Article

DeepSeek and GPT Fall Behind: Claude Leads in Zero-Shot Consumer Complaints Classification

Konstantinos I. Roumeliotis *, Nikolaos D. Tselikas, Dimitrios K. Nasiopoulos

Posted Date: 10 February 2025

doi: 10.20944/preprints202502.0720.v1

Keywords: DeepSeek Chat; DeepSeek-V3; DeepSeek LLM; Claude Sonnet; GPT-4o; Consumer Complaints; Consumer Complaints Classification; Zero-Shot Classification

Article

DeepSeek and GPT Fall Behind: Claude Leads in Zero-Shot Consumer Complaints Classification

Konstantinos I. Roumeliotis [1,2,*], Nikolaos D. Tselikas [1] and Dimitrios K. Nasiopoulos [3]

[1] Department of Informatics and Telecommunications, University of the Peloponnese, 22131 Tripoli, Greece
[2] Department of Digital Systems, University of the Peloponnese, 23100 Sparta, Greece
[3] Department of Agribusiness and Supply Chain Management, School of Applied Economics and Social Sciences, Agricultural University of Athens, 11855 Athens, Greece
* Correspondence: k.roumeliotis@uop.gr

Abstract: Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing (NLP) tasks, but their effectiveness in real-world consumer complaint classification without fine-tuning remains uncertain. Zero-shot classification is particularly challenging in finance, where complaint categories often overlap, requiring a deep understanding of nuanced language. In this study, we evaluate the zero-shot classification performance of leading LLMs— DeepSeek-V3, OpenAI's GPT-4o and GPT-4o mini, and Anthropic's Claude 3.5 Sonnet and Claude 3.5 Haiku—on consumer complaints submitted to the Consumer Financial Protection Bureau (CFPB). These models were tasked with categorizing complaints into five predefined financial classes based solely on complaint text. Performance was measured using accuracy, precision, recall, F1-score, and heatmaps to identify classification patterns. While DeepSeek Chat and GPT-4o produced competitive results, Claude 3.5 Sonnet consistently outperformed all models, demonstrating superior classification accuracy and efficiency. These findings highlight the relative strengths and limitations of DeepSeek-V3 and other top-tier models in financial text processing, providing valuable insights into their practical applications.

Keywords: DeepSeek chat; DeepSeek-V3; DeepSeek LLM; Claude Sonnet; GPT-4o; consumer complaints; consumer complaints classification; zero-shot classification

1. Introduction

Understanding consumer complaints is crucial for businesses, regulators, and financial institutions. Effective complaint management not only improves customer satisfaction but also helps organizations identify systemic issues, ensure compliance, and enhance decision-making [1]. The Consumer Financial Protection Bureau (CFPB) serves as a key regulatory body in the U.S., collecting consumer complaints related to financial products and services [2]. These complaints, submitted in free-text narratives, contain critical insights but require accurate classification to be processed efficiently. The challenge lies in categorizing these complaints into well-defined financial classes, as many cases involve overlapping categories and nuanced financial terminology [3].

The classification of consumer complaints is a demanding natural language processing (NLP) task due to the complex, unstructured nature of financial narratives [4]. Traditionally, NLP models have been employed to automate and enhance complaint classification, improving efficiency over manual tagging [3]. Earlier studies have explored various machine learning (ML) and deep learning (DL) approaches, including fine-tuned models trained on financial datasets [5]. However, as large language models (LLMs) continue to advance, it becomes imperative to evaluate their ability to perform this classification task in a zero-shot setting—without any additional training or fine-tuning [6].

In this study, we explore the capabilities of state-of-the-art LLMs for zero-shot consumer complaint classification. Specifically, we evaluate DeepSeek Chat, OpenAI's GPT-4o and GPT-4o mini, and Anthropic's Claude 3.5 Sonnet and Claude 3.5 Haiku. These models are among the most powerful and widely discussed in the AI landscape, known for their advanced reasoning, language understanding, and text classification capabilities [7]. The objective is to determine how well DeepSeek-V3, GPT-4o, and Claude 3.5 Sonnet perform when tasked with classifying CFPB complaints into five predefined financial categories without any task-specific fine-tuning.

To assess the zero-shot classification performance of these models, we analyze consumer complaints submitted to the CFPB. Each complaint is categorized into one of five financial classes based solely on the complaint text. The models— DeepSeek Chat [8], GPT-4o [9], GPT-4o mini, Claude 3.5 Sonnet [10], and Claude 3.5 Haiku—are evaluated using key performance metrics, including accuracy, precision, recall, and F1-score. Additionally, we generate heatmaps to visualize classification patterns and identify strengths and weaknesses across different models.

Our primary research question is: How do top-tier LLMs, including DeepSeek-V3, GPT-4o, and Claude 3.5 Sonnet, perform in zero-shot consumer complaint classification, and which model achieves the highest accuracy in categorizing financial disputes? By addressing this question, we aim to provide insights into the practical application of state-of-the-art LLMs in financial text classification and identify the most effective model for this critical NLP task.

2. Literature Review

Consumer complaints are prevalent across various sectors, including financial services, e-commerce, transportation, and environmental concerns. Efficiently managing these complaints is crucial for businesses to enhance user satisfaction, build trust, and retain customers [11]. As digital platforms become the primary medium for consumer interactions, organizations must process vast complaint volumes effectively. Manual handling is often impractical due to sheer volume, making automated classification and prioritization essential. Complaint classification categorizes grievances into structured groups, streamlining customer service operations and aiding in sentiment analysis to improve service quality [12].

2.1. Automated Complaint Classification and Prioritization

Vairetti et al. (2024) [3] propose a DL and multi-criteria decision-making (MCDM) framework for complaint prioritization. Their study emphasizes efficient categorization to improve customer satisfaction (CSAT) and reduce churn rates. Using a BERT-based model, BETO, they achieve a 92.1% accuracy rate, showcasing the effectiveness of modern NLP models in complaint classification.

