Accept or Deny? Evaluating LLM Performance and Fairness in Loan Approval

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly employed in high-stakes decision-making tasks, such as loan approvals. Despite their expanding applications across various domains, LLMs continue to struggle with processing tabular data, ensuring fairness, and delivering reliable predictions. In this work, we assess the effectiveness of LLMs in loan approval, with a particular focus on their zero-shot and in-context learning (ICL) capabilities. Specifically, we evaluate the performance of several LLMs on loan approvals using datasets from three geographical locations, namely Ghana, Germany and the United States. We analyze the impact of different serialization formats, such as `JSON`, and natural language-like text, on model performance and fairness. Our results indicate that LLMs perform significantly worse than classical machine learning models in zero-shot classification tasks, often displaying a tendency to either approve or reject all loan applications. While ICL improves performances of models by 17-27% (relative), its impact on fairness remains inconsistent. Our work underscores the importance of effective tabular data representation methods and fairness-aware models to improve the reliability of LLMs in financial decision-making.

1 Introduction

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have led to their widespread adoption across various industries, enabling automated decision-making in fields such as healthcare, education, and finance (Jindal et al., 2024; Nguyen et al., 2023; Munir et al., 2022). Among these advancements, Large Language Models (LLMs), trained on vast amounts of textual data, have demonstrated remarkable potential to generalize across tasks and provide accurate predictions (Naveed et al., 2023; AI4Science and Quantum, 2023). Given their growing presence in critical domains like financial decision-making and loan approval, it is crucial to understand the behavior and ethical implications of these systems due to their direct impact on individuals.

However, despite the benefits LLMs bring to various areas, challenges still persist. i) While traditional ML models are designed for tabular data, LLMs are not natively equipped for such tasks. Converting tabular data into textual formats for LLMs can introduce challenges, as the transformation may lose important structures and relationships inherent in the original data (Singha et al., 2023; Sui et al., 2024). ii) LLMs trained on large datasets often inherit biases present in the data. The model may unintentionally amplify these biases, particularly in critical areas like financial decision-making. This raises concerns about fairness, highlighting the urgent need for strategies to mitigate such biases. iii) There is limited research on how in-context learning (ICL), which embeds task-relevant examples in the input prompt, can enhance the accuracy and fairness of LLMs in financial decision-making. The challenge lies in whether embedding these examples effectively improves the model's ability to make unbiased decisions, especially in sensitive areas like finance. This gap in understanding raises questions about how ICL can be leveraged to ensure both accurate and fair outcomes. iv) LLMs are not currently evaluated on datasets from diverse regions, limiting the ability to assess how well they adapt to varying financial behaviors shaped by cultural and economic differences (Myung et al., 2024). This lack of evaluation raises concerns about their effectiveness and fairness when applied to global financial contexts.

In light of these challenges, our work aims to make the following contributions [1]:

1. We investigate the capability of LLMs in financial decision-making, focusing on loan approval tasks. This includes a comprehensive

[1] We will make our code public upon acceptance.

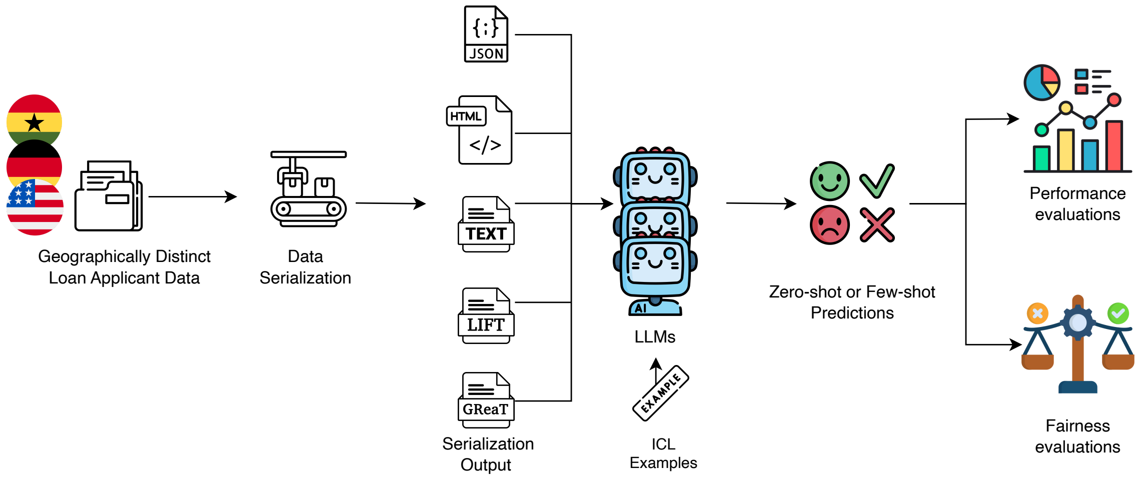

Figure 1: Our methodology involved utilizing datasets sourced from three distinct countries, and applying various serialization methods. For each serialization approach, we conducted experiments with zero-shot or few-shot learning techniques, assessing both model performance and fairness outcomes.

Data Name	Size	#Features	Output
Ghana	614	13	Yes/No
German	1000	21	Good/Bad
United States	1451	18	Yes/No

Table 1: Summary of the datasets used in the study. Ghana (Sackey and Amponsah, 2018), German (Statlog) and United States (Kaggle). Details of the feature description of each dataset are provided in the Appendix C.

zero-shot benchmark evaluation of various LLMs and an analysis of the features they prioritize in their decision-making process.

2. We analyze the impact of different tabular serialization formats on the decision-making process of LLMs.
3. We evaluate the effectiveness of techniques, such as in-context learning, that aim to improve LLM performance in financial decision-making, with particular attention to their impact on accuracy and fairness.
4. We examine the presence of gender-related biases in LLM-generated financial decisions, assessing their implications and associated risks.

2 Related Work

LLMs in Decision-Making Systems. Large Language Models (LLMs) have demonstrated significant promise in decision-making across a wide range of domains, including healthcare (Kim et al., 2024), education (Hendrycks et al., 2021), research (Kwiatkowski et al., 2019), and supply chain management (Li et al., 2023). In healthcare, for instance, LLMs have been instrumental in diagnosing diseases and recommending treatment plans by processing vast amounts of medical literature and patient data (Thirunavukarasu et al., 2023; Yang et al., 2023). Similarly, in agriculture, LLMs have been utilized to optimize crop yields and resource management by analyzing data from various sources (Rezayi et al., 2022; Silva et al., 2023). More specifically, within the financial sector, LLMs are used to analyze market trends, expedite loan approvals, and offer investment advice (Li et al., 2023; Huang et al., 2023; Wu et al., 2023). Despite the clear advantages in efficiency and decision-making speed, there is a noticeable gap in research investigating the impact of these decisions on individuals. A false positive, such as approving a loan for an undeserving applicant, may lead to financial difficulties for the borrower. Over time, the applicant could be unable to repay the loan, leading to a decrease in their credit score and exacerbating their financial burdens. This situation underscores the importance of considering the potential long-term consequences of LLM-driven decisions in finance.

Serialization in LLMs. To enable the use of LLMs for tabular data, the table must be serialized into a natural text representation, a process referred to as *serialization* (Jaitly et al., 2023). Serialization methods, which convert tabular data into a format that LLMs can process, can introduce their own biases and limitations. For instance, (Hegselmann et al., 2023) discusses how different seri-

Serialization	Example Template
JSON (default)	{age: 32, sex: female, loan duration: 48 months, purpose: education}
GReaT (Borisov et al., 2022)	age is 32, sex is female, loan duration is 48 months, loan purpose is education
LIFT (Dinh et al., 2022)	A 32-year-old female is applying for a loan for 48 months for education purposes.

Table 2: **Comparison of serialization formats for loan applicant information.** This table presents example templates for representing loan applicant data with four features (age and sex, loan duration and purpose). JSON is assumed as the default format. The selected serialization formats ensure diverse data representation, balancing availability across different formats, naturalness, and alignment with prior work. Table 7 in Appendix D shows examples for the List, Text, HTML and Latex format.

alization formats can lead to variations in LLMs performance. Their study highlights that the choice of serialization method can influence how effectively an LLMs understands and processes the data. A number of studies have proposed different serialization methods, including Text and List formats (Hegselmann et al., 2023), the GReaT format (Borisov et al., 2022), natural-like serialization as used in LIFT (Dinh et al., 2022), and HTML-like formatting (Sui et al., 2024). Additionally, works like TabPFN (Hollmann et al., 2022) introduce tabular foundation models specifically designed for tabular datasets. However, in this work, we focus on the capabilities of general-purpose LLMs and their financial domain variants. We do not cover tabular foundation models due to the broad range of serialization formats considered in our study, which may not align well with such models.

Bias and Unfairness of LLMs. The field of machine learning has long contended with biases and ethical issues, which have become even more pronounced with the rise of LLMs. These models are trained on large corpora of human-generated text, which often contain inherent societal biases (Garg et al., 2018; Navigli et al., 2023; Sun et al., 2019; Kotek et al., 2023). As a result, these biases can be encoded into the models and perpetuated in their decisions, leading to discriminatory outcomes. For instance, gender, racial, and cultural biases present in the training data can result in unfair treatment of certain groups (Bolukbasi et al., 2016; Abid et al., 2021). Addressing these biases is crucial to ensure fair and ethical use of LLMs in decision-making processes.

Our study examines the use of large language models (LLMs) for loan approval decisions across datasets from three geographical regions. We explore two key dimensions: the impact of serialization methods and the effect of zero-shot and few-shot prompting on decision accuracy and fairness.

3 Methodology

3.1 Problem and Dataset Description

3.1.1 Problem Formalization

Given the tabular dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i is a d-dimensional feature vector and y_i belongs to a set of classes C, the columns or features are named $F = \{f_1, \ldots, f_d\}$. Each feature f_i is a natural-language string representing the name of the feature, such as "age" or "sex". For zero-shot learning, we provide the LLMs with features F and task it to predict the class C. For our k-shot classification experiments, we use a subset D_k of size k—sampled from the training set.

3.1.2 Datasets

We provide a summary of the dataset we used in the study in Table 1 with a detailed description in Appendix C. For each dataset, we split the dataset into 80% train and 20% test using stratified sampling.

