Accept or Deny? Evaluating LLM Fairness and Performance in Loan Approval across Table-to-Text Serialization Approaches

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly employed in high-stakes decision-making tasks, such as loan approvals. While their applications expand across domains, LLMs struggle to process tabular data, ensuring fairness and delivering reliable predictions. In this work, we assess the performance and fairness of LLMs on serialized loan approval datasets from three geographically distinct regions: Ghana, Germany, and the United States. Our evaluation focuses on the model's zero-shot and in-context learning (ICL) capabilities. Our results reveal that the choice of *serialization*[1] format significantly affects both performance and fairness in LLMs, with certain formats such as `GReaT` and `LIFT` yielding higher F1 scores but exacerbating fairness disparities. Notably, while ICL improved model performance by 4.9-59.6% relative to zero-shot baselines, its effect on fairness varied considerably across datasets. 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

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, it is crucial to understand the behavior and ethical implications of these systems due to their direct and severe impact on individuals (Aguirre et al., 2024). Financial decision-making is the systematic process of analyzing information to make informed choices in financial tasks such as investment, loan approval, and more (Kazemian et al., 2022).

[1]-Serialization refers to the process of converting tabular data into text formats suitable for processing by LLMs.

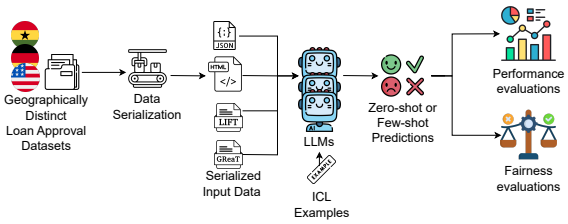

Figure 1: **Overview of our approach.** We investigated the impact of different serialization approaches on model fairness and performance by applying zero- and few-shot learning to datasets from three geographically distinct regions: Ghana, Germany and the United States.

In this work, we focus on loan approval, where a bank must decide whether or not to grant a loan based on the applicant's creditworthiness. This task is typically performed by loan officers who consider various input factors to make informed decisions. Loan approval is a critical task to explore as it directly impacts financial inclusion, borrower outcomes, and institutional risk management, making it an ideal domain for assessing the effectiveness and fairness of LLM-driven decision-making systems. Moreover, given the diversity in financial practices and socioeconomic contexts, evaluating loan approval across datasets from three distinct geographical regions (Ghana, Germany, and the United States) provides valuable insights into how LLMs manage data diversity and fairness within varying economic environments. Additionally, tabular nature of the datasets in this study underscores the importance of selecting an appropriate serialization method before feeding data into LLMs, as it can significantly influence model performance and fairness (Singha et al., 2023; Sui et al., 2024).

Building upon these observations, we frame our study around the following research questions: i) How do different serialization formats (e.g., `JSON`, `Text`, `GReaT`, `LIFT`) impact the fairness and performance of LLMs in loan approval tasks across diverse geographical datasets? ii) What effect does

in-context learning (ICL) have on the fairness and predictive performance of LLMs in loan approval scenarios, particularly when applied to datasets from Ghana, Germany, and the United States? iii) How do financial domain-specific LLMs compare to general-purpose LLMs in their ability to accurately and fairly assess loan applications, especially under zero-shot and few-shot learning settings? iv) What key factors contribute to fairness disparities in LLM-generated loan approval predictions, and how do these factors vary across different serialization methods and geographical regions?

To address the research questions outlined above, this work makes the following contributions:[2]

1. Investigate the capability of LLMs in financial decision-making, focusing on loan approval tasks. This includes a comprehensive zero-shot benchmark evaluation of various LLMs and an analysis of the features they prioritize in their decision-making process.
2. Analyze the impact of different tabular serialization formats on the decision-making process of LLMs.
3. 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. Examine the presence of gender-related biases in LLM-generated financial decisions, assessing their implications and associated risks.

Data Name	Size	#Features	Output
Ghana	614	13	Yes/No
Germany	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), Germany (Statlog) and United States (Kaggle). See Appendix D for details of the feature description of each dataset.

2 Related Work

LLMs in financial decision-making. Large Language Models (LLMs) have been employed to support various financial decision-making tasks, encompassing diverse applications such as stock trading (Ding et al., 2024), investment management (Kong et al., 2024), and credit scoring (Feng et al., 2023). These models either provide recommendations on optimal investment strategies to maximize returns or assess an individual's financial reliability and creditworthiness (Haque and Hassan, 2024). Loan approval tasks, in particular, carry significant risk due to their direct impact on financial inclusion and access to capital, making the evaluation of fairness and predictive accuracy in such models critically important (Kanubala et al., 2024).

Serialization in LLMs. LLMs require tabular data to be serialized into natural text, a process known as serialization (Jaitly et al., 2023). However, 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 serialization 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 Hegselmann et al. (2023) `Text` and `List` formats, 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 Hollmann et al. (2022) introduce TabPFN, a 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. LLMs 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 or racial biases present in the training data can result in unfair treatment of certain groups (Bolukbasi et al., 2016; Abid et al., 2021). Additionally, (Aguirre et al., 2024) highlight that the choice of in-context examples significantly influences model fairness, particularly when these examples are not demographically representative. Addressing these biases is crucial to ensure fair and ethical use of LLMs in decision-making processes.

Our study examines the use of LLMs for loan

[2]The evaluation code and data for each serialization strategy will be made publicly available upon acceptance.

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. Table 8 in Appendix E shows examples for the List, Text, HTML and Latex format.

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 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. Few-shot examples are top-n examples balanced by gender to align with fairness metrics.

3.2 Datasets

Dataset choice. Guided by data availability and relevance, we selected three distinct datasets representing the region's socioeconomic context. We posit that geographical, political, and ideological differences across regions directly influence financial practices, such as loan acquisition. The regions examined were arbitrarily chosen for this study; while expanding to more diverse regions is feasible, we have limited our scope to maintain a focused analysis. The distinct differences in data properties highlight the geographical variations central to this study. Although the task remains the same, subtle disparities within datasets from specific groups may introduce biases that can impact decision-making.

A comparison of dataset characteristics reveals distinct patterns across the Germany, Ghanaian, and U.S. datasets, as further detailed in the Appendix D. Only the Germany and Ghana datasets include age as a feature, with Germany applicants predominantly in their 20s and Ghanaian applicants in their 40s. The U.S. dataset primarily emphasizes employment status, whereas the other datasets provide additional information on the number of years employed. Across all datasets, male applicants consistently outnumber female applicants. Notable variations are also observed in loan amount distributions: the Germany dataset presents a broader and more evenly distributed range of loan amounts, while the U.S. and Ghana datasets are concentrated on smaller loan amounts with higher frequency.