Roy et al. (2024) [5] explore complaint classification in the railway sector, highlighting the role of social media in customer feedback. Their system employs ML models such as Random Forest and Support Vector Machines (SVM) to classify tweets based on urgency, ensuring timely responses to critical complaints. Sentiment analysis plays a key role in identifying negative tweets that require immediate attention.

2.2. Sentiment and Emotion Analysis in Complaints

Kumar et al. (2024) [13] examine consumer complaints in multilingual e-commerce settings, integrating sentiment, emotion, and severity analysis. Their hierarchical attention-based DL model processes complaints at word and sentence levels, outperforming benchmark models and emphasizing the importance of emotional and severity considerations in complaint handling.

Das et al. (2024) [14] focus on financial complaints, proposing an explainable AI approach to differentiate between negative reviews and formal complaints. Their dyadic attention-based model facilitates sentiment detection, emotion recognition, and severity classification, offering a comprehensive understanding of customer dissatisfaction in financial services.

2.3. Machine Learning and AI for Complaint Processing

Jondhale et al. (2024) [15] develop an ML-based pipeline for predicting disputed financial complaints. Utilizing PySpark ML and feature engineering, their system enhances complaint classification accuracy and supports proactive dispute resolution. Correa et al. (2024) extend the role of generative AI in consumer complaint handling, integrating classification, summarization, and response generation, achieving an 88% classification accuracy and demonstrating AI's potential in customer service operations.

Djahongir Ismailbekovich (2024) [16] investigates AI-driven chatbots for consumer complaints, noting their efficiency benefits. However, challenges such as bias, lack of nuance interpretation, and regulatory compliance necessitate a hybrid approach that combines AI automation with human oversight.

2.4. Broader Implications of Consumer Complaints

Zhou et al. (2024) [17] analyze environmental complaints in China, illustrating how consumer feedback influences public policy and environmental quality. Their study underscores the role of complaint reporting in pollution control efforts, broadening the scope of complaint management applications.

Sharma et al. (2024) [4] introduce a multimodal NLP feedback system that enables silent feedback collection in shopping malls. Their model processes audio feedback to extract customer sentiments, reducing communication barriers and improving feedback quality.

2.5. Topic Modeling and Sentiment Analysis

Khadija et al. (2024) [18] apply Latent Dirichlet Allocation (LDA) combined with BERT embeddings to analyze Indonesian customer complaints. Their findings suggest that BERT-based models enhance LDA's topic coherence and silhouette scores, making them effective for extracting meaningful topics from short-text complaints, thus aiding businesses in structured complaint categorization.

Song et al. (2024) [19] propose a textual analysis framework integrating guided LDA and sentiment polarity for quantifying service failure risks. By incorporating CRITIC and TOPSIS methodologies, their approach enhances traditional Failure Mode and Effects Analysis (FMEA), offering a data-driven risk assessment method. Applied to the hotel industry, their findings improve accuracy in identifying and prioritizing service failure risks.

2.6. Deep Learning and AI-Driven Monitoring

Seok et al. (2024) [20] introduce a DL-based customer complaint monitoring system using explainable AI (XAI) techniques. Their approach integrates BERT-based models to extract service-related features from online reviews and analyze sentiment. Their study employs a staged p-chart for continuous monitoring of complaints in seasonal industries, addressing limitations in traditional NLP techniques.

Jia et al. (2024) [21] explore generative AI (GAI) in customer complaint resolution, comparing AI-generated responses with human-authored ones. Their research highlights GAI's potential for automating complaint responses while maintaining empathy and coherence, emphasizing the need for balancing AI integration with human oversight to ensure authentic interactions.

2.7. Consumer Behavior in Complaints

Lee et al. (2024) [22] develop the "3D" model for temperament-centered complaints. Their e-Customer Complaint Handling (e-CCH) system collects, processes, and classifies complaints based on consumer temperaments, providing personalized solutions. Their open-sourced dataset offers valuable insights for industrial service management and complaint handling research.

Wang et al. (2024) [23] analyze consumer complaint behavior in live-streaming e-commerce using a two-staged SEM-ANN approach. Their study identifies key complaint factors, including consumer confusion, emotional venting, and altruistic motives. The research highlights the impact of group complaints on consumer behavior and suggests strategies for e-commerce platforms to mitigate negative experiences.

2.8. AI-Powered Chatbots for Complaint Resolution

Juipa et al. (2024) [24] introduce a sentiment analysis-based chatbot for telecommunications complaint management. Their study demonstrates that incorporating GPT-3.5 for sentiment analysis improves complaint resolution efficiency and customer satisfaction. The chatbot achieves an 86% satisfaction rate, exceeding industry standards and supporting AI-driven chatbots in enhancing complaint-handling processes.

Overall, the literature underscores the growing reliance on AI, ML, and NLP for consumer complaint classification and sentiment analysis. These advancements have significantly improved accuracy, efficiency, and scalability in handling consumer grievances across industries.

3. Materials and Methods

This section outlines the systematic approach employed to explore and evaluate the capabilities of DeepSeek-V3, GPT-4o, and Claude 3.5 Sonnet in the challenging task of zero-shot consumer complaint classification. Given the complexity of financial complaints—where categories often overlap and subtle linguistic nuances influence classification—this study follows a structured methodology to ensure a fair and comprehensive assessment of each model's performance.

We begin by describing the dataset used in our analysis, which consists of consumer complaints submitted to the CFPB [2]. Since these complaints are written in free-text form, preprocessing is a crucial step to standardize and prepare the data for classification. In the preprocessing steps, we detail the cleaning procedures applied to the complaint narratives to improve input consistency across all models.

Next, we discuss prompt engineering, a critical component of zero-shot classification. As these models have not been fine-tuned for this specific task, the prompt structure plays a significant role in guiding the models toward accurate classifications. We describe the design and refinement of prompts tailored to extract the best possible performance from each model.

