prompt engineering and model calibration affected the accuracy of classification, especially for issues that were similar in nature but had subtle differences in their descriptions. As a result, further work is required to improve prompt specificity and ensure better differentiation between closely related complaint categories. See exhibit 3 for more details on accuracy.

Planned actions to address these challenges:

- **Refine prompts:** Iteratively refined prompts by testing various levels of specificity and context inclusion. In addition, we'll include specific definitions for the issues in order to provide more context to the model. This approach will help the model classify nuanced complaint language more accurately.
- **Address overlapping complaint categories**: Considering that many complaint issues might have overlapping meanings, such as "Incorrect information" vs "Improper use of report," which can make it difficult for the model to assign a single accurate label. We will explore introducing a hierarchical classification within the issue level itself, where related issues are grouped under broader umbrella categories. This might allow the model to identify issues at a more general level, and improve the accuracy.

GitHub: https://github.com/rjcontrerasr/IS883-LLM_Project-Financial_complaints/blob/main/IS883_G5_LLM_Financial_Classification.ipynb

4. Evaluation Plan

To measure the effectiveness of the financial complaints classification solution, we will use both quantitative and qualitative metrics. Quantitative metrics include accuracy, which will measure the correctness of complaint categorization across different levels, such as issue, product and

subproduct. We will also evaluate the consistency of the outputs of the model to compare the accuracy across different evaluations (see Exhibit 3 for an estimation with the current prototype).

Time savings for processing time will also be evaluated. These will measure the efficiency of the model by comparing the classification time against human effort (e.g., processing 10,000 complaints monthly).

Qualitative metrics will involve user feedback collected through surveys, which will be administered after interactions to assess final user satisfaction and determine whether the process was efficient according to the customer's opinion. Combining these qualitative metrics with quantitative evaluations will provide a comprehensive view of the model's effectiveness, focusing on both operational gains and user experience.

5. Updated Timeline and Next Steps

The initial stages, including data gathering, cleaning, and initial prototype development, were completed as expected. The next steps involve continuing to improve the model, designing the chatbot workflow, setting up classification and notification chains, building the user interface, and preparing for deployment and user testing. For further details, please refer to the Jira project timeline at:

https://jmhu.atlassian.net/jira/core/projects/ACMFFIC/timeline[1]

6. Presentation of Work

GitHub: https://github.com/rjcontrerasr/IS883-LLM_Project-Financial_complaints/blob/main/IS883_G5_LLM_Financial_Classification.ipynb

[1] Please review your email (elhamod@bu.edu) to accept the invitation as a team member of the project.

Exhibit 1

On average, each complaint contains approximately 200 tokens, and the prompt instructions another 200 tokens. With our current model, we assume handling over 10,000 complaints per month, averaging 4M tokens, for a total cost of $10 per 4M input tokens **(10)**. Similarly, for output tokens, we average about 40 tokens for product and issue categorization. This would result in a total output cost of $4 for those 10,000 complaints, assuming we can deduct it from the $10 charge per 1 million output tokens.

Model	Pricing	Pricing with Batch API*
gpt-4o	$2.50 / 1M input tokens	$1.25 / 1M input tokens
	$1.25 / 1M cached** input tokens	
	$10.00 / 1M output tokens	$5.00 / 1M output tokens