Data processing. We provide a summary of the dataset we used in the study in Table 1 with a detailed description in Appendix D. For each dataset, we split the dataset into 80% train and 20% test using stratified sampling based on gender feature. To convert each dataset to the formats shown in Table 2 we created custom functions and also used pandas [3] functions that change dataframe to HTML and Latex. See Table 8 in Appendix E for examples of Latex, Text, HTML and List formats.

3.2.1 Table-to-Text serialization

Converting tabular data to text (*serialization*) is essential, as the format can significantly influence LLM decision-making (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 8 in Appendix E. 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] https://pandas.pydata.org/

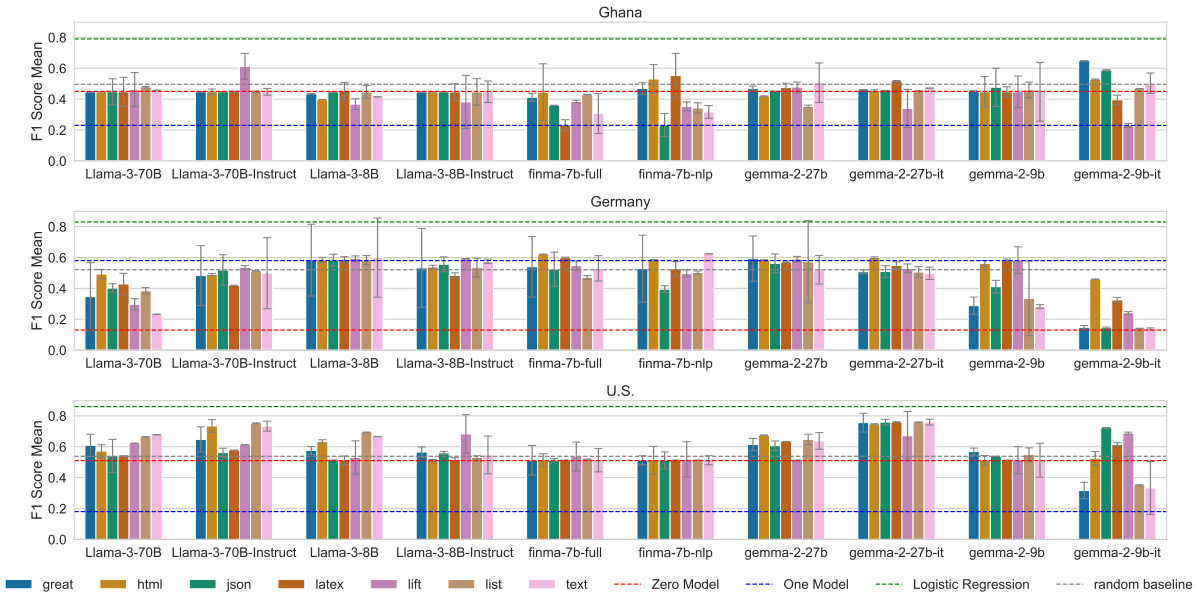

Figure 2: **Zero-shot weighted average F1 score performance of LLMs on loan approval tasks.** Evaluated across three prompts (variation shown by error bars) and multiple table-to-text serialization methods. The *Logistic Regression* model baseline (green dashed line) uses default `JSON` serialization with variables as individual features. Most LLMs underperform relative to this baseline, with only `GReaT` on Ghana, `List/Text` on Germany, and `Gemma-2-27b-it` on the U.S. showing modest improvements.

3.3 Models

3.3.1 Baseline and benchmark models

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

Baseline models. The *zero* model, *one* model and *Random* model serve as our simple baselines, as shown in Figure 2. 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. Finally, the *Random* model serves as a baseline by comparing the model's performance against randomly generated predictions[4].

Benchmark model. 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. Additionally, we used default parameters of scikit-learn[5] implementation for logistic regression to be used as basic comparison baseline. 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.

3.3.2 Large Language Models (LLMs)

We evaluated a total of ten (10) LLMs selected based on their open-source availability, instruction tuning, parameter size, and domain relevance (Table 3). To assess the effect of domain relevance, we included models specifically fine-tuned for financial tasks: `FinMA-7B-NLP` and `FinMA-7B-full`, introduced by Xie et al. (2023). To examine the effect of instruction tuning, we incorporated 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`. Each of these instruction-tuned variants was paired with its corresponding base model (`LlaMA-3-70B`, `LLaMA-3-8B`, `Gemma-2-27b`, and `Gemma-2-9b`) sourced from Touvron et al. (2023); Meta (2024); Team et al. (2024). This selection allows us to ex-

[4]We use NumPy with a fixed seed for reproducibility

[5]https://scikit-learn.org/

amine both the impact of instruction tuning and the role of model size, while also testing whether financial fine-tuning improves decision-making in domain-specific tasks such as loan approval. See Appendix C for model evaluation setup.

Model Name	Training Type	Params	Financial Dataset only
LLaMA-3	Pretrained & Instruction-tuned	8B & 70B	✗
Gemma-2	Pretrained & Instruction-tuned	9B & 27B	✗
FinMA-full	Fine-tuned	7B	✓
FinMA-NLP	Fine-tuned	7B	✓

Table 3: Overview of the ten LLMs evaluated, including models fine-tuned specifically on financial datasets.

3.4 Approaches to LLMs improvement

3.4.1 In-Context Learning (ICL)

In-context learning involves providing examples that 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 $n = 2, 4, 6, 8$. Our few-shot examples are strategically selected to ensure representational equity. For instance, when using two examples, one will correspond to a male and the other to a female, aligning with our fairness score metrics, which are based on gender representation.

3.5 Model and fairness evaluation

We use the weighted-average F1 score to evaluate model performance on the loan prediction task (see Appendix B for definitions). To assess fairness, we employ two standard metrics: equality of opportunity (EO) and statistical parity (SP). EO aligns with the goals of loan approval by ensuring that qualified applicants, regardless of group membership, have an equal chance of approval (Kozodoi et al., 2022). In contrast, SP measures whether approval rates are independent of sensitive attributes. 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 positive outcome is independent of the sensitive attribute.*

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

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:*

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

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

4 Results and Analysis

In this section, we present our results and analysis, structured around a set of research questions that guide the discussion. We begin by comparing the performance of different serialization methods across models for each dataset, as shown in Figure 2. We observe that the *zero* model outperforms the *one* model on the Ghana and United States (US) datasets, while the reverse is true for the Germany dataset. This suggests that the Germany dataset has a higher proportion of non-defaulters compared to the other two datasets. We also conducted experiments on model token attribution which is detailed in the Appendix I.

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

In Figure 2, we compare the zero-shot performance of LLMs against baseline models and analyse 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 of the models for the US, 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.