Finally, we present our model evaluation strategy, where we define the metrics used to assess classification accuracy. The performance of DeepSeek Chat, GPT-4o, and Claude 3.5 Sonnet is measured using standard classification metrics, including accuracy, precision, recall, and F1-score. Additionally, heatmaps are employed to analyze misclassification patterns and compare trends across models.

By implementing a rigorous methodology, this study aims to provide a data-driven comparison of DeepSeek-V3, GPT-4o, and Claude 3.5 Sonnet in real-world consumer complaint classification, offering valuable insights into their effectiveness for financial text processing.

3.1. Description of the Dataset

The dataset used in this study is the Consumer Complaints Dataset for NLP, curated by Shashwat Tiwari and hosted on Kaggle [25]. This dataset is derived from consumer complaints originally sourced from the CFPB website [2]. It contains consumer-submitted complaints about financial products and services, making it a highly relevant resource for evaluating LLMs in the demanding task of financial text classification.

The dataset includes one year's worth of consumer complaints, covering the period from March 2020 to March 2021. Additionally, the dataset creator supplemented the complaints by utilizing the CFPB's API to fetch up-to-the-minute submissions, ensuring a mix of historical and more recent complaints. Each complaint is associated with one of nine original financial product categories, but due to similarities between certain classes and class imbalances, the dataset has been consolidated into five broader financial categories Credit Reporting, Debt Collection, Mortgages and Loans (includes car loans, payday loans, student loans, etc.), Credit Cards, and Retail Banking (includes checking/savings accounts, money transfers, Venmo, etc.).

The dataset consists of approximately 162,400 consumer complaints, each containing a free-text narrative describing the consumer's financial issue. The text length varies significantly, with an average length of 588.49 characters and a maximum length of 20,596 characters, posing a challenge for classification models.

Additionally, the dataset is highly imbalanced, with the following distribution:

- Credit Reporting: 56.14%
- Debt Collection: 14.25%
- Mortgages and Loans: 11.69%
- Credit Cards: 9.58%
- Retail Banking: 8.33%

Given its 9.41 usability score on Kaggle, this dataset is widely regarded as a valuable benchmark for consumer complaint classification using NLP. By leveraging this dataset, we aim to evaluate the zero-shot classification capabilities of DeepSeek Chat, GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku, assessing their ability to process real-world financial complaints efficiently and accurately.

3.2. Preprocessing Steps

To ensure a high-quality dataset for evaluating DeepSeek-V3, GPT-4o, and Claude 3.5 Sonnet in zero-shot consumer complaint classification, several preprocessing steps were applied. These steps aimed to improve data consistency, address class imbalance, and optimize the dataset for model evaluation. Given the complexity and overlapping nature of financial complaints, preprocessing was crucial in ensuring fair and meaningful model comparisons.

3.2.1. Removing Excessively Long Narratives

The dataset includes consumer complaints of varying lengths, with some narratives reaching 20,596 characters. To manage computational efficiency and avoid outliers affecting classification results, complaints exceeding 500 characters were removed. This step ensured that all models, including DeepSeek-V3, GPT-4o, and Claude 3.5 Sonnet, processed narratives of a more typical length, aligning with real-world classification constraints.

3.2.2. Standardizing Narrative Text

To enhance textual consistency, the narrative column underwent standardization, including:

- Converting text to lowercase
- Removing special characters, excessive whitespace, and inconsistent formatting
- Normalizing text structure to ensure cleaner inputs

This standardization prevented inconsistencies that could impact classification accuracy and helped models better understand complaint content.

3.2.3. Removing Entries with Empty Fields

Some complaints were missing key information in critical columns such as "product" and "narrative". These incomplete entries were removed to maintain data integrity and ensure that each record contained a valid complaint description and category. This step eliminated potential biases arising from incomplete data affecting model predictions.

3.2.4. Creating a Balanced Subset Using Undersampling

The original dataset was highly imbalanced, with credit reporting complaints making up over 56% of submissions, while retail banking complaints accounted for only 8.33%. To create a balanced evaluation set, an undersampling technique was applied, selecting a stratified subset of 1,000 records. The resulting dataset ensured equal representation across all five complaint categories, with:

- 200 complaints per category
- 20.00% representation per category

This balanced subset allowed for an unbiased assessment of models in zero-shot classification, ensuring that no single category disproportionately influenced model performance.

Through these preprocessing steps, the dataset was refined to provide clean, standardized, and balanced inputs, allowing for a robust comparison of DeepSeek Chat, GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku in financial complaint classification.

3.3. Prompt Engineering

Crafting effective prompts is essential for optimizing the performance of LLMs, particularly in zero-shot classification tasks where the model hasn't been specifically fine-tuned. This technique involves structuring input queries in a way that guides the model to generate precise and relevant responses. The effectiveness of a prompt largely determines how well the model understands the task and produces accurate results, making prompt design a crucial skill for leveraging LLMs efficiently.

In this research, prompt engineering played a key role in enabling models to classify consumer complaints into predefined categories based only on the complaint text, without prior exposure to those classifications. Since LLMs are inherently flexible, developing effective prompts required an iterative process, where refinements were made to enhance clarity, reduce ambiguity, and improve classification accuracy. Elements such as the specificity of instructions, data presentation format, and overall structure were carefully adjusted to optimize real-world application [26].

To establish an effective prompt, we conducted a trial-and-error process using the GPT model's chat interface. Through multiple refinements, we devised a structured prompt that improved interpretability. Given that structured input generally enhances LLM performance, we initially experimented with XML formatting to organize the information. While this approach made the data more accessible to the model, it also increased the token count, making API interactions more resource-intensive and expensive.

To address these limitations, we leveraged Anthropic's console prompt generator—a tool designed to help users craft effective prompts by providing pre-built templates based on best practices [27]. Our experimentation revealed that JSON formatting was a more efficient alternative to XML, as it streamlined the prompt structure and reduced token usage, making the process more cost-effective.

The final prompt was carefully refined to strike a balance between clarity and efficiency, ensuring it was universally applicable across different LLMs while minimizing computational overhead. **Listing 1** illustrates an example of the optimized prompt. This structured approach allowed us to enhance model performance while effectively managing API constraints.