To convert each dataset to the formats shown in Table 2 we created custom functions and also used pandas [2] functions that change dataframe to HTML and Latex. See Table 7 in Appendix D for examples of Latex, Text, HTML and List formats.

3.2 Models

3.2.1 Baseline and Benchmark Models

To comprehensively understand and accurately evaluate the investigated LLMs, we incorporated simple baseline models and a benchmark model.

[2] https://pandas.pydata.org/

- The **zero** model and **one** model serve as our simple baselines, as shown in Figure 3. The **zero** model assumes that no one will repay the loan (i.e. zero output for all predictions), while the **one** model assumes that everyone will repay the loan (one output for all predictions). These models provide initial reference points for our experiment, illustrating the performance metrics under these extreme assumptions.
- Additionally, we trained a Logistic Regression model on the training set to serve as our benchmark model. This model allows us to compare the performance of the LLMs against traditional and well-understood machine learning models. In training the Logistic Regression model, we preprocessed the dataset by dropping missing values, applying label encoder to the categorical features, and scaling all numerical features using a standard scaler.

We acknowledge that other classical models, such as decision trees or support vector machines, might be optimized for this task and potentially yield better performance. However, our primary objective was to establish a straightforward benchmark for comparison.

3.2.2 Large Language Models

In this work, we investigated *ten* (10) LLMs. Our selection criteria for these LLMs focused on their i) open-source nature, ii) popularity, iii) size, and iv) specific domain coverage. The open-source nature and popularity of the models are important because they indicate the potential for broad adoption across various domains. We deliberately excluded our closed-source model due to resource constraints.

We considered models that have been trained purposely for financial applications and these included FinMA-7B-NLP and FinMA-7B-full from the work of (Xie et al., 2023). To incorporate open-source models optimized for instruction tuning, we considered Meta's LLaMA-3-70B-Instruct and LLaMA-3-8B-Instruct, as well as Google's Gemma-2-27b-it and Gemma-2-9b-it. Additionally, we included both the smaller and base variants of Gemma-2-9b, LLaMA-3-8B, Gemma-2-27b, and LLaMA-3-70B from the work of (Team et al., 2024; Touvron et al., 2023; Meta, 2024). .

Detail model evaluation setup is presented in the Appendix B along with detailed model token attribution extended result in Appendix H.

3.3 Approaches to LLMs Improvement

3.3.1 In-Context Learning (ICL)

In-context learning involves providing examples to enhance the capabilities of LLMs (Zhang et al., 2024; Agarwal et al., 2024). This approach is widely used because it eliminates the need for parameter updates, reducing computational costs associated with training. Following a similar approach utilized by the work of (Zhang et al., 2024) we experimented with different numbers of examples, specifically $n = 2, 4, 6, 8$.

3.3.2 Table-to-Text Serialization

Given that LLMs are trained on textual datasets, and our datasets are in tabular format, we need to convert these tables to text. This process, often referred to as *serialization*, is crucial because the format in which data is presented can significantly impact the decision-making ability of LLMs (Hegselmann et al., 2023). To investigate how this behaviour transfers to our loan approval task, we explored *six* serialization formats as shown in Table 2 and Table 7 in Appendix D. These formats ranged from straightforward default values, such as JSON and List, to more structured and natural language text-like formats, such as HTML, Latex, Text (Hegselmann et al., 2023), GReaT (Borisov et al., 2022) and LIFT (Dinh et al., 2022).

3.4 Model and Fairness Evaluation

To evaluate the performance of the models on the loan prediction task, we considered standard metrics such as the weighted-average F1 score and the model accuracy (see Appendix A for definitions). However, due to the imbalanced nature of the datasets, we present only the results based on the weighted average F1 score. Furthermore, to assess whether the LLMs encode potential bias, we employed two popular fairness metrics. We select equality of opportunity as it better aligns with the objectives of loan approval tasks, ensuring that qualified applicants, regardless of group membership, have an equal chance of approval (Kozodoi et al., 2022). We also consider statistical parity, which assesses whether approval rates are independent of sensitive attributes. The formal definitions of these metrics are provided below:

Definition 1 (Statistical Parity (SP)) *(Dwork et al., 2012) A trained classifier's predictions $\hat{Y}$ satisfies this definition if the probability of a*

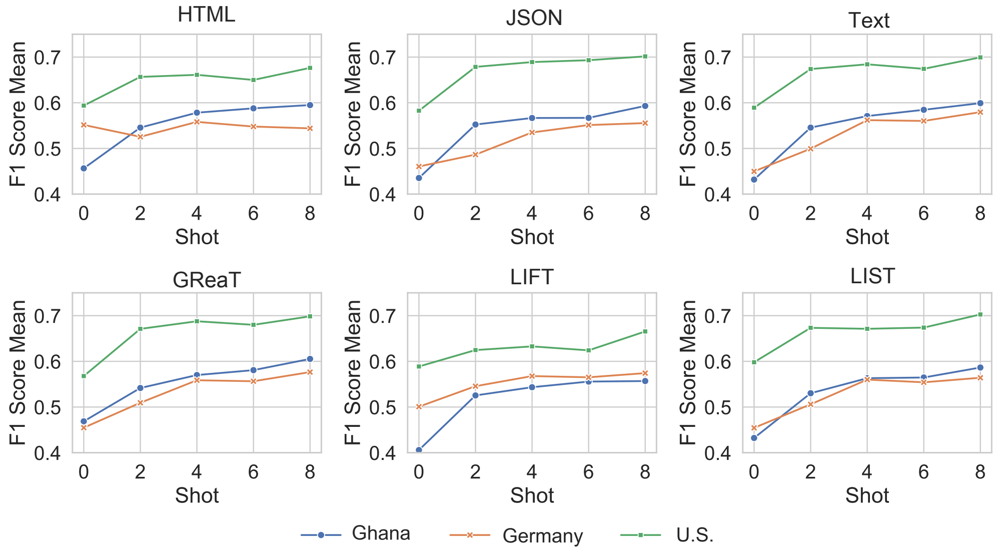

Figure 2: **Comparison of weighted F1 score trends across serialization formats for few-shot examples averaged over all ten (10) models.** This table illustrates how in-context learning (ICL) enhances loan prediction tasks, with varying performance trends across different serialization formats.

positive outcome is independent of the sensitive attribute.

$$P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1] \qquad (1)$$

A represents the sensitive attribute to be protected. In this work, we consider the *gender* attribute as the sensitive attribute and for simplicity, we assumed it to be binary (i.e. male or female). The notation $\hat{Y}$ represents the predictions of the classifier, while Y refers to the true target label.

Definition 2 (Equality of Opportunity (EO)) *(Hardt et al., 2016) Equality of opportunity ensures that the true positive rate is the same across different demographic groups. A classifier $\hat{Y}$ satisfies equality of opportunity if:*

$$P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1) \qquad (2)$$

For all our experiments we considered females as the protected group and males as the non-protected group.

4 Results

Figure 3 compares the performance of different serialization methods across models for each dataset. The **zero** model outperforms the **one** model in the Ghana and United States (U.S.) datasets, while on the German dataset, the reverse is true. This shows that the German dataset has a higher rate of non-defaulters as compared to the other two datasets.

In the following subsection, we investigate these findings by addressing key research questions:

4.1 Does LLMs perform better than baseline/benchmark models on the default format (JSON)?

In Figure 3, we compare the zero-shot performance of LLMs against baseline models. Analyzing the results by country, the general trend indicates that most models do not outperform either the **zero** model or the **one** model. Some models achieved marginally higher F1 scores, including `Gemma-2-9b-it` for Ghana and seven models for the U.S., while none did so for Germany. Importantly, none of the selected LLMs were able to outperform the simple `Logistic Regression model`, which serves as the benchmark.

> For `JSON` serialization method financial domain-specific models (`FinMA-7B-full`, `FinMA-7B-NLP`) do not demonstrate significantly better performance under zero-shot decision-making compared to models trained for general applications. Also, none of the models outperform the baseline `Logistic Regression model`.

4.2 Does natural language improve performance?

We explored multiple serialization techniques, each requiring varying levels of effort to implement, as shown in Table 2. The effort ranges from moder-

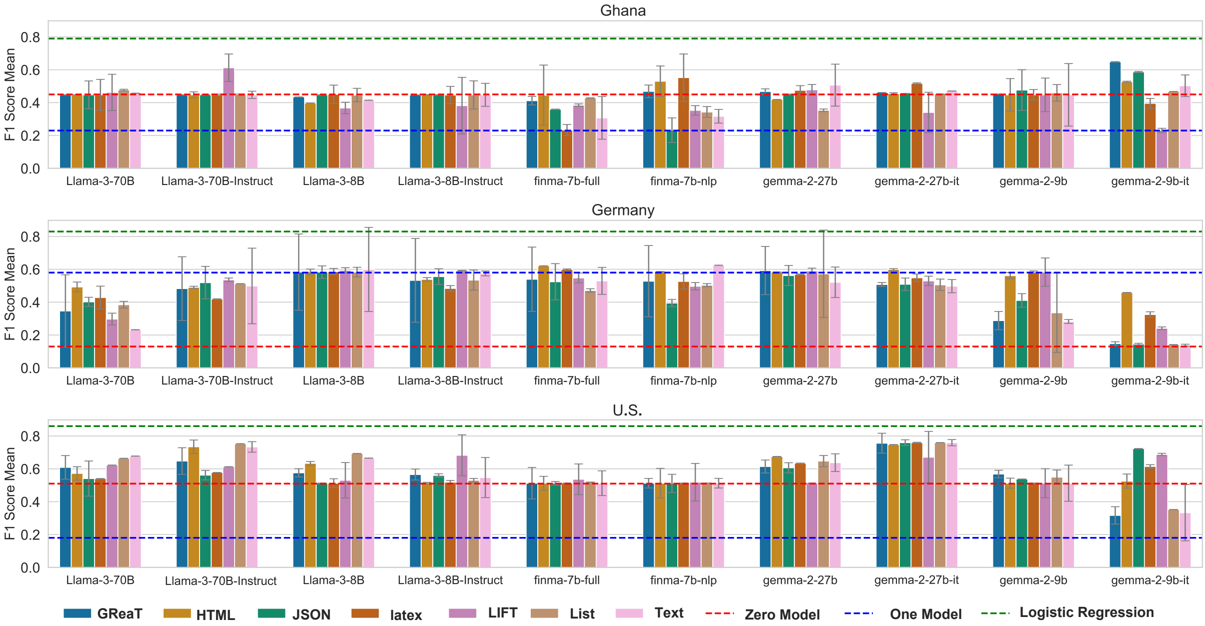

Figure 3: **Serialization types performance.** Zero-shot performance of LLMs in loan approval, measured by the weighted average F1 Score, across different table-to-text serialization methods. The results are an average of three different prompts. Logistic regression benchmark uses default JSON serialization where variable as indivisual features.

ate refinement for serializations like "GReaT" to more detailed modifications for "lift" serialization. We hypothesize that making the text sound more natural would lead to performance improvements.