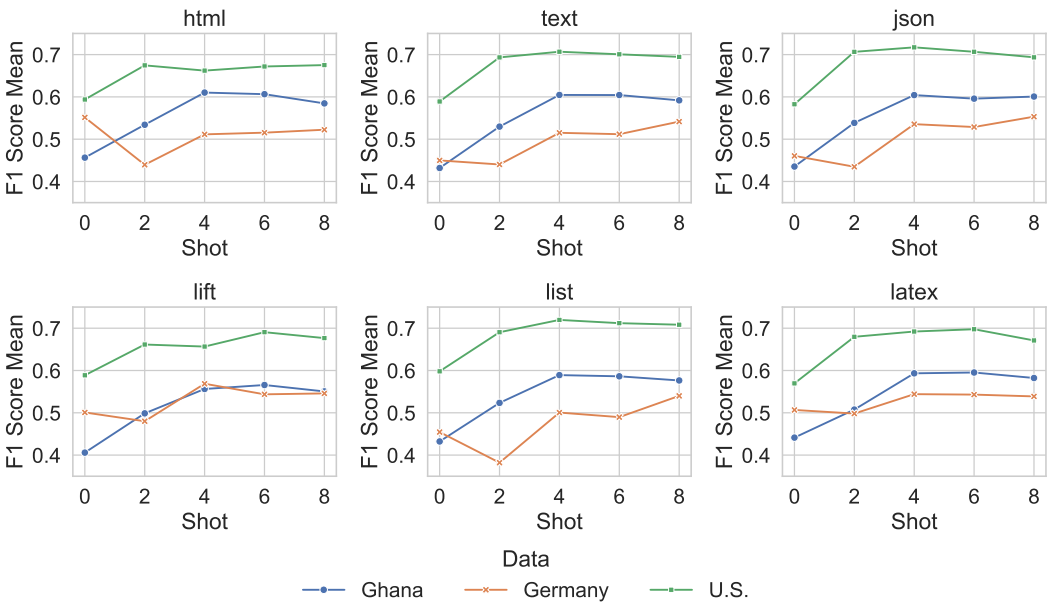

Figure 3: **Average weighted F1 score** trends across serialization formats for few-shot examples, showing higher gains in U.S. data across formats, while Germany lags consistently despite increasing shot numbers.

💡 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 *Logistic Regression* model.

4.2 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 2: For the *Ghana* dataset, the best performances are achieved using the `GReaT` (`Gemma-2-9b-it`) and `LIFT` (`LLaMA-3-70B-Instruct`) serialization method. In the *Germany* 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`.

💡 Serialization methods can significantly influence loan approval or denial, which, in turn, may have long-term consequences for individuals wrongly denied loans.

4.3 Does serialization using natural language texts improve performance?

We hypothesized that using more natural input text would improve model performance, which motivated our inclusion of the `LIFT` and `GReaT` serialization method (see Table 2). `LIFT` produced the best results for `LLaMA-3-70B-Instruct` on the Ghana dataset. However, this improvement did not hold consistently across all models and datasets, indicating that while natural language formats can be beneficial, their effectiveness is context-dependent.

💡 Our results show that increasing the naturalness of input formatting does not consistently enhance model performance.

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

Given LLMs' subpar performance in the zero-shot experiments, we explored various methods to improve their decision-making capabilities through in-context learning(ICL). Figure 3 presents the results from our ICL experiment, where we provide the model with varying numbers of n-shot exam-

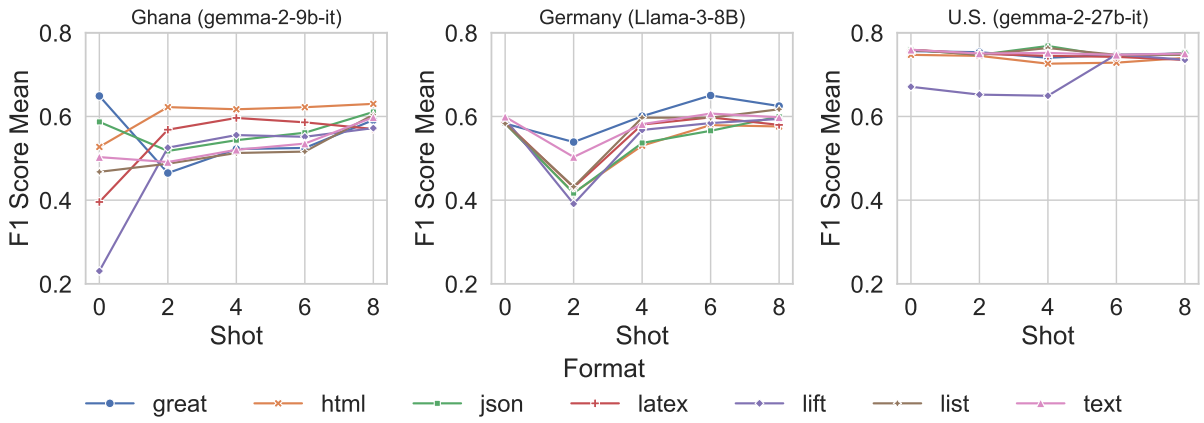

Figure 4: **Few-shot weighted F1 trends.** Adding a small number of in-context examples improves performance, while differences among serialization formats remain modest across datasets.

ples, ranging from zero-shot ($n = 0$) to 8-shot across datasets and serialization formats. From Figure 3, providing more examples improves the loan approval task. Similarly, in Figure 4, we see average improvement with more examples. This is shown across all the serialization methods.

> 💡 Model performance improves with more example shots, improving LLM decision-making for loan approval.

Datasets:	Germany		Ghana		U.S.	
Fairness Metrics:	SP	EO	SP	EO	SP	EO
Baseline models						
Zero	0.00	0.00	0.00	0.00	0.00	0.00
One	0.00	0.00	0.00	0.00	0.00	0.00
Random	0.15	0.38	0.02	-0.03	0.04	0.35
Benchmark model						
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 4: **Zero-shot fairness metrics across regions for `JSON` serialization.** The red color shows high bias across comparing models excluding baselines.

4.5 How does model fairness vary across datasets?

Baseline models from Table 4 all show no discrimination in terms of equality of opportunity (EO) and statistical parity (SP) except the *Random* model. However, we see high discrimination in terms of both EO and SP with the `FinMA-7B-full` for the Germany 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 Germany data. Further highlighting that different models penalize sensitive groups differently. Additionally, examing fairness by conducting few-shot experiments showed that few-shots examples (e.g., $n = 8$) can introduce significant fairness disparities in Equality of Opportunity (EO), with differences exceeding 0.10 for certain serialization methods in the Ghana dataset (see Figure 5).