Listing 1. Model-agnostic prompt.

```
conversation.append({
        'role': 'user',
        'content':
        'You are an AI assistant specializing in consumer complaint classification.'
        'Your task is to analyze a consumer complaint and classify it into the most'
        'appropriate category from the predefined list:'                                  (1)
        '["retail_banking", "credit_reporting", "credit_card", "mortgages_and_loans", "debt_collection"]'
        'Provide your final classification in the following JSON format without explanations:'
        '{"product": "chosen_category_name"}.\nComplaint: '
        '...'
})
```

3.4. Model Predictions

For this study, we utilized the gpt-4o, gpt-4o-mini, claude-3-5-sonnet-20241022, claude-3-5-haiku-20241022 and deepseek-chat models to perform zero-shot classification of consumer complaints. The goal was to assess how well each model could categorize financial disputes without fine-tuning, relying solely on their pre-trained knowledge.

To ensure a structured and efficient evaluation process, we developed a reusable Python class that systematically handled model interactions while separating the core logic from the dataset. This class performed the following key operations:

- Iterating through the dataset – Each complaint was processed row by row.
- Constructing the prompt – The complaint text was dynamically combined with the predefined complaint categories to generate a structured prompt for classification.
- Managing API communications – The appropriate API was called for each respective model, ensuring seamless interaction.
- Handling responses – The models were instructed to return their predictions in JSON format. However, DeepSeek-V3 frequently included extra text outside the expected JSON structure, requiring additional processing to extract and clean the valid classification output. A separate method was implemented to search for and capture the valid JSON response, ensuring uniformity across all model outputs.
- Storing results – All predictions were saved within the same dataset file, making it easier to analyze classification performance across different models.

Additionally, to measure the computational efficiency of each model, we recorded the time taken for each prediction and stored this information in the same CSV file. This allowed us to compare not only accuracy but also latency, which is a crucial factor in real-world applications.

In line with our commitment to open science and unbiased research, we have made all code available as an open-source project on GitHub, licensed under MIT, allowing researchers and developers to replicate and extend our work [28].

3.5. Model Evaluation Strategy

To assess the effectiveness of DeepSeek-V3, GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku in zero-shot consumer complaint classification, a comprehensive evaluation strategy was implemented. The primary goal was to measure how accurately each model classified complaints into the five predefined categories and to analyze their performance trends.

3.5.1. Accuracy-Based Comparison

A straightforward metric for evaluation was exact match accuracy, which counts how many predictions exactly matched the true complaint category. This method provides a quick and direct comparison of classification performance across all models.

Beyond simple accuracy, four key classification metrics were computed for each model:

- Accuracy – The proportion of correctly classified complaints.
- Precision (weighted) – How often a model's predicted category was correct, considering class imbalances.
- Recall (weighted) – How well the model identified complaints belonging to each category.
- F1-score (weighted) – The harmonic mean of precision and recall, balancing both metrics.

Each model's evaluation results were stored in a dedicated CSV file, allowing easy comparison and further analysis.

3.5.2. Confusion Matrix and Heatmap Analysis

To visualize classification patterns, a heatmap was generated based on a confusion matrix. This helped identify:

- Common misclassifications – Whether certain categories were frequently confused with others.
- Model biases – If a model tended to favor certain categories over others.

These heatmaps provided deeper insights into the strengths and weaknesses of models, highlighting their classification tendencies in financial consumer complaints.

By leveraging these evaluation techniques, we established a quantitative foundation for comparing the zero-shot classification capabilities of these leading LLMs, offering valuable insights into their suitability for real-world financial NLP applications.

4. Results

In Section 3, we detailed the methodology used to prompt the models for this classification task. Here, we evaluate the performance of DeepSeek-V3, GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, and Claude 3.5 Haiku in zero-shot consumer complaint classification by examining their accuracy, precision, recall, F1-score, cost, and inference speed.

This section provides a detailed discussion of the findings, offering insights into each model's capabilities, strengths, and weaknesses. The complete outcomes of our study are summarized in **Error! Reference source not found.**.

Table 1. Comparison of model performance metrics.

Model	Accuracy	Precision	Recall	F1	Cost ($)
gpt-4o	0.736	0.7546	0.736	0.7358	0.45
gpt-4o-mini	0.691	0.7026	0.691	0.6933	0.03
claude-3-5-sonnet-20241022	0.763	0.7973	0.763	0.7617	0.64
claude-3-5-haiku-20241022	0.721	0.7287	0.721	0.7227	0.17
deepseek-chat	0.737	0.752	0.737	0.7368	0.01

4.1. Classification Performance

4.1.1. Overall Accuracy

Among the models tested, Claude 3.5 Sonnet achieved the highest accuracy (0.763), demonstrating its superior ability to correctly classify financial complaints. This suggests that it can better distinguish between overlapping complaint categories without requiring fine-tuning.

DeepSeek Chat (0.737) and GPT-4o (0.736) followed closely, showing they are competitive in handling financial texts. However, Claude 3.5 Haiku (0.721) and GPT-4o-mini (0.691) performed slightly worse, indicating a tradeoff between model size and classification effectiveness.

4.1.2. Precision, Recall, and F1-Score

Claude 3.5 Sonnet had the highest precision (0.7973), making its predictions the most likely to be correct. GPT-4o (0.7546) and DeepSeek-V3 (0.752) had similar precision, positioning them as reliable choices, while GPT-4o-mini had the lowest precision (0.7026), indicating a higher likelihood of incorrect classifications.

In terms of recall, which measures how well a model captures all instances of a category, Claude 3.5 Sonnet again led with 0.763, reinforcing its robustness, followed by DeepSeek Chat (0.737) and GPT-4o (0.736), both performing similarly.

The F1-score, which balances precision and recall, further confirmed Claude 3.5 Sonnet (0.7617) as the best overall performer, with DeepSeek-V3 (0.7368) and GPT-4o (0.7358) also achieving competitive scores. GPT-4o-mini had the lowest F1-score (0.6933), reflecting its reduced effectiveness in this classification task.

