In-Context Bias Propagation in LLM-Based Tabular Data Generation

Pol G.Recasens [1] **Alberto Gutierrez-Torre** [1] **Jordi Torres** [1 2] **Josep Ll. Berral** [2 1] **Anisa Halimi** [3] **Kieran Fraser** [3]

Abstract

Large Language Models (LLMs) are increasingly used for synthetic tabular data generation through in-context learning (ICL), offering a practical solution for data augmentation in data scarce scenarios. While prior work has shown the potential of LLMs to improve downstream task performance through augmenting underrepresented groups, these benefits often assume access to a subset of unbiased in-context examples, representative of the real dataset. In real-world settings, however, data is frequently noisy and demographically skewed. In this paper, we systematically study how statistical biases within in-context examples propagate to the distribution of synthetic tabular data, showing that even mild in-context biases lead to global statistical distortions. We further introduce an adversarial scenario where a malicious contributor can inject bias into the synthetic dataset via a subset of in-context examples, ultimately compromising the fairness of downstream classifiers for a targeted and protected subgroup. Our findings demonstrate a new vulnerability associated with LLM-based data generation pipelines that rely on in-context prompts within sensitive domains.

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable generalization capabilities through in-context learning (ICL), performing new tasks simply by conditioning on a number of prompt examples. These strong capabilities are often assumed to stem from a large number of model parameters and the pretraining on large amounts of data (Moreno-Muñoz et al., 2023). In practical domains such as healthcare or finance, however, data collection remains costly and challenging due to privacy concerns and the reliance on manual labeling. As a result, data augmentation has become a critical strategy for building high-quality datasets that drive more accurate ML models (Borisov et al., 2022). In such settings, LLMs are increasingly used for data generation, leveraging their strong prior knowledge to produce realistic and statistically coherent samples without requiring model fine-tuning.

Recent studies have shown that LLM-generated synthetic tabular data can improve downstream model performance, offering a practical alternative to traditional augmentation methods (Kim et al., 2024; Seedat et al., 2023). These methods typically rely on few-shot prompting, where a small number of in-context examples added to the prompt guide the generation. When combined with the prior knowledge of pretrained LLMs, this technique has been shown to effectively augment underrepresented groups, leading to state-of-the-art performance on downstream tasks. Notably, a larger number of in-context examples further improves the quality of the synthesized data (Seedat et al., 2023). However, this paradigm implicitly assumes that the prompt examples are unbiased, and independently drawn from the real training data distribution, which is a strong assumption that often fails in real-world settings.

In practice, available examples are often noisy and skewed. Consider a community hospital attempting to expand its training dataset using an LLM, with available records being demographically imbalanced, over-representing a particular racial or gender group. The statistical imbalances in the prompt can inadvertently bias the synthetic data towards the same underrepresented distribution without unduly influencing quality, potentially amplifying disparities in the downstream task. Moreover, in collaborative workflows, multiple agents may contribute ICL examples to a shared LLM service, each introducing subtle or even adversarial biases. Despite the growing use of LLMs for tabular data generation, how such in-context prompt skew affects the resulting synthetic distribution remains unexplored, as well as whether these biases persist in downstream models trained on the generated data.

Contributions. In this paper, we study how in-context examples influence the statistical properties of LLM-generated tabular data. Our contributions are:

[1]Barcelona Supercomputing Center, Barcelona, Catalonia, Spain [2]Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain [3]IBM Research Ireland. Correspondence to: Pol G.Recasens <pol.garcia@bsc.es>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada.

- We empirically demonstrate that statistical biases present in ICL examples systematically propagate to the distribution of LLM-generated tabular data.
- We introduce a novel perspective on prompt injection, an adversarial setting where a malicious agent *injects* a small number of biased in-context examples into a collaborative data generation pipeline. We show that this targeted manipulation alters the synthetic distribution and induces fairness violations in downstream models.