> 💡 LLMs fine-tuned on financial datasets have the potential to amplify existing historical gender bias.

4.6 What is the fairness F1 score tradeoffs?

Following the best-performing models, as shown in Figure 4, we assess the fairness of these

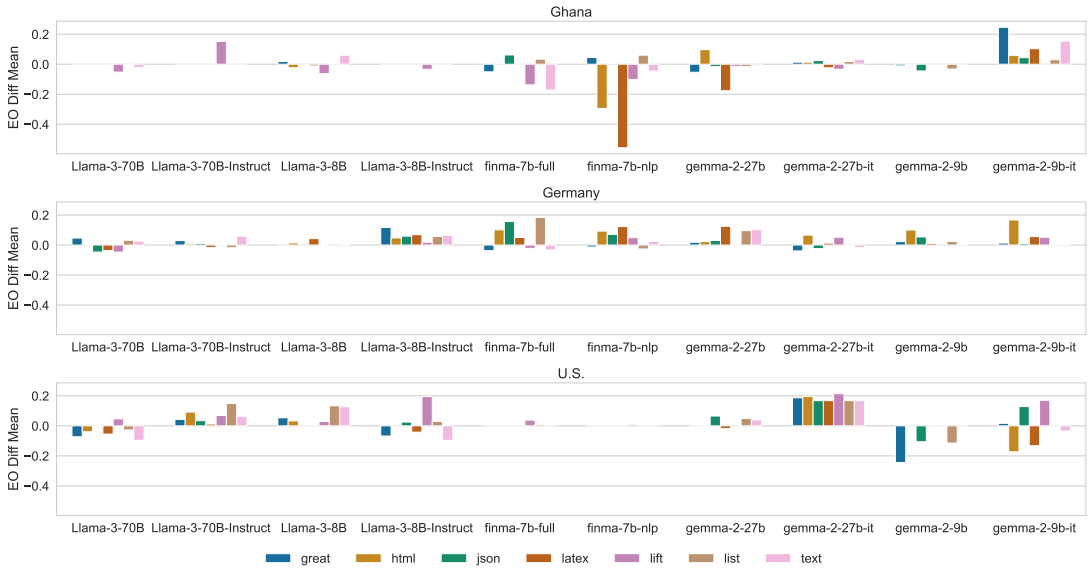

Figure 5: **Mean difference in EO for different serialization methods and models.** Finance-based models show higher gender-based disparity for certain serializations while the results are highly region and format-dependent.

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. The `Gemma-2-9b-it` model shows a higher disparity in EO difference. 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 and high performance does not equate to fairness.

5 Discussion and Conclusion

Summary. The ability of LLMs to handle structured tabular data for high-stakes tasks like loan approvals remains under-explored. This work evaluates how different serialization methods (`JSON`, `LIFT`, `Text`) and in-context learning (ICL) impact the fairness and accuracy of LLMs across diverse regional datasets (Ghana, Germany, United States). We find that, in zero-shot scenarios, all LLMs perform worse than a *Logistic Regression* model baseline, frequently defaulting to uniform approval or denial. Modest improvements only emerge with a few in-context examples, largely influenced by serialization format and dataset rather than model.

Fairness implications of LLMs in finance. The results indicate that LLMs fine-tuned on financial datasets cannot yet be fully trusted for high-stakes financial decisions. Therefore, careful attention to data representation is at least as critical as the choice of model. To further address fairness concerns, employing more balanced datasets and ensuring a transparent decision-making process could be beneficial. This transparency is particularly important, as prior decisions made by banks can significantly impact the long-term creditworthiness of applicants (Majumdar et al., 2025).

Recommendation for practitioners. We recommend that practitioners retain thoroughly validated tabular models as a baseline and treat LLM outputs only as decision support until they demonstrably exceed that baseline in both accuracy and fairness. During and after deployment, models should be stress-tested on multiple serialization approaches and on regionally diverse datasets to ensure robustness. Benchmarking must extend beyond raw performance scores to include a suite of fairness and accuracy metrics so that improvements in prediction quality do not mask emerging biases.

Future work. Explore serialization-robust training, fairness-aware optimization, interpretability methods that expose feature reliance, and broader multilingual datasets that capture diverse regions.

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.

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 due to resource constraints and limited flexibility for experimentation, particularly around fine-grained control of inputs and internal mechanisms.

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 I.

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*.

Carlos Alejandro Aguirre, Kuleen Sasse, Isabel Alyssa Cachola, and Mark Dredze. 2024. Selecting shots for demographic fairness in few-shot learning with large language models. In *Proceedings of the Third Workshop on NLP for Positive Impact*, Miami, Florida, USA. Association for Computational Linguistics.

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*.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

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.

Duanyu Feng, Yongfu Dai, Jimin Huang, Yifang Zhang, Qianqian Xie, Weiguang Han, Zhengyu Chen, Alejandro Lopez-Lira, and Hao Wang. 2023. Empowering many, biasing a few: Generalist credit scoring through large language models. *arXiv preprint arXiv:2310.00566*.

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.

FM Haque and Md Mahedi Hassan. 2024. Bank loan prediction using machine learning techniques. *arXiv preprint arXiv:2410.08886*.

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.

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*.

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*.

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

Deborah D Kanubala, Isabel Valera, and Kavya Gupta. 2024. Fairness beyond binary decisions: A case study on german credit. *European Workshop on Algorithmic Fairness*.

Siavash Kazemian, Cosmin Munteanu, and Gerald Penn. 2022. A taxonomical NLP blueprint to support financial decision making through information-centred interactions. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 89–98, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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*.

Yaxuan Kong, Yuqi Nie, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. Large language models for financial and investment management: Applications and benchmarks. *Journal of Portfolio Management*, 51(2).

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.

Ayan Majumdar, Deborah D Kanubala, Kavya Gupta, and Isabel Valera. 2025. A causal framework to measure and mitigate non-binary treatment discrimination. *arXiv preprint arXiv:2503.22454*.

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].

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.

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.

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.

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*.

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.

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 Additional Takeaways

How does prompt sensitivity vary across different regions and models?

The results in Figure 2 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.

How does model size relate to performance and fairness?

We assess the effect of model scale by evaluating multiple sizes of both the LLaMA and Gemma families in Figure 2. Across LLaMA variants, expanding parameter counts yields only marginal performance changes. In contrast, Gemma exhibits pronounced performance gains as size increases a pattern that re-emerges in the fairness analysis (Figure 5), where the Gemma models' equality-of-opportunity scores are highly sensitive to model scale.

Does Instruction Tuning Affect Model Performance and Fairness Scores?