> Inputs that resemble natural language more closely do not necessarily yield the best performance.

4.3 How does the zero-shot performance of LLMs vary across different serialization methods compared to baseline models?

Examining region-specific results, we observe the following from Figure 3:

For the *Ghana* dataset, the best performances are achieved using the GReaT serialization method (Gemma-2-9b-it) and LIFT serialization method (LLaMA-3-70B-Instruct).

In the *German* dataset, Gemma-2-9b-it shows the poorest performance, with three out of four models performing as poorly as the zero model. Financial domain-trained models (FinMA-7B-full and FinMA-7B-NLP) deliver the best results with List and Text serialization methods.

For the *U.S.* dataset, results are generally more promising across all models, with Gemma-2-27b-it consistently achieving the best performance across all serialization methods tested except LIFT.

With reference to Figure 8, we observe that decisions are more dependent on datasets than models. Particularly, finance-based models tend to show low performance in U.S. and Ghana data while Gemma-2-9b-it shows lower performance in German data. Looking at the average across the formats Gemma-2-27b-it performs best for the U.S., LLaMA-3-8B performs well for Germany. The details of this experiment have been included in Appendix G.

> Serialization methods can significantly affect model performance, emphasizing the need for careful selection. Furthermore, LLM performance on the loan approval task depends on the data source.

4.4 Does using few-shot examples improve the decision-making abilities of LLMs?

Given the LLMs' sub par performance in the zero-shot experiments, we explored various methods to improve their decision-making capabilities through in-context learning(ICL). Figure 2 presents the results from our ICL experiment, where we provide the model with varying numbers of n-shot examples, ranging from zero-shot (n=0) to 8-shot across datasets and serialization formats. We can observe

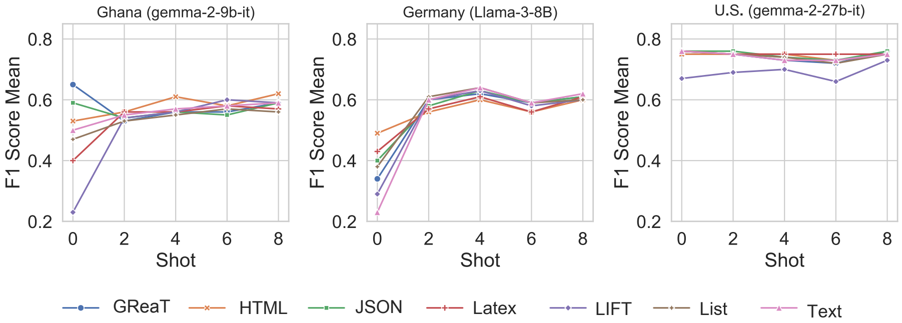

Figure 4: **Mean weighted F1 scores for few-shot results using different serialization methods for best models on each dataset.** This table presents the mean weighted F1 scores for few-shot learning across different datasets, using the best-performing model for each data source. The reported results are averaged over three different prompts.

presenting examples can improve all the tasks. This trend continues in Figure 4 where we see average improvement when with more examples we show across all serialization methods.

> 💡 Performance of the models for all serialization increases as the number of example shots increases. Thus, few-shots examples can improve the decision-making of LLMs for loan approval, to an extent.

	Germany		Ghana		U.S.	
	SP	**E0**	**SP**	**E0**	**SP**	**E0**
Baseline models						
Zero Model	0.00	0.00	0.00	0.00	0.00	0.00
One Model	0.00	0.00	0.00	0.00	0.00	0.00
Logistic Regression	-0.03	-0.08	-0.04	0.05	-0.02	-0.01
Models Fine-tuned for Finance						
FinMA-7B-full	**0.13**	**0.16**	0.03	**0.06**	0.00	0.00
FinMA-7B-NLP	0.07	0.07	0.00	0.01	0.00	0.00
Mid range open-source base models						
LLaMA-3-8B	0.00	0.00	0.00	0.00	0.00	0.00
Gemma-2-9b	0.05	0.05	-0.03	-0.04	**-0.06**	-0.11
Mid range open-source instruction tuned models						
LLaMA-3-8B-Instruct	0.03	0.06	0.00	0.00	0.01	0.02
Gemma-2-9b-it	0.01	0.01	0.03	0.04	-0.04	0.13
Large range open-source instruction tuned models						
LLaMA-3-70B-Instruct	-0.03	0.01	0.00	0.00	-0.01	0.03
Gemma-2-27b-it	-0.01	-0.02	0.00	0.02	0.04	**0.17**
Large range open-source base models						
LLaMA-3-70B	-0.05	-0.05	0.00	0.00	0.00	0.00
Gemma-2-27b	0.00	0.03	0.00	-0.02	0.01	0.07

Table 3: **zero-shot performance JSON**(default), displaying the Weighted-Average Statistical Parity (SP), and Equality of Opportunity (EO) metrics. It compares performance across the baseline models (**zero** model and **one** model), the benchmark model (Logistic Regression model), and the zero-shot models for all three datasets.

4.5 Does decision-making impact vary across different demographic groups?

From Table 3, our baseline models all show no discrimination in terms of equality of opportunity (EO) and statistical parity (SP). However, we see high discrimination in terms of both EO and SP with the FinMA-7B-full for the German dataset. Similarly, we see this model also returns the highest disparity in terms of EO in the Ghana dataset. It is interesting to note that this model among the other models selected in this study is the only one fine-tuned for finance. This therefore opens up interesting research directions on further investigating the fairness of downstream tasks that have been trained with this model. In a similar light, Gemma-2-27b-it returns the highest disparity in terms of EO for the U.S. dataset. On the contrary, LLaMA-3-8B has no disparity in terms of both fairness metrics on the German data. Further highlighting that different models penalize sensitive groups differently.

> 💡 Finance based models shows higher gender based disparity.

4.6 Do few-shot examples improve fairness?

In the German dataset, With reference to Figure F, few-shot examples (e.g., $n = 4$) can lead to significant fairness disparities in equality of opportunity (EO), reaching differences of up to 0.10 for some serialization methods in the German datasets. For statistical parity (SP), disparities generally stabilize or narrow as the number of examples increases, though the extent of improvement varies by dataset and serialization method.

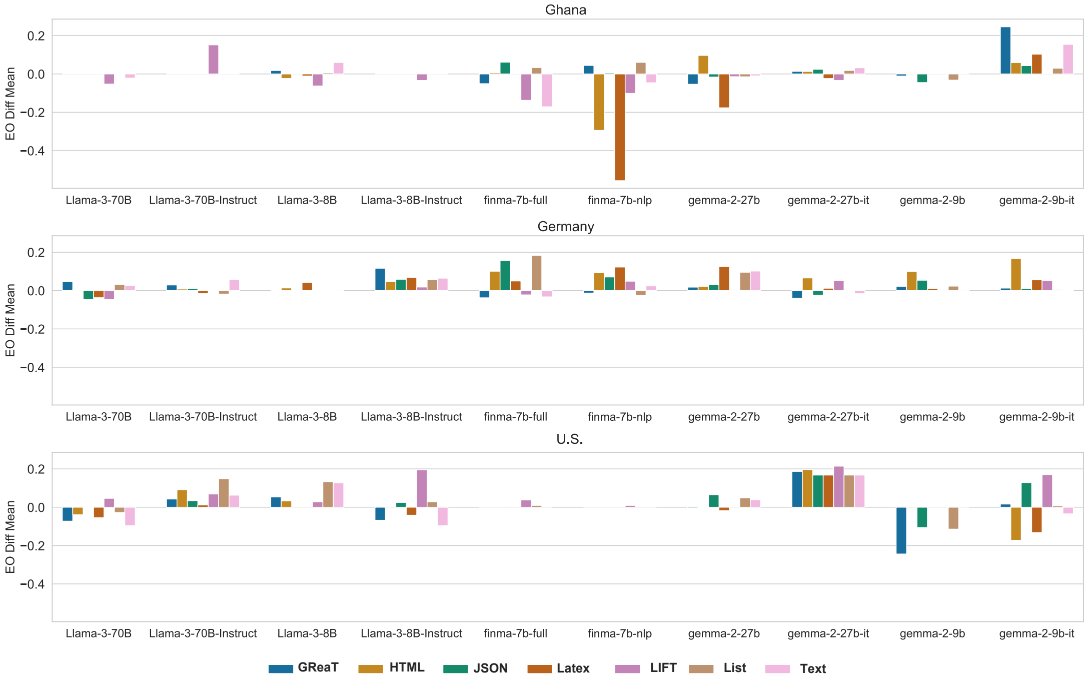

Figure 5: The mean difference in EO for different serialization methods and models. Finance-based models show higher gender-based disparity for certain serializations while the results are highly region and format-dependent.

Fairness in few-shot learning is highly context-dependent. While more examples can sometimes reduce disparities, the impact is not universal, underscoring the importance of carefully selecting and evaluating serialization methods to ensure fairness.

4.7 How does prompt sensitivity vary across different regions and models?

The results in Figure 3 represent the average performance across three different prompts, with error bars indicating the sensitivity to prompt variations. We observe relatively low prompt sensitivity in the U.S. and Ghana datasets, whereas the German dataset exhibits significantly higher sensitivity to prompt differences.

LLM performance sensitivity to prompts varies across data sources—some datasets exhibit stable results across prompts, while others show significant variability.