2. Related Work

Tabular data generation with LLMs. Pretrained language models have recently been used to synthesize tabular data, leveraging their broad priors to augment datasets in domains with limited or sensitive data. While fine-tuning based approaches such as GReaT (Borisov et al., 2022) have proven effective for structured data generation, prompting-based methods offer a resource efficient alternative and have shown promise in augmenting underrepresented groups. For example, Seedat et al. (2023) use ICL for tabular data augmentation, followed by manual curation to refine the synthetic dataset. Kim et al. (2024) introduce a prompting method that relies solely on in-context examples and categorical mappings to generate realistic tabular data. LLMs have also been deployed in real-world healthcare scenarios (Tornqvist et al., 2024), successfully addressing data scarcity in sensitive domains. However, these methods often assume that in-context examples are independent and identically distributed (iid) from the training dataset, which is a strong assumption that rarely holds in practice. In many cases, the available data is skewed or demographically imbalanced. How such biases propagate through LLM generation remains poorly understood.

Fairness in language models. LLMs are known to inherit and amplify biases from their training data and prompts. Prior work has identified various in-context vulnerabilities, including recency, majority-label, and token frequency biases. Wei et al. (2023) show that larger models tend to override semantic priors when exposed to biased in-context examples, whereas smaller models are more robust. To counteract such effects, Ma et al. (2023) introduce fairness-guided prompt search, and Zhou et al. (2024) locate biased attention heads and feed-forward layers, mitigating biases through targeted parameter modification. In structured-data settings, Liu et al. (2023) flip in-prompt labels to address offset demographic bias, and Cherepanova et al. (2024) combine curated exemplars with masking to improve group fairness in tabular classification. While recent work addresses LLM fairness in tabular classification tasks, our work shifts the focus to data generation, studying how prompt-level bias propagates into the synthetic data distribution.

3. Experimental Set-up

Models. We evaluate 4 open-source LLMs with sizes ranging from 8 billion to 70 billion parameters. Specifically, we consider Granite-8b, Mixtral-8b, Mixtral-22b, and Llama-3-3-70b to generate synthetic tabular data. Each LLM is served using vLLM on an internal cluster of A100 GPUs. We generate 1000 rows for each experimental setting via multiple independent API calls, where each call yields only two synthetic samples to preserve statistical robustness across the generation. We select a number of models from different families and sizes to better assess in-context bias propagation, although they are not equally performant at generating tabular data due to different pretraining and alignment with human feedback.

Prompts. We adopt the prompt structure from CLLM (Liu et al., 2023), which decomposes the prompt into *role*, *task*, and *in-context examples* components. However, we modify this structure to explicitly instruct the model to generate one synthetic example per subgroup of the biased attribute. This ensures a balanced frequency count across subgroups of the synthetic dataset. We provide the prompts in Appendix C.

Datasets. We use the ADULT (Asuncion & Newman, 2007) and COMPAS (Angwin et al., 2016) datasets, widely used in fairness research, to study in-context bias propagation. COMPAS contains recidivism risk assessments, while ADULT includes census income data.

Evaluation Metrics. We evaluate the *quality* of the synthetic datasets by training a downstream Random Forest classifier and reporting both accuracy and F1 score on the target prediction task. To assess *fidelity*, we compute the Total Variation Complement (TVC) and the Jensen–Shannon Divergence, applied to categorical and numerical variables, respectively, to quantify distributional similarity between the real and synthetic datasets. *Fairness* is assessed using the Statistical Parity Difference (SPD) and the Disparate Impact (DI) metrics. We detail *fairness* and *fidelity* metrics in Appendix B.

4. Bias Propagation in LLM-generated Data

4.1. In-context Bias Propagation

We formalize how statistical biases in in-context prompts propagate to the synthetic tabular data generated by language models. Let $\mathcal{M}$ be a pretrained language model that generates synthetic structured data via in-context learning (ICL). A prompt of k demonstrations is defined as

$$\mathcal{P} = \{x_1, \ldots, x_k\}, \quad x_i = (a_i, f_i, t_i) \sim \mathcal{D}_P, \qquad (1)$$

where $a_i \in A$ is a protected attribute, $f_i \in \mathcal{F}$ non-sensitive attributes, and $t_i \in T$ the target attribute subject to bias.