We further investigated the impact of instruction tuning on accuracy and fairness by comparing the base and instruction-tuned variants of the LLaMA and Gemma families (Figure 2). Instruction tuning has little effect on LLaMA, but its influence on Gemma depends on model size: the 9 B version loses accuracy. These shifts are also shaped by the choice of serialization. For example, the instruction-tuned Gemma improves fairness on the United States dataset in some formats yet becomes more biased in others.

Do few-shot examples improve fairness?

In the Germany dataset, with reference to Figure 7, few-shot examples (e.g., $n = 8$) can lead to significant fairness disparities in equality of opportunity (EO), reaching differences of over 0.10 for some serialization methods in the Ghana datasets. The U.S. dataset shows greater sensitivity to few-shot examples, with models exhibiting a decline in fairness scores.

> 💡 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.

B 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 (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.

C 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.

D Dataset Description and Analysis

Table 5, 6, and 7 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 5: 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 6: 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 7: Description of Features in German Credit Dataset

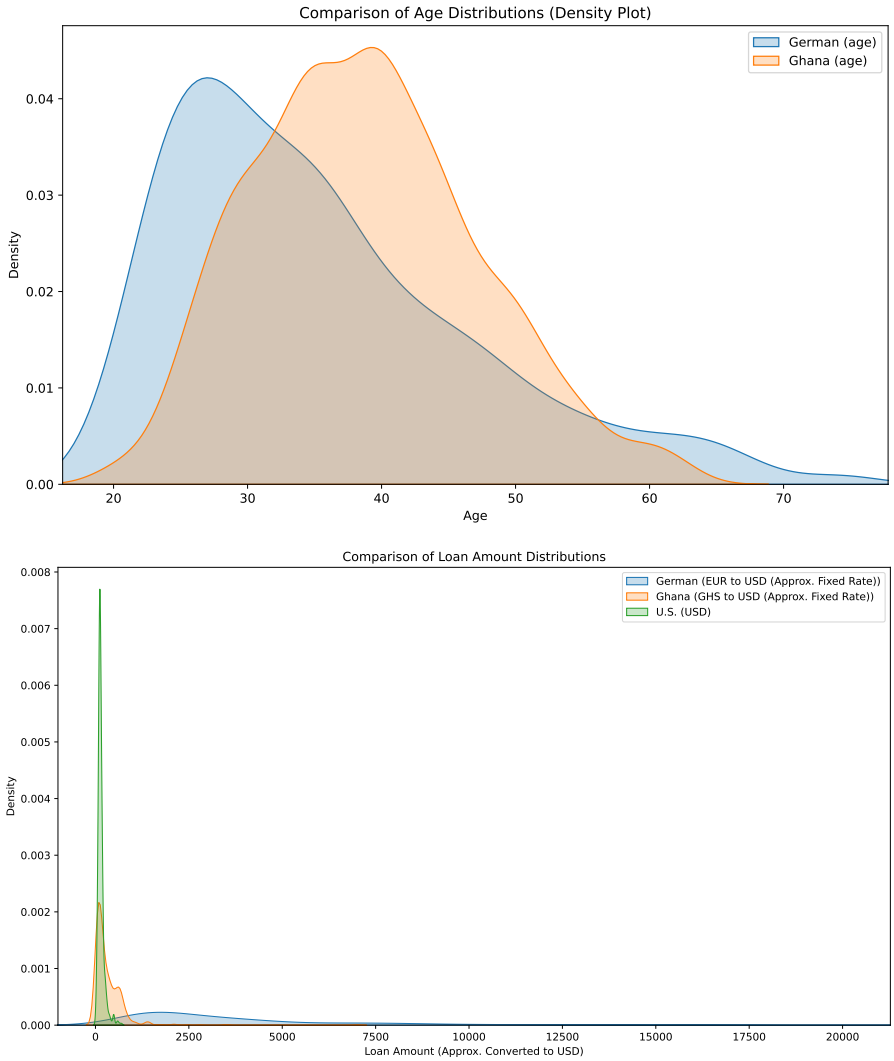

Figure 6: KDE plot comparing age and loan amount distributions across datasets, highlighting inherent socio-economic and cultural disparities. The age distribution reveals that the Ghana dataset skews older, with a concentration in the 30-50 age range, while the German dataset shows a relatively younger distribution peaking around the 20-30 age range. Loan amounts are predominantly smaller in both Ghana and U.S. datasets, with the German dataset exhibiting a broader distribution range, indicating socio-economic and lending disparities across regions.

E Serialization

Table 8 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 8: **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.

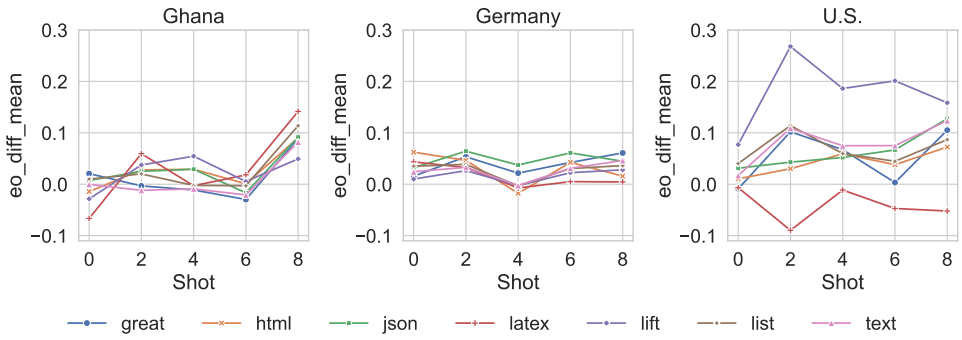

Figure 7: **Equality of Opportunity Difference for Few-Shot Learning Across Serialization Methods and Datasets.** In-context learning (ICL) does not consistently reduce bias; in some cases, models exhibit significantly unfair behavior, particularly in certain shot configurations.

F More Fairness Scores

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 7. In Figure 8, 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 8: **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 Prompt Examples

In Figure 9, 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.

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 9: **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 In-Context Learning (ICL)

In the In-Context Learning (ICL) experiment shown in Figure 9, we selected balanced few-shot examples from the training set, ensuring that each set of n examples was predetermined and included a balanced representation of the gender feature. 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).

Looking at Figure 9, 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.

I 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 10 and Figure 11, 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 2. 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 15 and Figure 14, 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.