4.8 What is the fairness F1 score tradeoffs?

Following the best-performing models, as shown in Figure 4, we assess the fairness of these models in Figure 5. The `Gemma-2-27b-it` model shows a degree of disparity for the U.S. data. In the case of the best-performing model for Germany, `LLaMA-3-70B-Instruct` does not show a higher level of unfairness compared to the `LLaMA-3-70B-Instruct` and `FinMA-7B-full` models. Similarly, the `Gemma-2-9b-it` model does not show a higher disparity in EO difference. Nevertheless, the `FinMA-7B-full` model shows a higher disparity in terms of EO in both the Ghana and Germany datasets. The negative EO difference highlights that the model discriminates against the non-protected group, which in this case is males.

Financial-based models exhibit greater disparities in EO mean difference, highlighting higher levels of unfairness.

5 Conclusion

Our study assessed the performance of LLMs in the loan approval task across various model settings, including general open-source models and those trained specifically for the financial domain. We examined the impact of in-context learning and explored the effects of different serialization methods. Future work aims to develop models that adapt to diverse serialization techniques, enhance performance while maintaining fairness, and provide deeper insights into how different few-shot examples influence fairness.

Limitations

Dataset Differences. In our work, we examined data sources from different regions, but a detailed study and analysis of the differences between these datasets are crucial. We used the default column names and values for all datasets. However, some of our serialization methods, such as `LIFT`, aimed to improve column names by correcting spelling errors and related mistakes inherent in the datasets. We acknowledge that there may still be variances that have not been captured and need further investigation. Due to the black-box nature of LLMs and computational constraints, we did not analyze the effect of individual features.

More Datasets. This study focused on three datasets from distinct geographical regions. While incorporating additional datasets with greater variability could improve the research, we maintained this scope to align with the study's objectives and constraints.

LLMs Covered in the Work. This work covers a limited number of LLMs and we mostly focused on models that we believed to the best of our knowledge would be adapted to several use cases because of popularity, open source and continued support by organizations that release them. We purposefully left our closed-sourced model because of a lack of resources and the difficulty in understanding what decisions or generations are made.

Prompt Design. In this study, we generated prompts by referencing similar research works. While certain prompt structures may outperform others, a comprehensive exploration of prompt engineering techniques is beyond this work's scope due to the extensive number of experiments conducted. We acknowledge the importance of this aspect and propose it as a direction for future research.

Explaining Model Behavior. We conducted token token attribution experiments to better understand the reasoning behind model behavior. However, as the results were inconclusive, we have not included a detailed discussion in the main text. Instead, a comprehensive account of the findings can be found in Appendix H.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. 2025. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966.

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. 2022. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2022. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Sukriti Jaitly, Tanay Shah, Ashish Shugani, and Razik Singh Grewal. 2023. Towards better serialization of tabular data for few-shot classification. *arXiv preprint arXiv:2312.12464*.

Jenelle A Jindal, Matthew P Lungren, and Nigam H Shah. 2024. Ensuring useful adoption of generative artificial intelligence in healthcare. *Journal of the American Medical Informatics Association*, 31(6):1441–1444.

Kaggle. Loan approval prediction dataset. https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset. Accessed: 2024-07-19.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. 2022. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. 2023. Large language models for supply chain optimization. *arXiv preprint arXiv:2307.03875*.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. https://ai.meta.com/blog/meta-llama-3/. [Accessed 01-06-2024].

Hussan Munir, Bahtijar Vogel, and Andreas Jacobsson. 2022. Artificial intelligence and machine learning approaches in digital education: A systematic revision. *Information*, 13(4):203.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Duc Khuong Nguyen, Georgios Sermpinis, and Charalampos Stasinakis. 2023. Big data, artificial intelligence and machine learning: A transformative symbiosis in favour of financial technology. *European Financial Management*, 29(2):517–548.

Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, Chen Zhen, Tianming Liu, and Sheng Li. 2022. Agribert: Knowledge-infused agricultural language models for matching food and nutrition. In *IJCAI*, pages 5150–5156.

Frank Gyimah Sackey and Peter Nkrumah Amponsah. 2018. Gender discrimination in commercial banks' credit markets in ghana: a decomposition and counterfactual analysis. *African Journal of Business and Economic Research*, 13(2):121–140.

Bruno Silva, Leonardo Nunes, Roberto Estevão, and Ranveer Chandra. 2023. Gpt-4 as an agronomist assistant? answering agriculture exams using large language models. *arXiv preprint arXiv:2310.06225*.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in llms. *arXiv preprint arXiv:2310.10358*.

Statlog. Statlog (german credit data). `https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29`. Accessed: 2024-07-19.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Preprint*, arXiv:2306.05443.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba O Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. *arXiv preprint arXiv:2402.12976*.

Appendix

A Metrics

In evaluating the performance of Large Language Models (LLMs), we employ several key metrics to assess their predictive accuracy. These metrics provide a comprehensive view of how well the models align with ground truth labels.

Definition 3 (Accuracy) *In evaluating the performance of the models, we estimated how well the LLM predictions matched the actual ground truth labels. Mathematically, we measure accuracy as:*

$$Acc = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{Y}_i = Y_i)$$

Where: N is the total number of instances; $\hat{Y}_i$ is the predicted value for the i-th instance; Y_i is the ground truth value for the i-th instance; $\mathbb{I}(\hat{Y}_i = Y_i)$ is an indicator function that equals 1 if the prediction $\hat{Y}_i$ matches the ground truth Y_i, and 0 otherwise.

Definition 4 (F1 Score:) *The F1 score is a harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful for imbalanced datasets. The F1 score is calculated as follows:*

$$\text{F1 Score} = 2 \cdot \frac{\textit{Precision} \cdot \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

Definition 5 (Weighted-Average F1 Score:) *The weighted average F1 score calculates the F1 score for each class independently and then combines them using weights that are proportional to the number of true labels in each class.*

$$\text{Weighted-Average F1 Score} = \sum_{i=1}^{C} w_i \times \text{F1 Score}_i$$

where

$$w_i = \frac{\text{No. of samples in class } i}{\text{Total number of samples}}$$

and C is the number of classes in the dataset.

B Model Evaluation Setup

For this task, we utilized EleutherAI's open-source Language Model Evaluation Harness (lm-eval) framework (Gao et al., 2024). We created custom configurations for each task and looked at log-likelihood prediction for each possible token and decided possible generation from the possible class outputs. we created 3 different prompts for each data sources and evaluated on same generation settings.

C Dataset Description

Table 4, 5, and 6 present the features included in the datasets. We use the target features as output classes, and for serializations that convert feature names to text, we correct spelling to improve clarity and expressiveness.

Feature Name	Description
Loan_ID	Unique identifier for the loan
Gender	Gender of the applicant
Married	Marital status of the applicant
Dependents	Number of dependents of the applicant
Education	Education level of the applicant
Self_Employed	Whether the applicant is self-employed
ApplicantIncome	Income of the applicant
CoapplicantIncome	Income of the co-applicant
LoanAmount	Loan amount requested
Loan_Amount_Term	Term of the loan in months
Credit_History	Credit history of the applicant
Property_Area	Area type of the property
Loan_Status	Status of the loan (e.g., Loan paid or not)

Table 4: Description of Features for US Loan Predictions Dataset

Feature Name	Description
sex	Gender of the applicant
amnt req	Amount requested for the loan
ration	Ratio of the amount granted to the amount requested
maturity	Maturity period of the loan
assets val	Value of the applicant's assets
dec profit	Decision on the profit potential
xperience	Experience of the applicant
educatn	Education level of the applicant
age	Age of the applicant
collateral	Collateral provided for the loan
locatn	Location of the applicant
guarantor	Guarantor for the loan
relatnshp	Relationship with the financial institution
purpose	Purpose of the loan
sector	Economic sector of the applicant
savings	Savings of the applicant
target	Loan amount requested granted or not

Table 5: Description of Features for Ghana Credit Rationing Dataset

Feature Name	Description
gender	The gender of the individual
checking_status	The status of the individual's checking account
duration	Duration of the credit in months
credit_history	Credit history of the individual
purpose	Purpose of the credit
credit_amount	Amount of credit requested
savings_status	Status of the individual's savings account
employment	Employment status of the individual
installment_commitment	Installment commitment as a percentage of disposable income
other_parties	Other parties related to the credit
residence_since	Number of years the individual has lived in their current residence
property_magnitude	Value or magnitude of property
age	Age of the individual
other_payment_plans	Other payment plans that the individual has
housing	Housing status of the individual
existing_credits	Number of existing credits at this bank
job	Job status of the individual
num_dependents	Number of dependents
own_telephone	Whether the individual owns a telephone
foreign_worker	Whether the individual is a foreign worker
class	Classification of the credit (e.g., good or bad)

Table 6: Description of Features in German Credit Dataset

D Serialization

Table 7 shows examples of the six (6) different serialization methods employed in this work. We considered straightforward default values, such as `JSON` and `List`, to more structured and natural language text-like formats, such as `HTML`, `Latex`, `Text` (Hegselmann et al., 2023), `GReaT` (Borisov et al., 2022) and `LIFT` (Dinh et al., 2022).

Serialization	**Example Template**
JSON (default)	{age: 32, sex: female, loan duration: 48 months, purpose: education}
List	- age: 32 - sex: female - loan duration: 48 months - purpose: education
GReaT (Borisov et al., 2022)	age is 32, sex is female, loan duration is 48 months, loan purpose is education
Text	The age is 32. The sex is female. The loan duration is 48 months. The purpose is education.
LIFT (Dinh et al., 2022)	A 32-year-old female is applying for a loan for 48 months for education purposes.
HTML	<table><thead> <tr><th>age</th> <th>sex</th> . . . <tr><td>32</td><td>female</td> . . . </tr> </tbody></table>
Latex	\begin{tabular}{lrrr} \toprule age & sex & loan duration & purpuse \\ \midrule 32 & female & 48 month & education \\ \end{tabular}

Table 7: **Comparison of serialization formats for loan applicant information.** This table presents example templates for representing loan applicant data with four features (age and sex, loan duration and purpose). JSON is assumed as the default format. The selected serialization formats ensure diverse data representation, balancing availability across different formats, naturalness, and alignment with prior work.

E Prompt Examples

In Figure 8, we employed straightforward and minimally complex prompts for the task to maintain simplicity and consistency. For each task, we carefully adapted the prompt while ensuring alignment with the specific requirements of the evaluation. However, we intentionally chose not to modify the output classes or introduce entirely new prompts across different tasks, as doing so could have introduced unintended variables that might influence the evaluation outcomes.