Conditioning $\mathcal{M}$ on $\mathcal{P}$ induces a generated synthetic data distribution $\mathcal{D}_G = \mathcal{G}_{\mathcal{M}}(\mathcal{P})$, such that each generated sample $\mathbf{x} = (a, f, t) \sim \mathcal{D}_G$. Our goal is to quantify how biases, from subtle imbalances to deliberate attacks, in the prompt distribution $\mathcal{D}_P$ propagate through $\mathcal{M}$ to $\mathcal{D}_G$.

We focus on the case where the prompt $\mathcal{P}$ is skewed with respect to a subgroup $a* \in A$ and value $c \in T$. We define the bias parameter π as

$$\pi = P_P(t = c | a = a*). \tag{2}$$

which quantifies the conditional probability of label c given a subgroup $a*$ in the prompt distribution $\mathcal{D}_P$. By varying π, we analyze how different levels of prompt bias influence the synthetic data distribution $\mathcal{D}_G$.

4.2. Experimental Design

We consider two symmetric experimental configurations in which in-context examples are *synthetically* generated within the Adult dataset context. Within each configuration we vary the proportion $\pi \in [0, 1]$ of African-American (AA) instances in the prompt while keeping the gender fixed. The remaining features are randomly sampled within realistic value ranges. This set-up allows us to isolate the effect of racial bias and examine whether it it propagates across gender subgroups. Specifically, we consider the setting where $A = \{\text{Male}, \text{Female}\}$ is the protected attribute (gender), $T = \{\text{African-American}, \text{Non-African-American}\}$ is the target attribute (race), and $a* \in A$ denotes the gender subgroup present in the prompt. While race is treated as the protected attribute in all threat-model experiments in Section 5, here we focus on gender to demonstrate how bias propagates across different demographic groups.

4.3. Experimental Results

We analyze how biases embedded in in-context examples influence the LLM-generated distribution of synthetic tabular data. Figure 1 presents results using *Granite-8b* across different levels of racial imbalance. As the probability that in-context samples are African-American increases, the generated data exhibits a corresponding shift in the racial subgroup representation. Interestingly, although the synthetic dataset is forced to be demographically balanced across genders (e.g., we generate one male and one female example per iteration), we observe that in-context biases from one gender group are reflected in the other. This behavior suggests that smaller LLMs capture and propagate in-context univariate statistical patterns, generalizing these correlations across subgroups. Moreover, we observe that generated male examples are less resilient than female examples to in-context bias, showing a disparity on which gender can be more easily biased, which in turn suggests SDG is also vulnerable to robustness biases (Nanda et al., 2021).

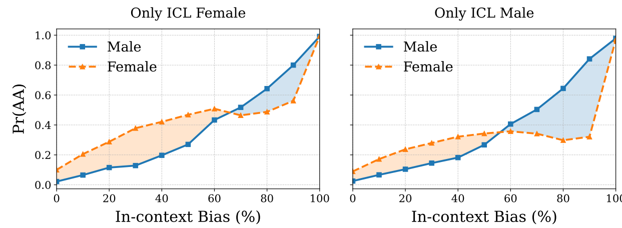

Figure 1. **In-context bias propagates to synthetic samples.** (*Granite-8B*) Increasing the bias parameter π, which controls the proportion of African-American (AA) individuals in the in-context prompt, leads to a corresponding increase in the proportion of AA individuals in the generated samples.

Overall, we observe a general sensitivity of LLMs to the statistical properties of their in-context examples. In the remainder of this section, we systematically characterize this propagation effect across different dimensions: prompt composition, number of in-context examples, and model scale. Together, our findings reveal that even mild biases in the prompt distribution can systematically shape the statistical parity and subgroup label rates of synthetic data, posing serious risks for fairness and security in collaborative generation scenarios.