4.2. Cost Efficiency

All predictions were made using official API versions to ensure a fair comparison. While DeepSeek is open-source and can be run locally, API-based inference was used to standardize the results.

The cost per 1,000 classifications varied significantly across models. DeepSeek-V3 was the most cost-efficient at $0.01, though this price is part of a limited-time offer valid until February 8, 2025. After this date, costs will increase to $0.14 per million input tokens and $0.28 per million output tokens, though it will still remain more affordable than GPT-4o and Claude models. GPT-4o-mini was also highly affordable at $0.03 per 1,000 classifications, with a pricing structure of $0.150 per million input tokens and $0.600 per million output tokens.

GPT-4o and Claude 3.5 Haiku offered a balance between cost and performance. GPT-4o was priced at $0.45 per 1,000 classifications, with a rate of $2.50 per million input tokens and $10.00 per million output tokens. Claude 3.5 Haiku, at $0.17 per 1,000 classifications, had a pricing structure of $0.80 per million input tokens and $4.00 per million output tokens.

Claude 3.5 Sonnet was the most expensive model at $0.64 per 1,000 classifications, with a pricing structure of $3.00 per million input tokens and $15.00 per million output tokens. Despite its higher cost, Claude 3.5 Sonnet delivered the best classification accuracy, making it a viable choice for scenarios where precision is critical.

For budget-conscious applications, DeepSeek-V3 remains the most economical option, but its significantly slower inference speed must be carefully considered.

4.3. Inference Speed and Latency

Inference time is crucial for real-time classification applications. The mean prediction time and total time for processing 1,000 complaints are summarized in **Error! Reference source not found.**. These results highlight the trade-offs between model speed and classification accuracy, which are essential considerations when selecting an LLM for time-sensitive tasks.

Table 2. Inference Speed Comparison of LLMs.

Model	Mean Prediction Time	Total Time (for 1,000 complaints)
gpt-4o	0.89s	889.25s
gpt-4o-mini	0.86s	860.44s
claude-3-5-sonnet-20241022	1.8s	1797.11s
claude-3-5-haiku-20241022	1.81s	1806.23s
deepseek-chat	26.69s	26686.94s

Observations:

- GPT-4o-mini was the fastest model (0.86s per prediction), followed closely by GPT-4o (0.89s). These models are ideal for low-latency applications requiring fast classification.
- Claude 3.5 Sonnet and Claude 3.5 Haiku were slower (~1.8s per prediction), suggesting higher computational demands.
- DeepSeek Chat was the slowest (26.69s per prediction), making it unsuitable for real-time applications despite its low cost.

Additionally, DeepSeek's API has faced recent performance issues due to high demand and attacks following its launch on January 20, 2025.

4.4. Heatmaps: Model Performance Analysis

Heatmaps are powerful visualization tools that aid in analyzing classification tasks by providing an intuitive and comprehensive way to interpret model performance [29]. They visually represent data distributions, feature correlations, and classification results using color gradients, making complex patterns more accessible. In classification tasks, heatmaps are commonly utilized for confusion matrices, feature importance analysis, and activation mapping in DL models, helping researchers and practitioners identify misclassifications, feature relevance, and decision boundaries. In particular, for LLMs, heatmaps can help researchers determine whether a model achieves sufficiently good zero-shot results or requires further fine-tuning. In this study, we employ heatmaps to gain deeper insights into the models' strengths and weaknesses across the five classification categories, as illustrated in **Error! Reference source not found.**.

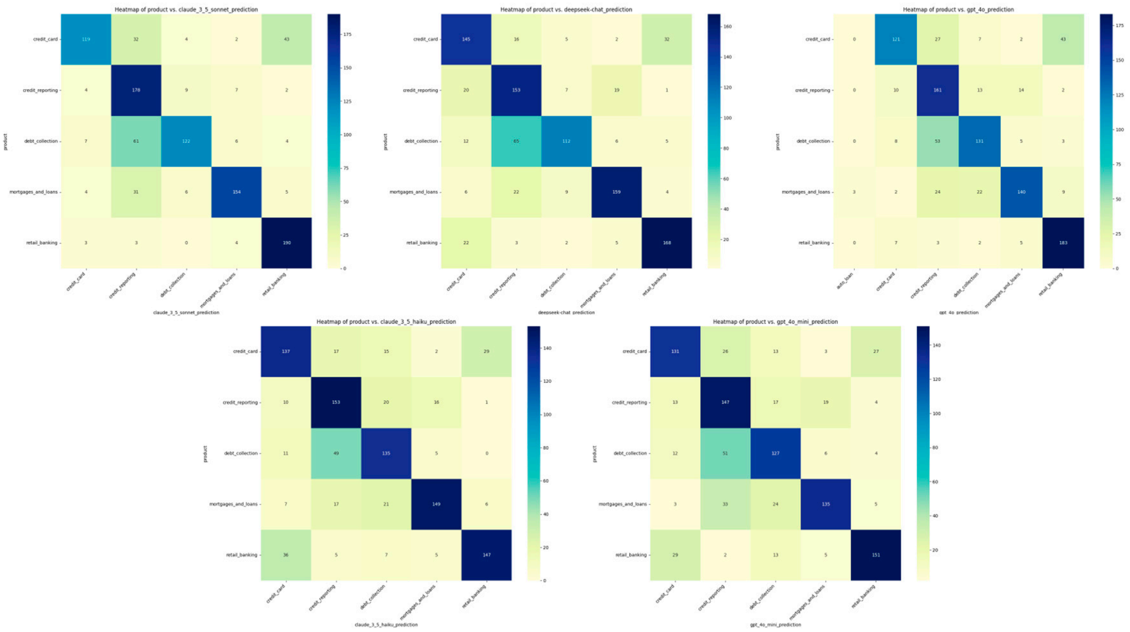

Figure 1. Comparison of Model Predictions with Actual Categories Using Heatmaps.