F More Fairness Scores

Below, we investigate additional questions, particularly the relationship between fairness scores and In-Context Learning (ICL) performance. Specifically, we analyze how variations in fairness scores impact ICL results, as illustrated in Figure 6.In Figure 7, we present the statistical parity difference across various serialization methods and models. This analysis aims to examine how different serialization techniques impact fairness, providing insights into potential biases introduced by these encoding strategies.. This exploration aims to provide deeper insights into potential biases and the extent to which fairness considerations influence model performance in different settings.

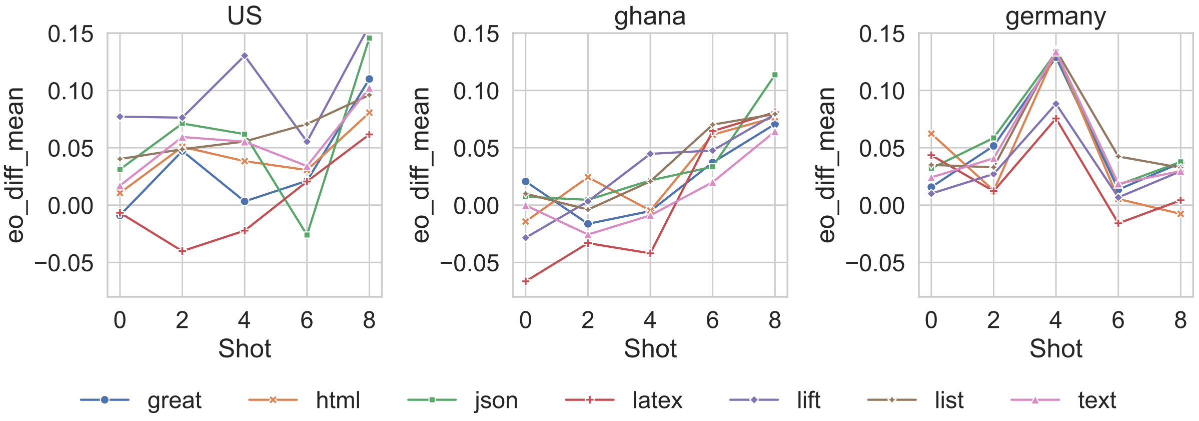

Figure 6: **Equality of Opportunity Difference for Few-Shot Learning Across Serialization Methods and Datasets** This figure shows the equality of opportunity difference for few-shot learning using different serialization methods across the three datasets. In-context learning (ICL) does not consistently reduce bias; in some cases, models exhibit significantly unfair behavior, particularly in certain shot configurations.

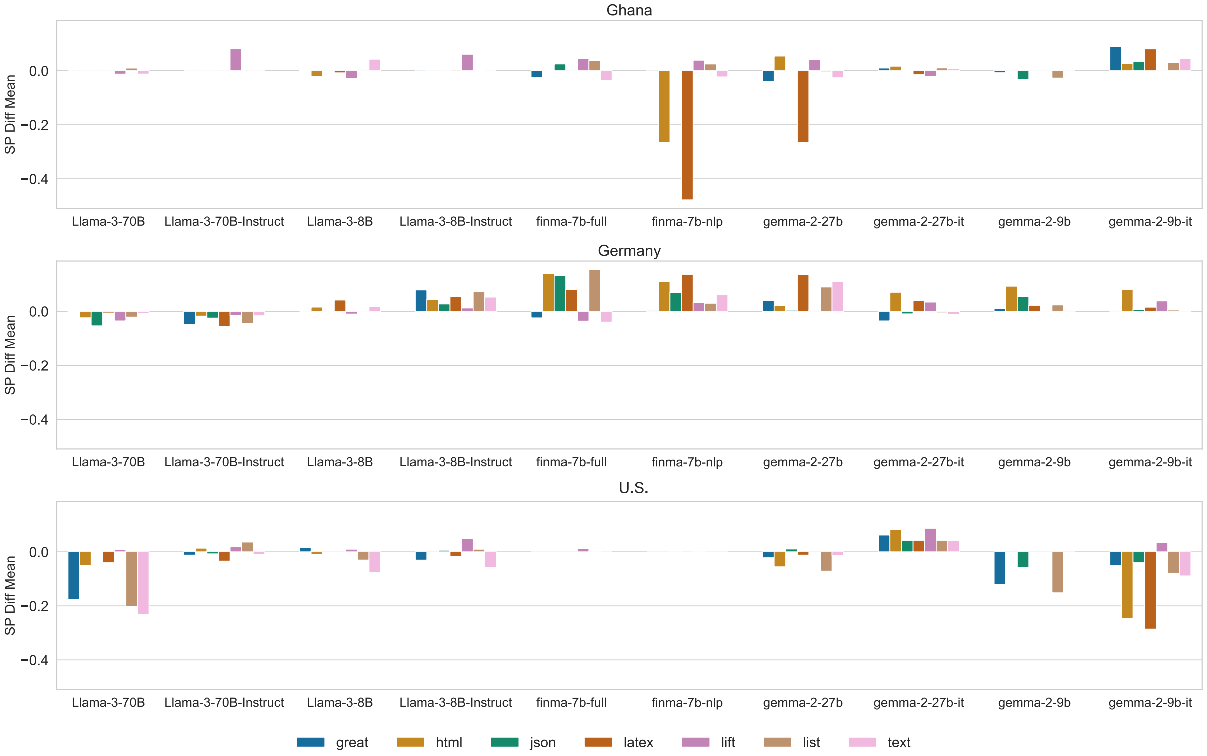

Figure 7: **Statistical Parity Difference Across Serialization Methods and Models** This figure illustrates the statistical parity difference for various serialization methods and models. We observe that financial models exhibit notably high bias, particularly for the Ghana and Germany datasets.

G In-Context Learning (ICL)

In the In-Context Learning (ICL) experiment depicted in Figure 8, we utilized the training set of our dataset to randomly select few-shot examples. Our findings indicate that ICL yields the most significant improvement when increasing from zero to two examples; however, subsequent increments in the number of examples does not result in similar returns. This observation aligns with existing research, which suggests that while ICL can be effective with a limited number of examples, its performance gains tend to plateau as more examples are added (Agarwal et al., 2025).

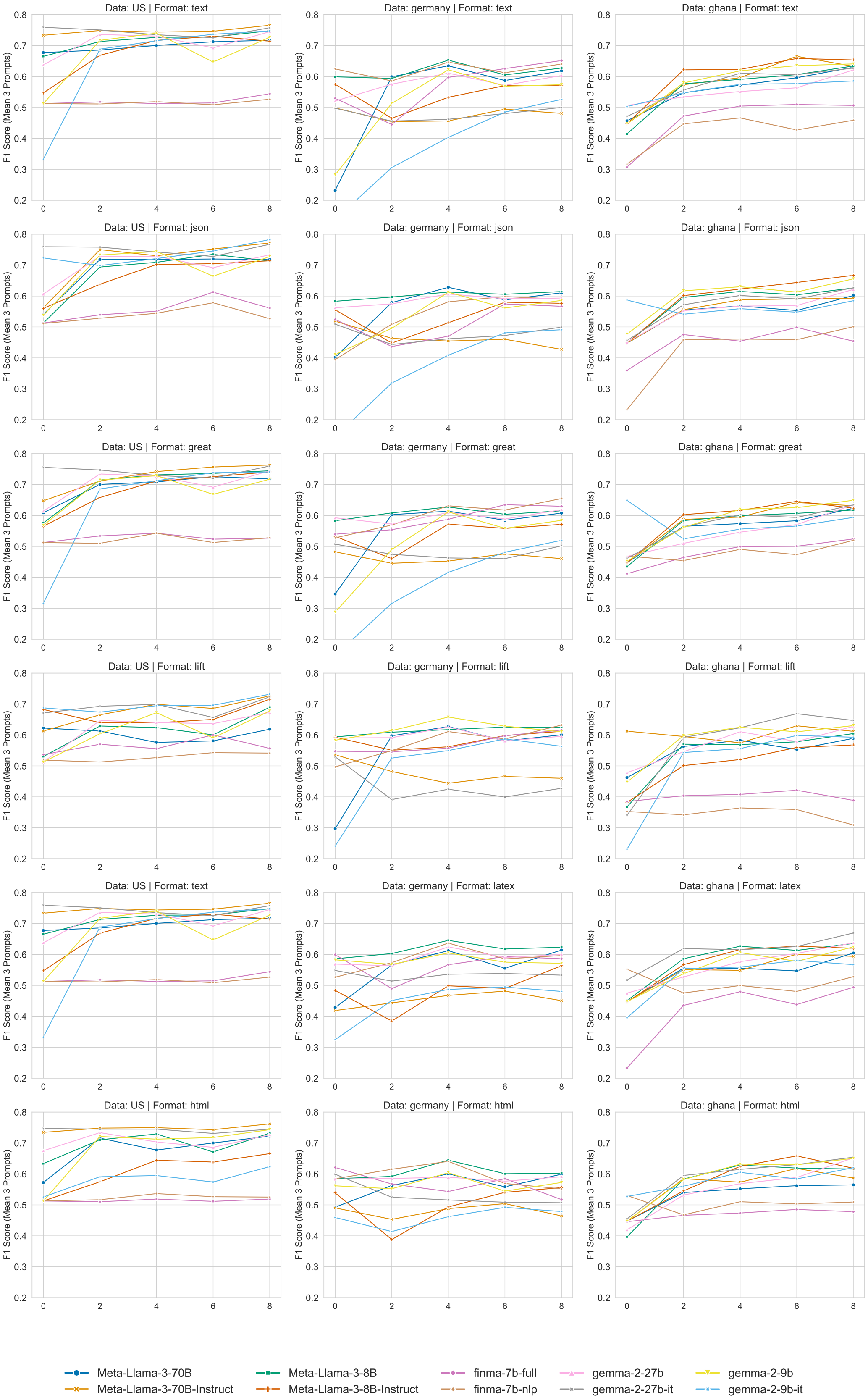

Figure 8: **Average F1 Score for Few-Shot Learning Across Different Serialization Methods** This figure presents the average F1 scores across various serialization methods for each dataset. We observe that the same models exhibit similar performance trends within each dataset, regardless of format. While the text format of the Ghana dataset may not share characteristics with the text format of the Germany dataset, Ghana's text and JSON formats display notable similarities.