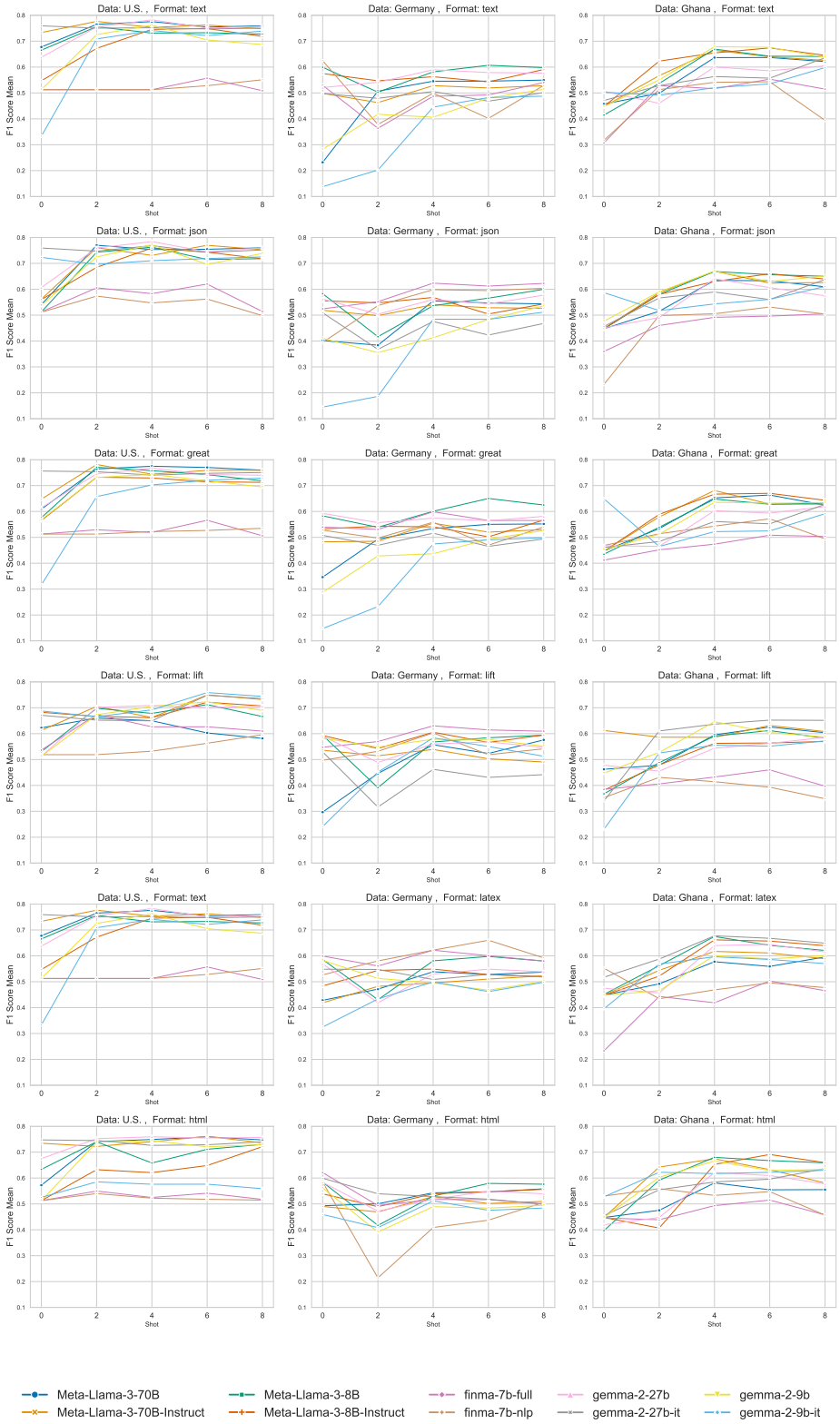

Figure 9: **Average F1 Score for Few-Shot Learning Across Different Serialization Methods** This figure presents the average F1 scores across various serialization methods for each dataset. We observe that the same models exhibit similar performance trends within each dataset, regardless of format. While 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.