4.4.1. Claude 3.5 Sonnet

The heatmap of Claude 3.5 Sonnet demonstrates strong classification performance, particularly in credit reporting (178 correctly classified instances) and retail banking (190 correctly classified instances). These results indicate the model's effectiveness in distinguishing these categories.

However, there is notable misclassification between debt collection and credit reporting, with 61 instances of debt collection complaints being misclassified as credit reporting. Additionally, 43 credit card complaints were incorrectly classified as retail banking, highlighting some linguistic overlap between these categories.

doi:10.20944/preprints202502.0720.v1

Further, the model struggled with mortgages and loans, misclassifying 31 instances as credit reporting and 6 instances as debt collection. This suggests that complaints related to mortgages and loans share linguistic similarities with these categories, impacting classification accuracy.

Overall, Claude 3.5 Sonnet demonstrates high classification accuracy but could benefit from fine-tuning to reduce confusion between overlapping financial categories.

4.4.2. DeepSeek-V3

The DeepSeek-Chat model exhibits high accuracy in retail banking (168 correct classifications) and mortgages and loans (159 correct classifications), suggesting robust performance in these domains.

However, a significant number of debt collection complaints (65 instances) were misclassified as credit reporting, and credit reporting complaints (7 instances) were confused with debt collection. This indicates potential linguistic similarities between these categories.

Furthermore, while credit card complaints were mostly classified correctly (145 instances), 32 cases were misclassified as retail banking. Additionally, credit reporting complaints were misclassified into credit card (20 cases) and debt collection (19 cases), showing that certain financial products share common terminology, leading to classification challenges.

Overall, DeepSeek-Chat demonstrates strong classification performance but exhibits noticeable misclassification between credit reporting and debt collection, as well as between credit card and retail banking.

4.4.3. GPT-4o Model

The heatmap for GPT-4o reveals a strong diagonal pattern, indicating high classification accuracy across most categories. The model performs well in classifying credit card (121), credit reporting (161), debt collection (131), mortgages and loans (140), and retail banking (183).

However, notable misclassifications include 43 instances of credit card complaints being classified as retail banking, 13 cases of credit reporting misclassified as debt collection, and 24 cases of mortgages and loans misclassified as credit reporting. These errors suggest overlapping linguistic characteristics across certain financial categories.

Interestingly, GPT-4o ignored explicit classification instructions, predicting the auto loan category three times, even though it was not part of the predefined categories. This behavior indicates that the model may have internal biases or a tendency to generalize beyond the given labels.

To enhance performance, further fine-tuning may be needed, particularly in distinguishing overlapping categories. Implementing context-aware embeddings or refining category definitions could help reduce misclassification errors.

4.4.4. GPT-4o Mini

The GPT-4o Mini heatmap exhibits high classification accuracy, with strong diagonal dominance. The model performs well but shows more confusion between credit reporting and debt collection compared to the full GPT-4o model.

Key misclassification patterns include:

- Credit card complaints misclassified as retail banking (27 instances).
- Credit reporting complaints confused with debt collection (17 instances) and mortgages and loans (19 instances).
- Debt collection complaints misclassified as credit reporting (51 instances).
- Retail banking misclassified into credit card (29 instances).

Compared to GPT-4o, the GPT-4o Mini model classifies credit card complaints more accurately (131 correct vs. 121 in GPT-4o) but exhibits increased misclassification with retail banking.

Overall, GPT-4o Mini shows competitive performance but demonstrates more misclassification across overlapping financial categories than GPT-4o, indicating potential areas for improvement.

4.4.5. Claude 3.5 Haiku

The Claude 3.5 Haiku model's heatmap indicates strong classification performance, with high accuracy in credit card (137 correct), credit reporting (153 correct), and mortgages and loans (149 correct).

Misclassification patterns include:

- Credit card complaints misclassified as retail banking (29 instances).
- Credit reporting complaints confused with debt collection (20 instances) and mortgages and loans (16 instances).
- Debt collection complaints misclassified as credit reporting (49 instances).
- Retail banking complaints misclassified into credit card (36 instances).

Compared to GPT models, Claude 3.5 Haiku exhibits more misclassification from retail banking into credit card but shows slightly less confusion in the debt collection category.

5. Discussion

5.1. Overview of Classification Performance

The results presented in Section 4 highlight the strong performance of various LLMs in classifying financial complaints. The models successfully categorized closely related financial terms into five distinct categories without task-specific fine-tuning. Given that a random classification approach would yield a 20% accuracy rate (1/5 probability), the observed accuracy levels exceeding 70% are a remarkable achievement. This performance can be attributed to the extensive pre-training phase of LLMs, which exposes them to a broad spectrum of financial data. However, while 70% accuracy is notable in an experimental setting, it may not be optimal for real-world applications, necessitating further fine-tuning for improved reliability.

5.2. Misclassification Patterns and Biases

Beyond accuracy, the heatmaps provided insights into misclassification tendencies. Notably, GPT-4o misclassified three mortgage and loan complaints as auto loan issues—a category not included in the prompt. Interestingly, GPT-4o-mini did not exhibit this behavior, suggesting that the larger GPT-4o model's pre-training may have introduced biases or a tendency to generalize beyond the provided labels. This finding underscores the importance of careful fine-tuning, especially for applications where misclassifications could lead to significant consequences, such as automated decision-making in financial or safety-critical domains.

5.3. Comparative Performance of LLMs

One key takeaway is the near-identical accuracy rates of GPT-4o (0.736) and DeepSeek-Chat (0.737). This similarity is particularly intriguing given allegations that DeepSeek may have harvested data from OpenAI's technologies [30]. Despite their comparable accuracy, the heatmaps reveal differences in classification behavior. Unlike GPT-4o, DeepSeek-Chat did not misclassify complaints into the auto loan category, and its classification patterns were distinctly different. This discrepancy suggests that model behavior cannot be fully understood through accuracy metrics alone, highlighting the importance of qualitative analysis through error visualization.