Example of Loanpred Prompt
You are an expert LLM with extensive knowledge in financial analysis. Your task is to evaluate a person’s profile and determine their likelihood of repaying a loan. Respond with ’Y’ if the person is likely to repay the loan, and ’N’ if they are not likely to repay. Profile: {profile} Answer:
You are a financial analysis expert with extensive knowledge in credit assessment. Review the individual’s profile to evaluate their likelihood of repaying the loan. Respond with ’Y’ if repayment is likely, or ’N’ if it is not. Profile: {profile} Answer:
As a financial expert skilled in assessing creditworthiness, analyze the following profile and indicate the likelihood of loan repayment. Use ’Y’ for likely repayment and ’N’ for unlikely. Profile: {profile} Answer:
Example Of Germany Prompt
You are an expert LLM with extensive knowledge in financial analysis. Your task is to evaluate a person’s profile and determine their likelihood of repaying a loan. Respond with ’good’ if the person is likely to repay the loan, and ’bad’ if they are not likely to repay. Profile: {profile} Answer:
You are a financial assessment specialist with deep insights into creditworthiness. Review the profile below and indicate the repayment likelihood with ’good’ if the individual is likely to repay the loan, or ’bad’ if they are not. Profile: {profile} Answer:
Imagine you are a loan assessment expert with extensive experience in evaluating repayment potential. Analyze the details provided to judge whether repayment is probable. Use ’good’ for likely repayment and ’bad’ for unlikely. Profile: {profile} Answer:
Example Of Ghana Prompt
You are an expert LLM with extensive knowledge in financial analysis. Your task is to evaluate a person’s profile and determine their likelihood of repaying a loan. Respond with ’Yes’ if the person is likely to repay the loan, and ’No’ if they are not likely to repay. Profile: {profile} Answer:
You are a financial risk evaluator with expertise in creditworthiness. Review the individual’s profile and indicate their repayment likelihood. Use ’Yes’ for likely repayment, or ’No’ if repayment is unlikely. Profile: {profile} Answer:
As an expert in financial analysis, assess the following profile to determine the likelihood of loan repayment. Respond with ’Yes’ if repayment is probable, and ’No’ if it is not. Profile: {profile} Answer:

Table 8: **Example Prompts Used for the Task.** For each task, we created three distinct prompts, and the reported results represent the average performance across all three.

H Token Attribution explainability experiments

In understanding the decision processes made by LLMs we used *captum* (Kokhlikyan et al., 2020), an open-source model explainability library that provides a variety of generic interpretability methods. Our main question of interest in this work was to understand the interesting features that are used by LLMs in decision-making. In addition, we seek to understand the different decision-making characteristics observed between each LLM.

In this work, the main questions we have are; if LLMs are looking at interesting attributes to make decisions and what different decision-making characteristics are observed between each LLM.

We calculated token attribution for examples by replacing them with every possible item in the test set and assuming specific generation output. The results reported show representative values for the whole test set since we built our baseline tokens to be representative of the whole test set. Detailed visualization of the attribution is shown in Figures below.

The models explored in this study are medium-sized open-source models, chosen to balance computational efficiency and feasibility. The inclusion of larger models was limited due to computational overhead, while architectural complexities in Captum prevented the integration of financial models.

For the Ghana dataset, as shown in Figure 9 and Figure 10, we observed that `Gemma-2-9b-it` models primarily exhibit negative or neutral attributions from surrounding features for both positive and negative predictions. This behavior results in a slight performance gain, as presented in Table 3. Additionally, we found no consistent feature that LLMs consistently focus on, making the decision-making process highly model-dependent.

For the US data, as shown in Figure 14 and Figure 13, we observed that most decisions are influenced by the Loan_ID column, which contradicts the patterns observed by manual decision-makers. Unlike other datasets, the US data exhibits more consistent feature selection by LLMs, indicating a stronger alignment in the features they prioritize.

```
{'sex': 1, 'amnt req': 1500, 'ration': 1, 'maturity': 30.0, 'assets val': 2000, 'dec profit': 300.0,
'xperience': 1.0, 'educatn': 1, 'age': 53, 'collateral': 1500, 'locatn': 0, 'guarantor': 0, 'relatnshp': 1,
'purpose': 1, 'sector': 4, 'savings': 0}
```

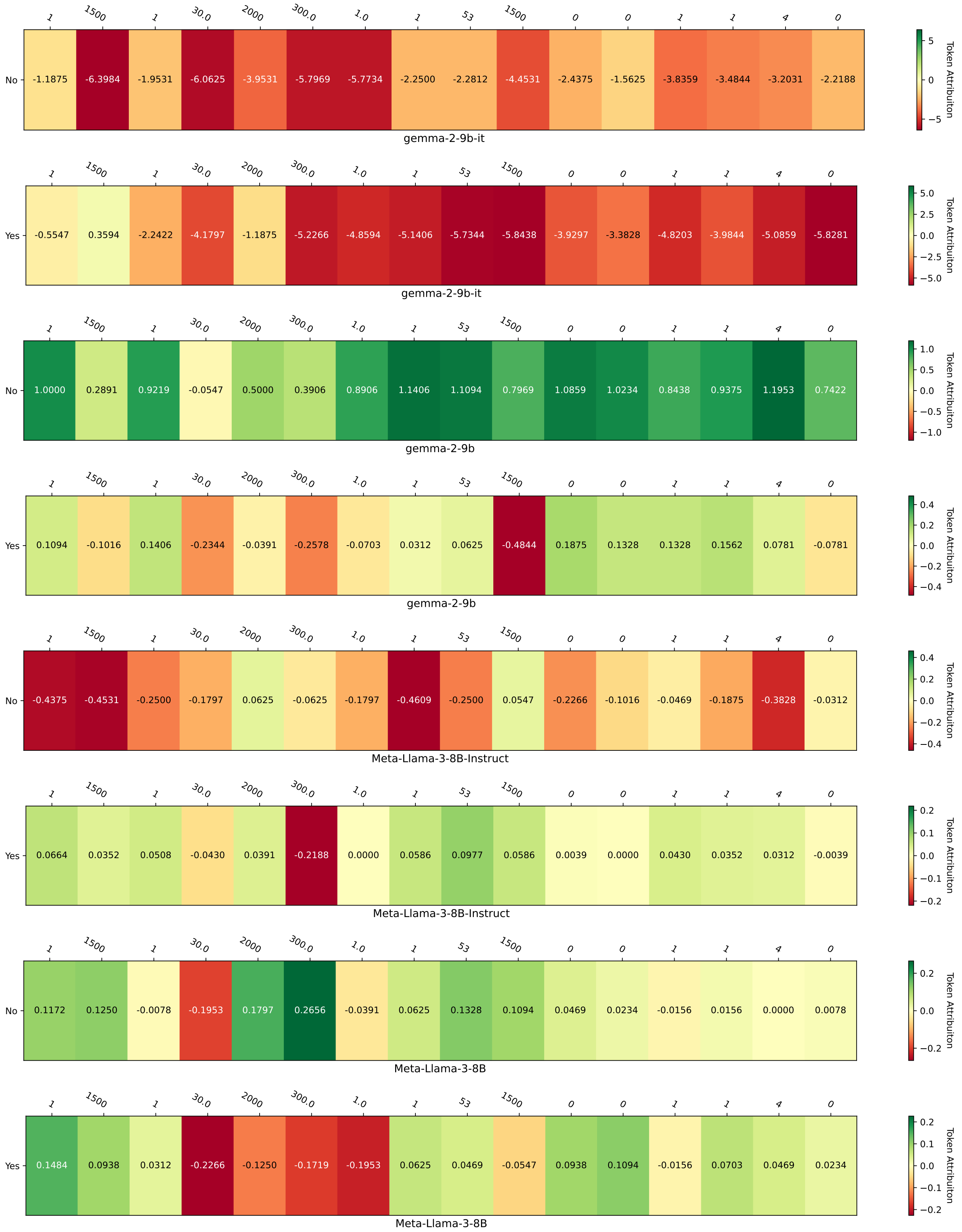

Figure 9: Attribution scores of Ghana data for example 1. Positive attribution scores are indicated in green, while negative scores are shown in red. We can see `Gemma-2-9b-it` models have more negative and neutral attribution scores completely different from their original model `Gemma-2-9b`.

```
{'sex': 0, 'amnt req': 9000, 'ration': 0, 'maturity': 30.0, 'assets val': 10000, 'dec profit': 900.0,
 'xperience': 3.0, 'educatn': 3,'age': 35, 'collateral': 9000, 'locatn': 1, 'guarantor': 0,
'relatnshp': 0, 'purpose': 1, 'sector': 4, 'savings': 1}
```