{'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}

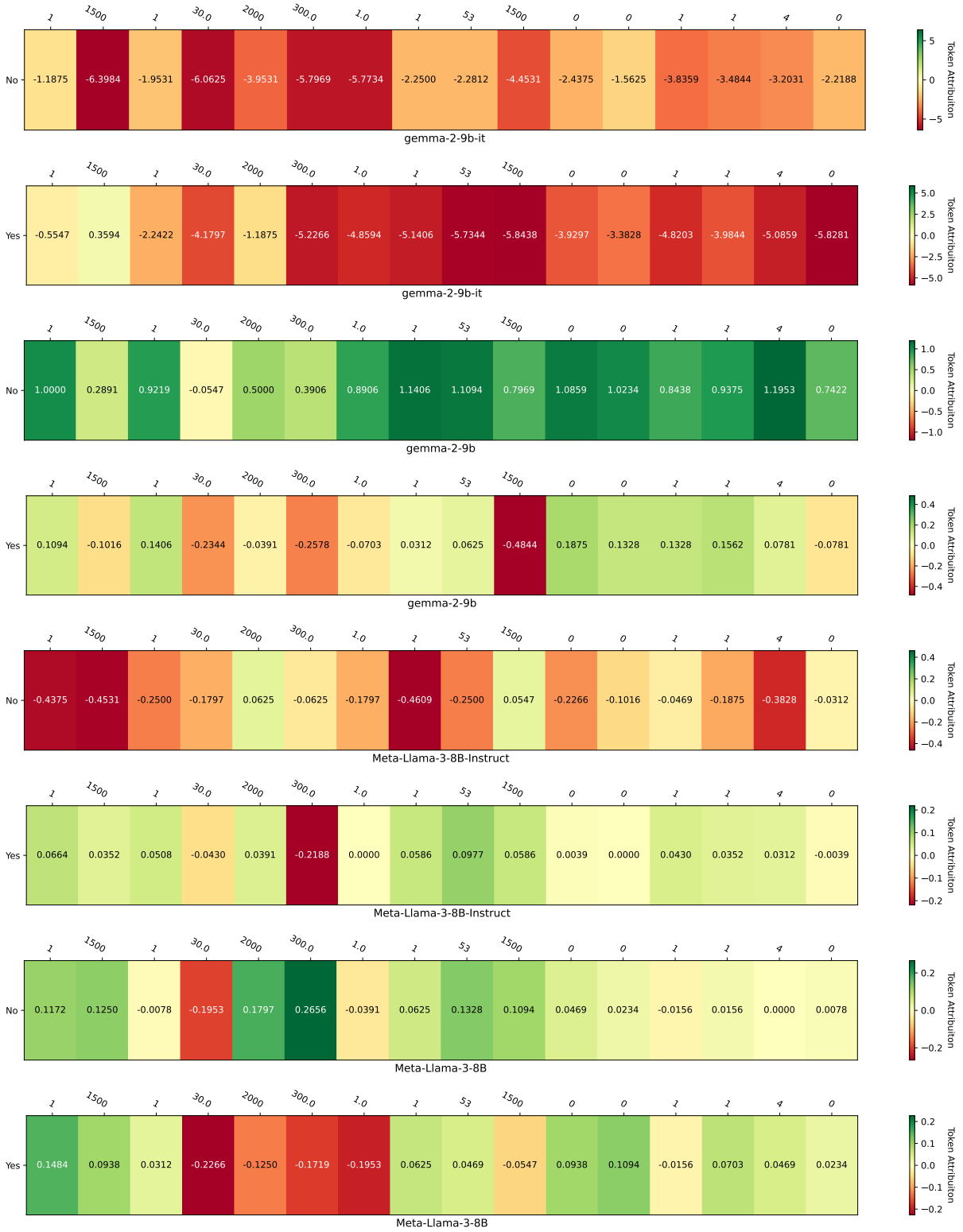

Figure 10: Attribution scores of Ghana data for example 1. Positive attribution scores are indicated in green, while negative scores are shown in red. We can see `Gemma-2-9b-it` models have more negative and neutral attribution scores completely different from their original model `Gemma-2-9b`.

\{'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\}

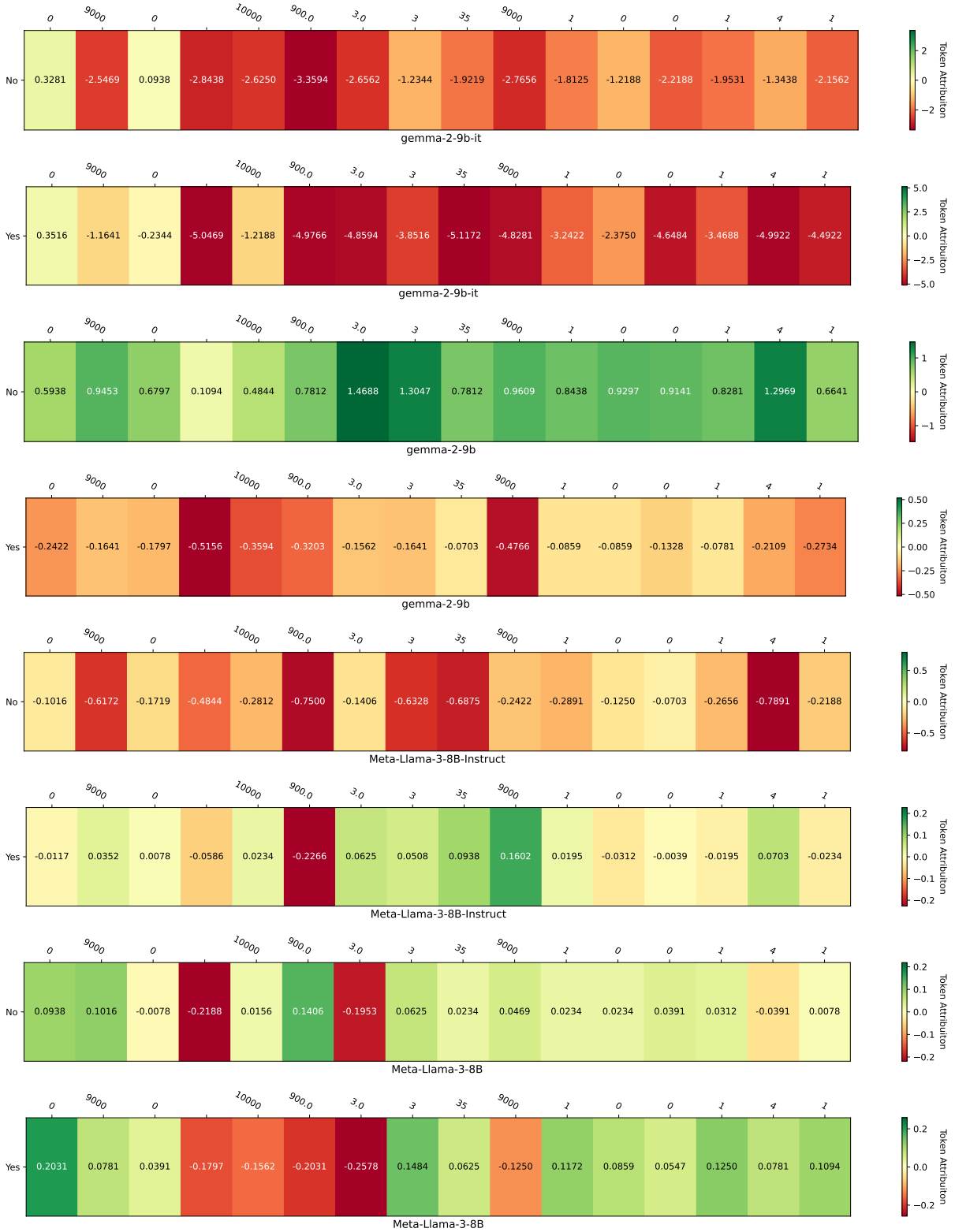

Figure 11: Attribution scores of Ghana data for example 2. Positive attribution scores are indicated in green, while negative scores are shown in red. `Gemma-2-9b-it` models show more negative and neutral token attribution.

```
{'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 12: 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 don't see a focus on specific features 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 13: 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 don't see a focus on specific features 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'}
```