5.4. Model Selection Based on Performance Metrics

Error! Reference source not found. summarizes the key takeaways from this study, providing a comparative analysis of the evaluated models.

Table 3. Summary of Key Takeaways.

Category	Best Model
Best for Accuracy	Claude 3.5 Sonnet (0.763 accuracy, highest precision & recall)
Best for Cost Efficiency	DeepSeek-V3 ($0.01 per 1,000 classifications) but slow
Best for Speed	GPT-4o-mini (0.86s per prediction)
Most Balanced Model	GPT-4o (good accuracy, reasonable cost, fast inference)
Worst in Accuracy	GPT-4o-mini (0.691 accuracy, lowest F1-score)

Key insights from this comparison include:

- Claude 3.5 Sonnet is the best choice when accuracy is the highest priority, particularly for financial complaint classification, where errors can impact case handling.
- GPT-4o provides a strong balance between accuracy, cost, and inference speed, making it ideal for real-time applications.
- DeepSeek Chat is the most cost-effective option but is hindered by slow inference speeds.
- GPT-4o-mini is well-suited for scenarios requiring fast classifications at minimal cost, although its lower accuracy may be a drawback.

5.5. Addressing Common Misclassification Challenges

The heatmaps further highlight recurring misclassification patterns across all models. Certain financial categories, such as credit reporting vs. debt collection and credit card vs. retail banking, exhibit significant overlap, leading to classification errors. Models with high accuracy, such as GPT-4o and DeepSeek-Chat, still struggle with these ambiguities. Notably, Claude 3.5 Sonnet and Haiku perform well in differentiating between debt collection complaints, suggesting that specific model architectures may have inherent strengths in handling particular types of financial data.

5.6. Recommendations for Improving Classification Accuracy

To enhance classification performance, the following strategies could be employed:

- Fine-tuning on domain-specific financial data to improve model understanding of nuanced complaint categories.
- Incorporating context-aware embeddings to mitigate misclassification in overlapping financial categories.
- Enhancing category definitions to provide clearer distinctions between similar complaint types.

By implementing these improvements, financial complaint classification models can achieve greater accuracy and reliability, ultimately enhancing their applicability in real-world financial analysis and customer service automation.

6. Conclusions

This study evaluated the zero-shot classification performance of leading LLMs in financial consumer complaint classification. While all models demonstrated notable capabilities, Claude 3.5 Sonnet consistently outperformed others in accuracy (0.763) and precision (0.7973), making it the most reliable choice for high-stakes financial applications. GPT-4o provided a well-balanced trade-off between accuracy, cost, and inference speed, making it a viable option for real-time classification tasks. DeepSeek-V3 emerged as the most cost-effective model, but its slow inference time limits its practical use in time-sensitive scenarios.

Despite their strengths, all models exhibited misclassification tendencies, particularly in cases where financial categories overlap. Heatmap analysis revealed distinct error patterns, emphasizing the need for further fine-tuning and contextual embeddings to enhance classification accuracy. The misclassification trends observed in GPT-4o suggest that larger models may introduce biases that require mitigation strategies in real-world deployments.

These findings contribute valuable insights into the practical application of LLMs in financial text classification. Future research should explore domain-specific fine-tuning and hybrid approaches that integrate structured financial knowledge to further improve classification performance. By addressing these limitations, LLMs can become more robust tools for automating financial complaint analysis, streamlining regulatory compliance, and enhancing consumer protection efforts.

Author Contributions: Conceptualization, K.I.R. and N.D.T.; methodology, K.I.R., N.D.T. and D.K.N.; software, K.I.R.; validation, K.I.R. and N.D.T.; formal analysis, K.I.R., N.D.T. and D.K.N.; investigation, K.I.R., N.D.T. and D.K.N.; resources, K.I.R.; data curation, K.I.R.; writing—original draft preparation, K.I.R. and N.D.T.; writing—review and editing, K.I.R., N.D.T. and D.K.N.; visualization, K.I.R.; supervision, D.K.N. and N.D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data supporting the reported results can be found at ref [28].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pio, P.G.C.; Sigahi, T.; Rampasso, I.S.; Satolo, E.G.; Serafim, M.P.; Quelhas, O.L.G.; Leal Filho, W.; Anholon, R. Complaint Management: Comparison between Traditional and Digital Banks and the Benefits of Using Management Systems for Improvement. *International Journal of Productivity and Performance Management* **2024**, *73*, 1050–1070. https://doi.org/10.1108/IJPPM-08-2022-0430/FULL/XML.
2. Consumer Financial Protection Bureau (CFPB) Consumer Complaint Database | Consumer Financial Protection Bureau Available online: https://www.consumerfinance.gov/data-research/consumer-complaints/ (accessed on 7 February 2025).
3. Vairetti, C.; Aránguiz, I.; Maldonado, S.; Karmy, J.P.; Leal, A. Analytics-Driven Complaint Prioritisation via Deep Learning and Multicriteria Decision-Making. *Eur J Oper Res* **2024**, *312*, 1108–1118. https://doi.org/10.1016/J.EJOR.2023.08.027.
4. Sharma, S.; Vashisht, M.; Kumar, V. Enhanced Customer Insights: Multimodal NLP Feedback System. *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2024* **2024**. https://doi.org/10.1109/SCEECS61402.2024.10481937.
5. Roy, T.S.; Vasukidevi, G.; Malleswari, T.Y.J.N.; Ushasukhanya, S.; Namratha, N. Automatic Classification of Railway Complaints Using Machine Learning. *E3S Web of Conferences* **2024**, *477*, 00085. https://doi.org/10.1051/E3SCONF/202447700085.
6. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study. *Big Data and Cognitive Computing 2024, Vol. 8, Page 63* **2024**, *8*, 63. https://doi.org/10.3390/BDCC8060063.
7. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models. *Future Internet 2025, Vol. 17, Page 28* **2025**, *17*, 28. https://doi.org/10.3390/FI17010028.
8. DeepSeek-AI; :; Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; et al. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. **2024**.
9. Models - OpenAI API Available online: https://platform.openai.com/docs/models#gpt-4o (accessed on 10 February 2025).
10. Models - Anthropic Available online: https://docs.anthropic.com/en/docs/about-claude/models (accessed on 10 February 2025).
11. Sakas, D.P.; Reklitis, D.P.; Terzi, M.C.; Glaveli, N. Growth of Digital Brand Name through Customer Satisfaction with Big Data Analytics in the Hospitality Sector after the COVID-19 Crisis. *International Journal of Information Management Data Insights* **2023**, *3*, 100190. https://doi.org/10.1016/J.JJIMEI.2023.100190.
12. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Leveraging Large Language Models in Tourism: A Comparative Study of the Latest GPT Omni Models and BERT NLP for Customer Review Classification