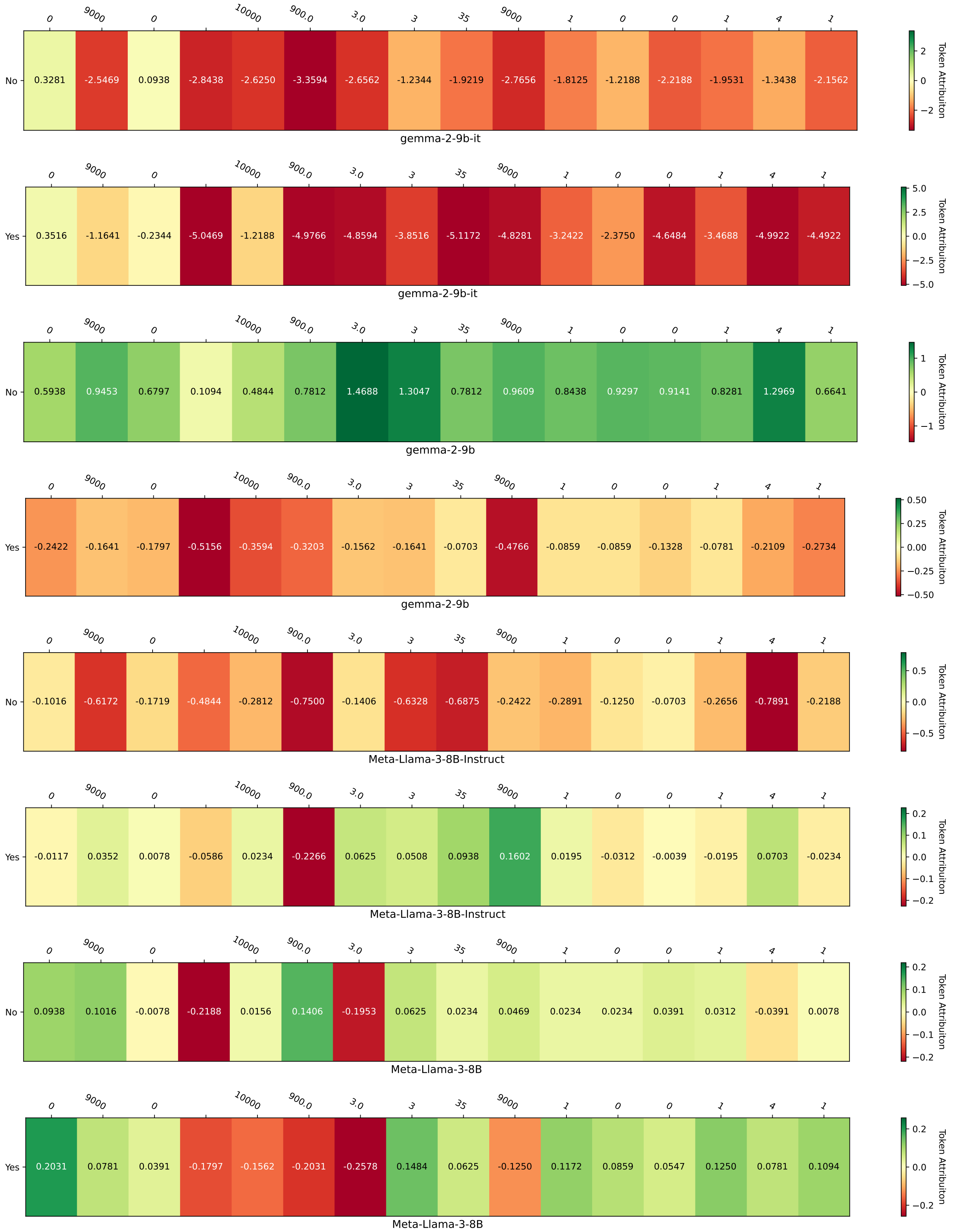

Figure 10: Attribution scores of Ghana data for example 2. Positive attribution scores are indicated in green, while negative scores are shown in red. `Gemma-2-9b-it` models show more negative and neutral token attribution.

```
{'gender': 'male','checking_status': "'no checking'", 'duration': 54, 'credit_history':
"'no credits/all paid'", 'purpose': "'used car'", 'credit_amount': 9436, 'savings_status':
"'no known savings'", 'employment': "'1<=X<4'",'installment_commitment': 2, 'other_parties': 'none',
'residence_since': 2, 'property_magnitude': "'life insurance'",'age': 39, 'other_payment_plans': 'none',
'housing': 'own', 'existing_credits': 1,'job': "'unskilled resident'", 'num_dependents': 2,
'own_telephone': 'none', 'foreign_worker': 'yes'}
```

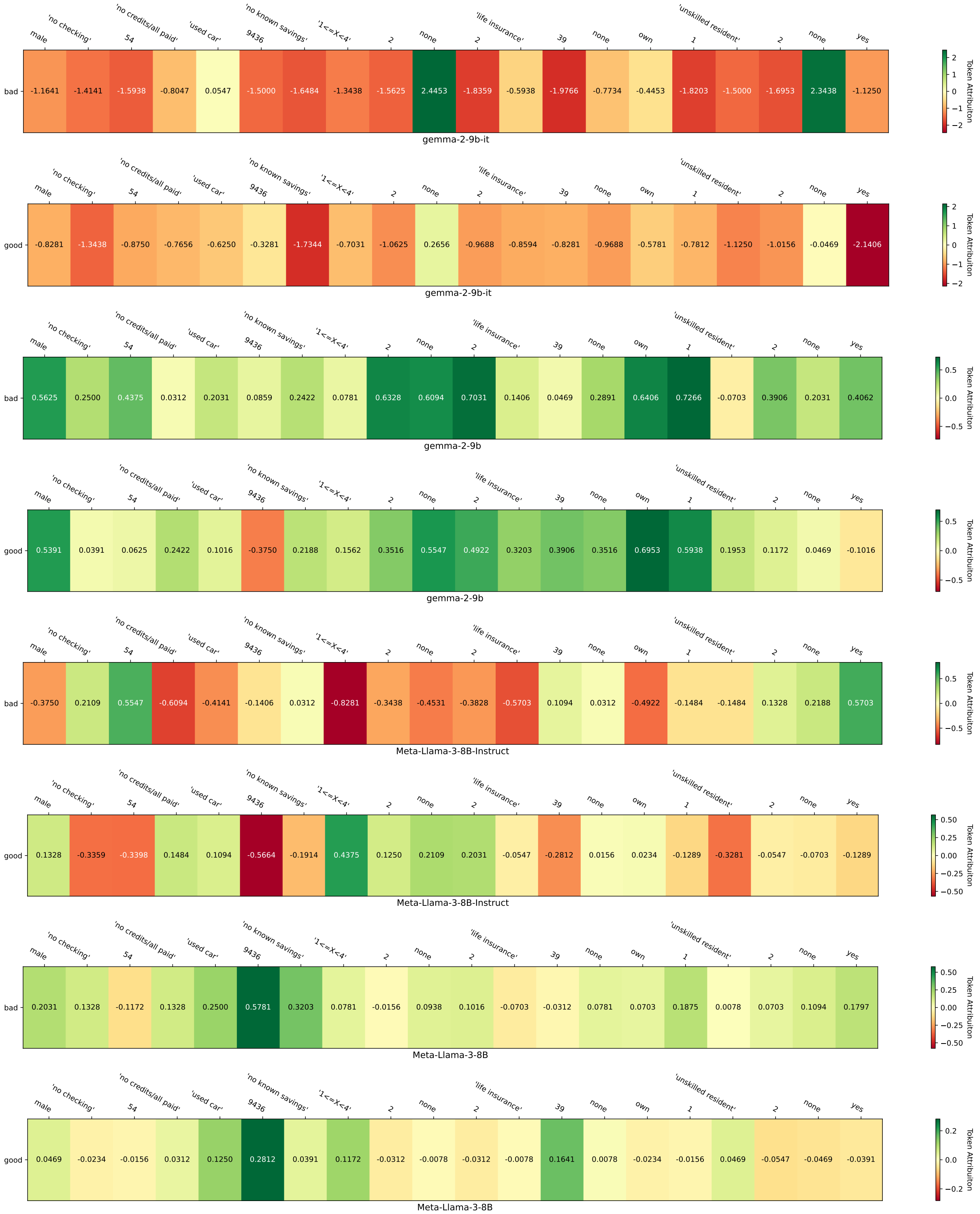

Figure 11: This figure displays the attribution scores for Example 1 of the Germany dataset. Positive attribution scores are indicated in green, while negative scores are shown in red. Gemma-2-9b-it models show high negative attribution from most features and we dont see focus on specific feature throughout the models.

```
{'gender': 'female','checking_status': "'<0'", 'duration': 18, 'credit_history': "'existing paid'",
'purpose': 'radio/tv', 'credit_amount': 3190, 'savings_status': "'<100'", 'employment': "'1<=X<4'",
 'installment_commitment': 2, 'other_parties': 'none', 'residence_since': 2,
 'property_magnitude': "'real estate'", 'age': 24, 'other_payment_plans': 'none','housing': 'own',
 'existing_credits': 1, 'job': 'skilled', 'num_dependents': 1,'own_telephone': 'none',
 'foreign_worker': 'yes'}
```