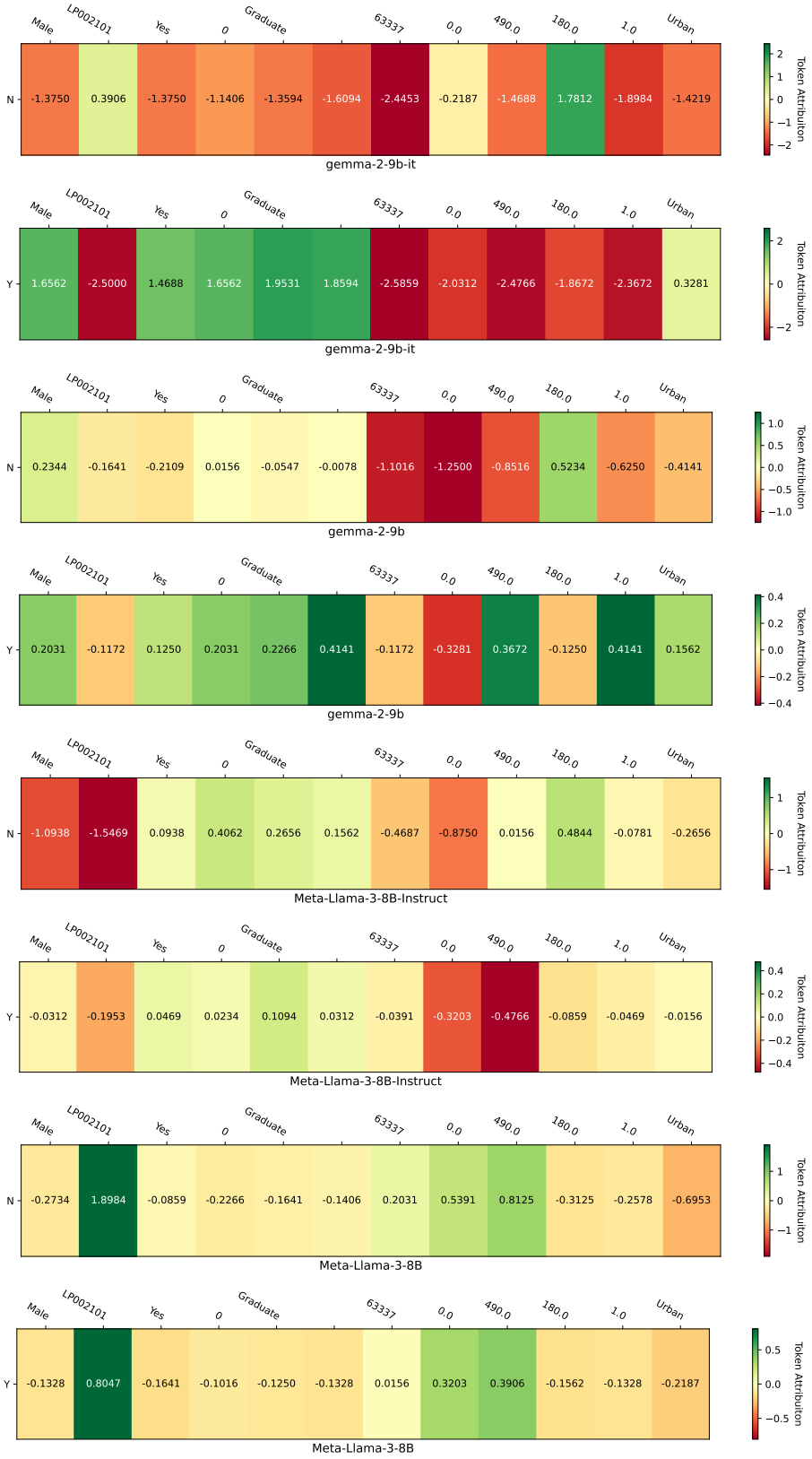

Figure 14: This figure displays the attribution scores for Example 1 of the US dataset. Positive attribution scores are indicated in green, while negative scores are shown in red. We can see the "Loan_ID" feature significantly influences the model's output.

```
{'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'}
```

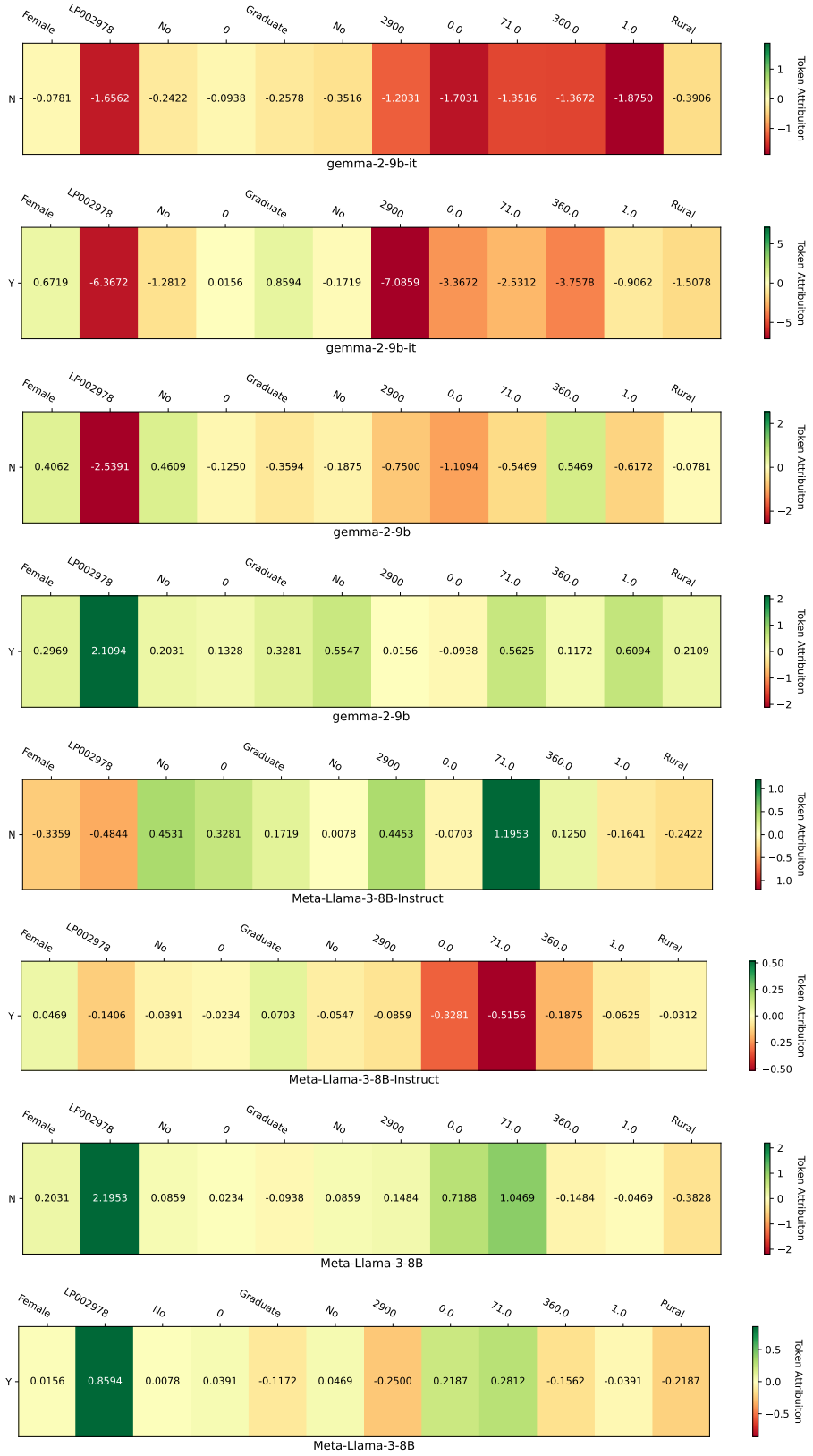

Figure 15: This figure displays the attribution scores for Example 2 of the US dataset. Positive attribution scores are indicated in green, while negative scores are shown in red. We can see the "Loan_ID" feature significantly influences the model's output.