and Sentiment Analysis. *Information 2024, Vol. 15, Page 792* **2024**, *15*, 792. https://doi.org/10.3390/INFO15120792.

13. Kumar, P.; Singh, A.; Saha, S. Navigating the Indian Code-Mixed Terrain: Multitasking Analysis of Complaints, Sentiment, Emotion, and Severity. **2024**. https://doi.org/10.2139/SSRN.4827145.
14. Das, S.; Singh, A.; Saha, S.; Maurya, A. Negative Review or Complaint? Exploring Interpretability in Financial Complaints. *IEEE Trans Comput Soc Syst* **2024**, *11*, 3606–3615. https://doi.org/10.1109/TCSS.2023.3338357.
15. Jondhale, R.; Patil, S.; Shinde, A.; Ajalkar, D.; Biradar, S. Predicting Consumer Complaint Disputes in Finance Using Machine Learning (AIOPS). *2nd IEEE International Conference on Advances in Information Technology, ICAIT 2024 - Proceedings* **2024**. https://doi.org/10.1109/ICAIT61638.2024.10690778.
16. IMPLEMENTING CHATBOTS FOR CONSUMER COMPLAINT RESPONSE | Proceedings of International Conference on Modern Science and Scientific Studies Available online: https://econferenceseries.com/index.php/icmsss/article/view/3992 (accessed on 9 February 2025).
17. Zhou, X.; Cao, G.; Peng, B.; Xu, X.; Yu, F.; Xu, Z.; Yan, Y.; Du, H. Citizen Environmental Complaint Reporting and Air Quality Improvement: A Panel Regression Analysis in China. *J Clean Prod* **2024**, *434*, 140319. https://doi.org/10.1016/J.JCLEPRO.2023.140319.
18. Khadija, M.A.; Nurharjadmo, W. Enhancing Indonesian Customer Complaint Analysis: LDA Topic Modelling with BERT Embeddings. *SINERGI* **2024**, *28*, 153–162. https://doi.org/10.22441/SINERGI.2024.1.015.
19. Song, W.; Rong, W.; Tang, Y. Quantifying Risk of Service Failure in Customer Complaints: A Textual Analysis-Based Approach. *Advanced Engineering Informatics* **2024**, *60*, 102377. https://doi.org/10.1016/J.AEI.2024.102377.
20. Seok, J.; Kim, C.; Kim, S.; Kim, Y.-M. Deep-Learning-Based Customer Complaints Monitoring System Using Online Review. **2024**. https://doi.org/10.2139/SSRN.4795530.
21. Jia, S.; Shan, G.; Chi, O.H. Leveraging Generative AI for Customer Complaint Resolution: A Comparative Analysis with Human Responses. *AMCIS 2024 Proceedings* **2024**.
22. Lee, C.H.; Zhao, X. Data Collection, Data Mining and Transfer of Learning Based on Customer Temperament-Centered Complaint Handling System and One-of-a-Kind Complaint Handling Dataset. *Advanced Engineering Informatics* **2024**, *60*, 102520. https://doi.org/10.1016/J.AEI.2024.102520.
23. Wang, R.; Wang, H.; Li, S. Predicting the Determinants of Consumer Complaint Behavior in E-Commerce Live-Streaming: A Two-Staged SEM-ANN Approach. *IEEE Trans Eng Manag* **2025**, 1–12. https://doi.org/10.1109/TEM.2025.3533921.
24. Juipa, A.; Guzman, L.; Diaz, E. Sentiment Analysis-Based Chatbot System to Enhance Customer Satisfaction in Technical Support Complaints Service for Telecommunications Companies .; International Conference on Smart Business Technologies (ICSBT), 2024.
25. Tiwari, S. Consumer Complaints Dataset for NLP Available online: https://www.kaggle.com/datasets/shashwatwork/consume-complaints-dataset-fo-nlp (accessed on 7 February 2025).
26. Zhang, K.; Zhou, F.; Wu, L.; Xie, N.; He, Z. Semantic Understanding and Prompt Engineering for Large-Scale Traffic Data Imputation. *Information Fusion* **2024**, *102*, 102038. https://doi.org/10.1016/J.INFFUS.2023.102038.
27. Anthropic PBC Automatically Generate First Draft Prompt Templates - Anthropic Available online: https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator (accessed on 30 November 2024).
28. Roumeliotis, K. GitHub - Applied-AI-Research-Lab/DeepSeek-LLM-and-GPT-Fall-Behind-Claude-Leads-in-Zero-Shot-Consumer-Complaints-Classification Available online: https://github.com/Applied-AI-Research-Lab/DeepSeek-LLM-and-GPT-Fall-Behind-Claude-Leads-in-Zero-Shot-Consumer-Complaints-Classification (accessed on 7 February 2025).

29. Zhao, S.; Guo, Y.; Sheng, Q.; Shyr, Y. Advanced Heat Map and Clustering Analysis Using Heatmap3. *Biomed Res Int* **2014**, *2014*, 986048. https://doi.org/10.1155/2014/986048.
30. Cade Metz OpenAI Says DeepSeek May Have Improperly Harvested Its Data - The New York Times Available online: https://www.nytimes.com/2025/01/29/technology/openai-deepseek-data-harvest.html (accessed on 9 February 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.