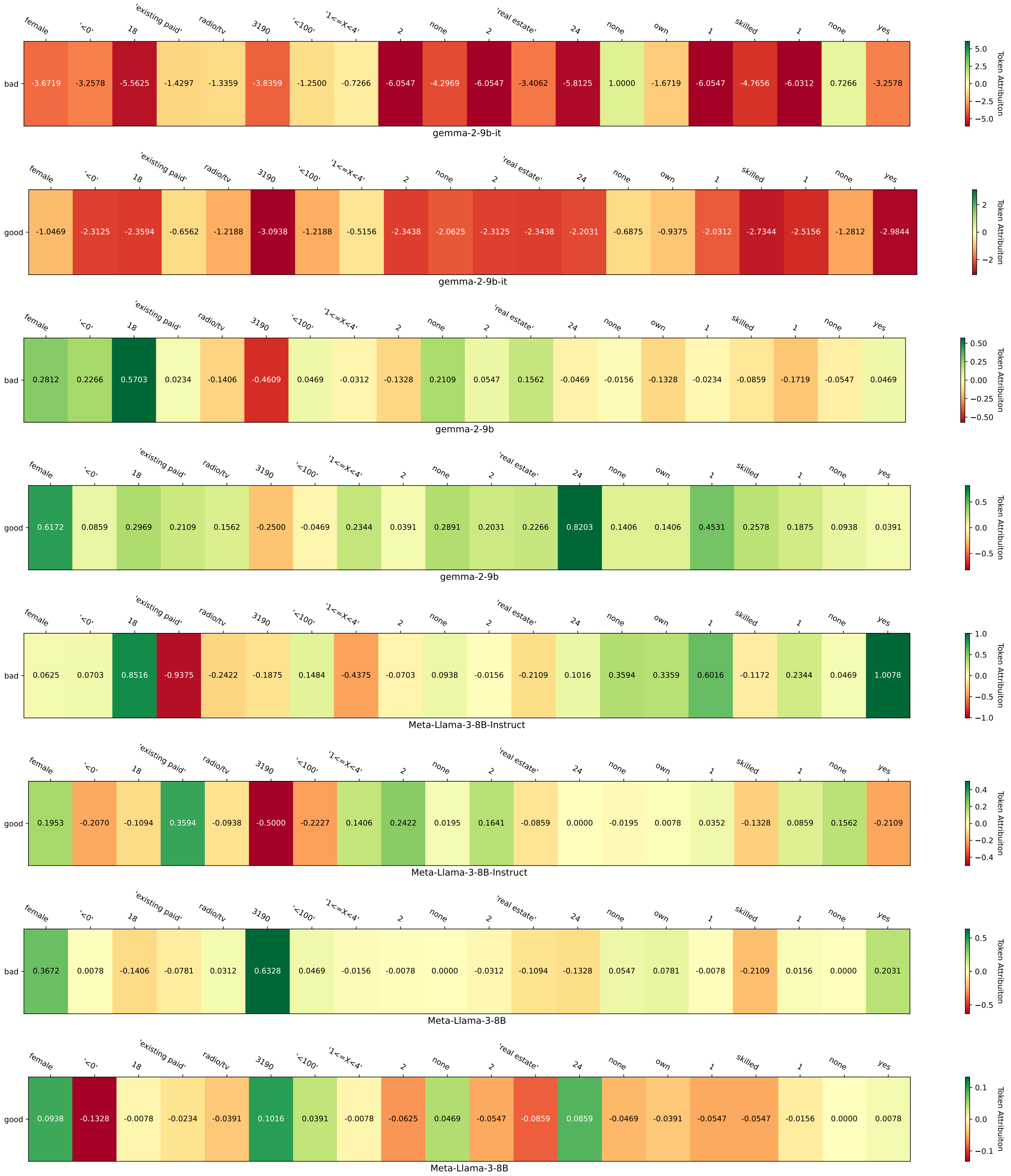

Figure 12: This figure displays the attribution scores for Example 2 of the Germany dataset. Positive attribution scores are indicated in green, while negative scores are shown in red. Gemma-2-9b-it models show high negative attribution from most features and we dont see focus on specific feature throughout the models.

```
{'Gender': 'Male','Loan_ID': 'LP002101', 'Married': 'Yes','Dependents': '0', 'Education': 'Graduate',
 'Self_Employed': None, 'ApplicantIncome': 63337,'CoapplicantIncome': 0.0, 'LoanAmount': 490.0,
 'Loan_Amount_Term': 180.0, 'Credit_History': 1.0,'Property_Area': 'Urban'}
```

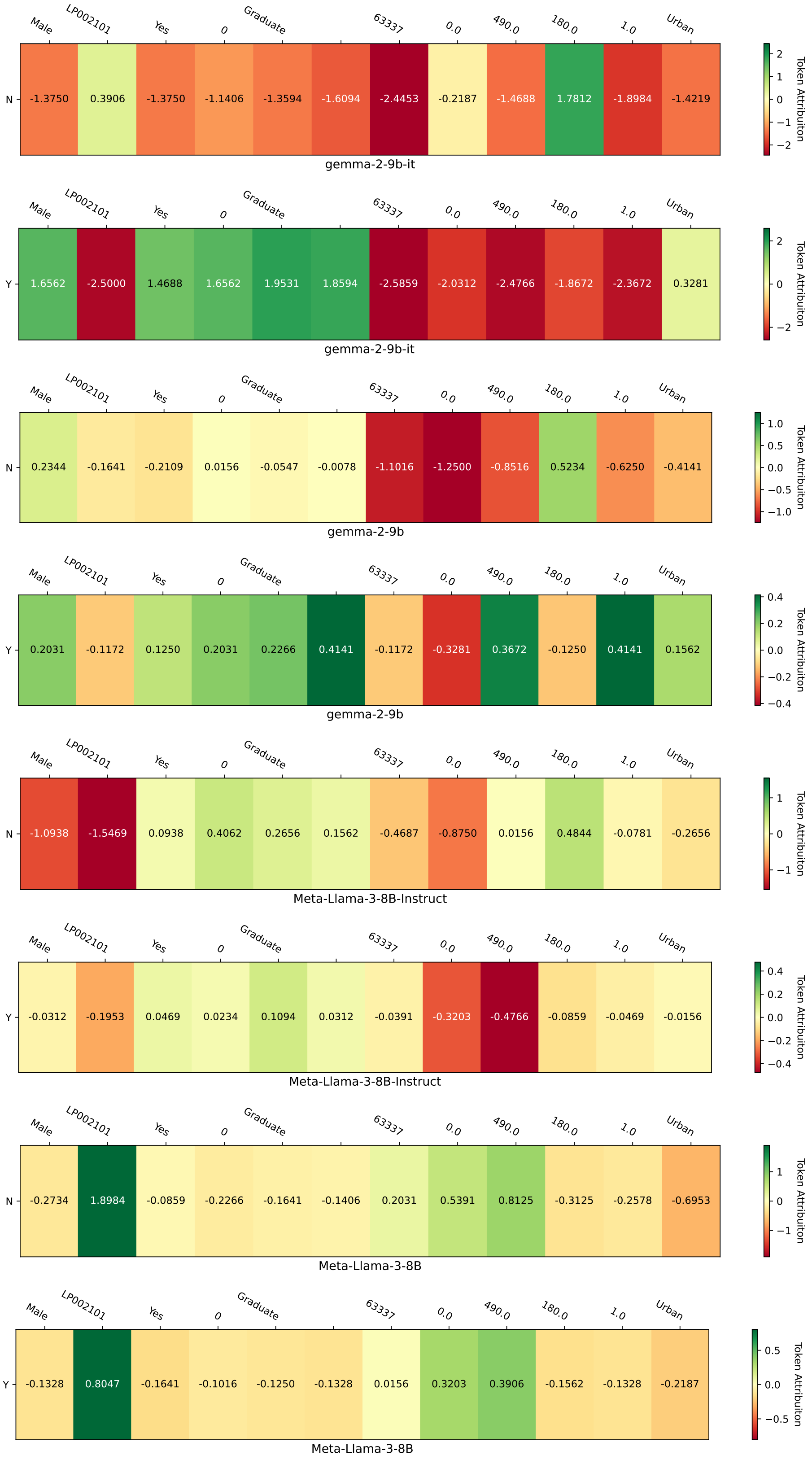

Figure 13: This figure displays the attribution scores for Example 1 of the US dataset. Positive attribution scores are indicated in green, while negative scores are shown in red. We can see the “Loan_ID” feature significantly influences the model’s output.

```
{'Gender': 'Female','Loan_ID': 'LP002978', 'Married': 'No', 'Dependents': '0','Education': 'Graduate',
 'Self_Employed': 'No', 'ApplicantIncome': 2900, 'CoapplicantIncome': 0.0,'LoanAmount': 71.0,
 'Loan_Amount_Term': 360.0, 'Credit_History': 1.0, 'Property_Area': 'Rural'}
```

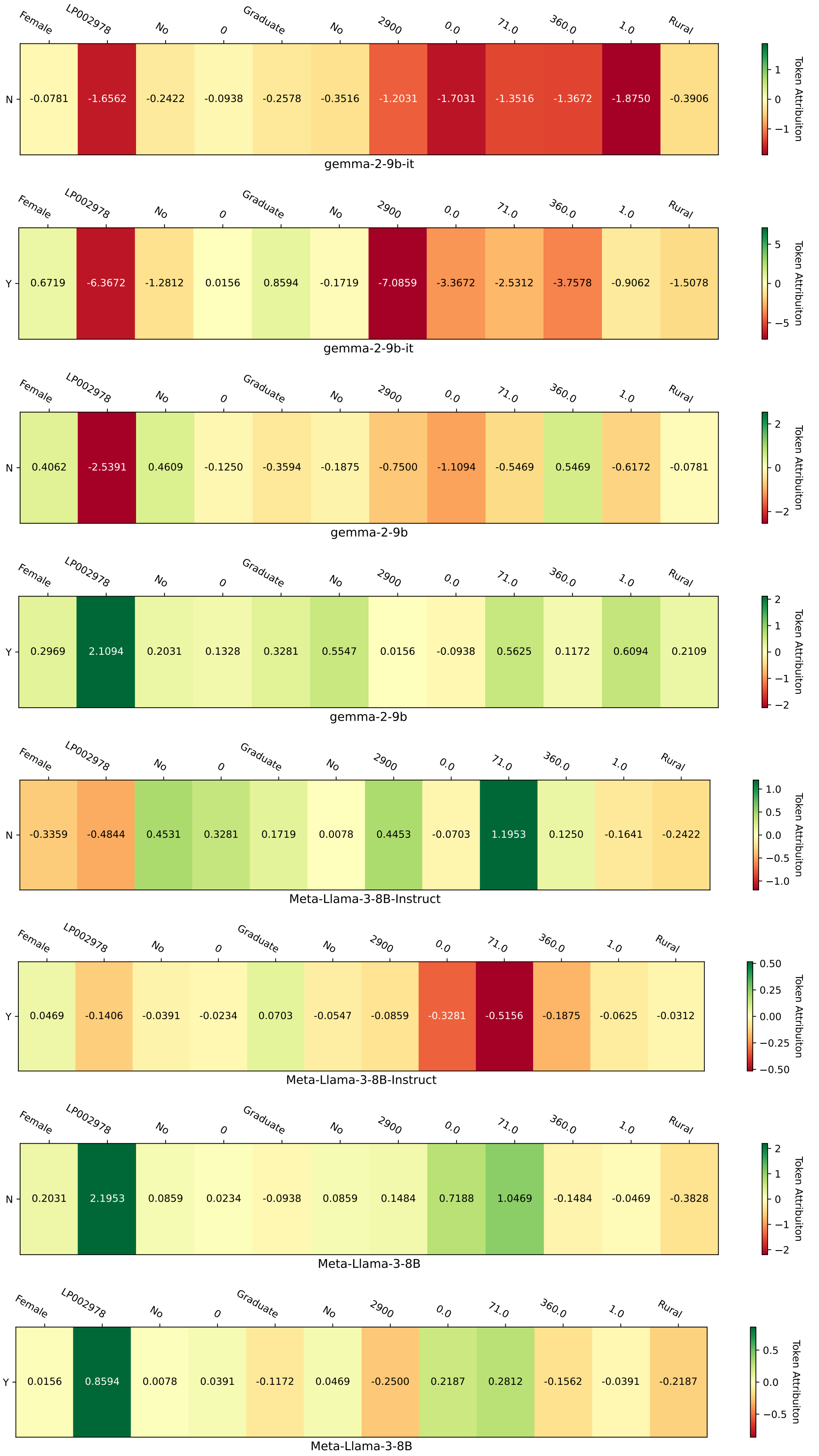

Figure 14: This figure displays the attribution scores for Example 2 of the US dataset. Positive attribution scores are indicated in green, while negative scores are shown in red. We can see the “Loan_ID” feature significantly influences the model’s output.