ref: https://openai.com/api/pricing/

Exhibit 2

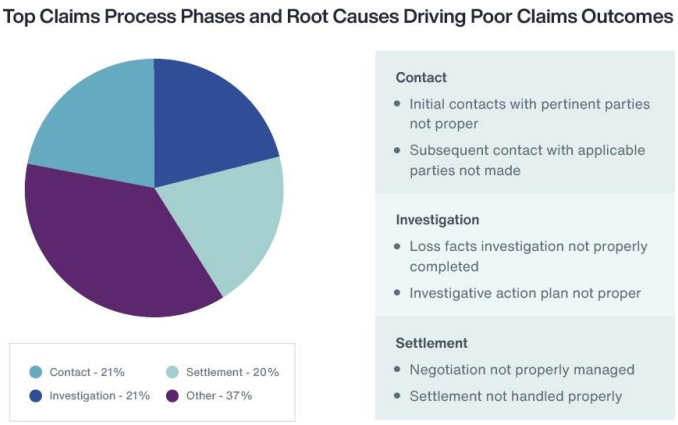

Exhibit 3

Performance across different modeling approaches

		Iteration 1			Iteration 2			Iteration 3		
Model Attempt	**Description**	**Product Accuracy**	**Sub-product Accuracy**	**Issue Accuracy**	**Product Accuracy**	**Sub-product Accuracy**	**Issue Accuracy**	**Product Accuracy**	**Sub-product Accuracy**	**Issue Accuracy**
Sub-product Classification	Classifies complaints into their respective sub-product	N/A	0.84	N/A	N/A	0.84	N/A	N/A	0.8	N/A
Issue Classification	Classifies customer complaints directly by issue	N/A	N/A	0.16	N/A	N/A	0.16	N/A	N/A	0.12
Sub-product and Issue Classification	Sequential classification into sub-product and then issue categories	N/A	0.85	0.12	N/A	0.85	0.12	N/A	0.79	0.3
Product, Sub-product, and Issue Classification	Hierarchical classification into product, sub-product, and issue categories	0.85	0.83	0.1	0.85	0.85	0.1	0.85	0.79	0.34
Filtered Hierarchical Classification	Hierarchical classification with filtered categories and prompt improvements	0.85	0.84	0.4	0.85	0.83	0.4	0.85	0.8	0.43

References:

(1) Stack Overflow**.** (n.d.). *Sending mail from Python using SMTP*. Retrieved from https://stackoverflow.com/questions/64505/sending-mail-from-python-using-smtp

(2) Kumar, S., Deep, S., & Kalra, P. (2024). *Enhancing customer service in banking with AI: Intent classification using DistilBERT*. International Journal of Current Scientific Research and Review. Retrieved from https://ijcsrr.org/wp-content/uploads/2024/05/32-1305-2024.pdf

(3) Hugging Face**.** (n.d.). *DistilBERT model documentation*. Retrieved from https://huggingface.co/docs/transformers/en/model_doc/distilbert

(4) McKinsey & Company. *Artificial intelligence in health insurance: Smart claims management with self-learning software*. https://www.mckinsey.com/industries/healthcare/our-insights/artificial-intelligence-in-health-insurance-smart-claims-management-with-self-learning-software

(5) Aon. (2024). *5 ways artificial intelligence can boost claims management*. https://www.aon.com/en/insights/articles/5-ways-artificial-intelligence-can-boost-claims-management#:~:text=Examples%20of%20AI%20use%20in,%2Dto%2Dend%20claims%20processing.

(6) Google Developers**.** (n.d.). *Gmail API guides*. Retrieved from https://developers.google.com/gmail/api/guides

(7) Python Software Foundation**.** (n.d.). *SMTP library*. Retrieved from https://docs.python.org/3/library/smtplib.html

(8) Consumer Financial Protection Bureau. (2024). *Consumer complaints search (from April 5, 2024, through October 5, 2024)*. Retrieved from https://www.consumerfinance.gov/data-research/consumercomplaints/search/?date_received_max=2024-10-05&date_received_min=2024-04-05&page=1&searchField=all&size=25&sort=created_date_desc&tab=List

(9) Consumer Financial Protection Bureau. (2023). *Consumer complaints search (from October 5, 2022, through April 4, 2023)*. Retrieved from https://www.consumerfinance.gov/data-research/consumercomplaints/search/?date_received_max=2023-04-04&date_received_min=2022-10-05&page=1&searchField=all&size=25&sort=created_date_desc&tab=List

(10) OpenAI**.** (n.d.). *API pricing*. Retrieved from https://openai.com/api/pricing/