Number of in-context examples. Figure 2 shows the effect of varying the number of in-context examples on bias propagation for Granite-8B and the ADULT dataset. When using a small context size $k = 10$, the conditional probability $\mathcal{P}_G(t = \text{AA}|a = \text{Female})$ in the generated data remains largely stable across π, showing minimal sensitivity to increasing bias in the provided examples. In contrast, with a larger context size ($k = 80$), this probability increases nearly proportionally with the probability $\mathcal{P}_P(t = \text{AA}|a = \text{Female})$ in the in-context examples, and propagates symmetrically across gender subgroups. While prior work highlighted the benefits of increasing the context size for improving data quality (Seedat et al., 2023), our findings reveal how using a larger number of biased in-context examples can significantly amplify the propagation of in-context statistical disparities to the generated outputs.

Model-specific behavior. Figure 3 compares bias propagation across model families. All models exhibit some degree of prompt-induced bias, although they vary in how they internalize and reproduce feature correlations present in the prompt. For instance, Granite-8b shows strong cross-subgroup bias propagation, with racial bias female examples influencing male examples, while bias propagation in Llama-3-3-70b primarily influences the gender subgroup explicitly present in the prompt. Also, we notice an increase in the racial bias gap between the two gender subgroups between Mixtral-8x7b and Mixtral-8x22b, which suggests that larger

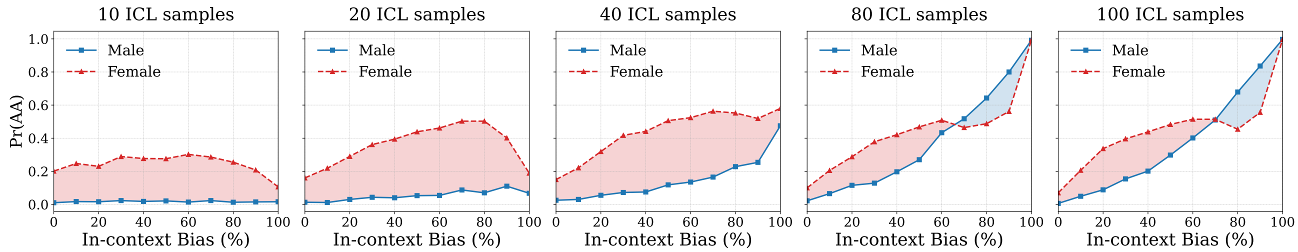

Figure 2. **Larger context size amplifies bias propagation.** (*ICL Female, Granite-8b*) Conditional probability of generating African-American (AA) examples for each gender subgroup across π and increasing context sizes. For a larger context size, the model more strongly internalizes in-context racial imbalances and propagates them to the generated data.

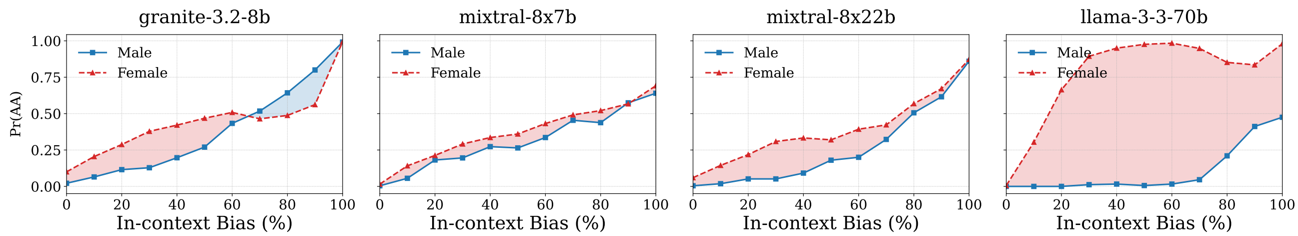

Figure 3. **Model-dependent sensitivity to prompt-induced bias.** (*ICL Female, 80 ICL samples*) Conditional probability of generating African-American (AA) examples for each gender subgroup across π for Granite-8B, Mixtral 8x7b, Mixtral-22B, and LLaMA-3-70B. While all models are vulnerable to prompt-induced bias, they vary in their degree of bias propagation and cross-gender generalization.

models within the same family are able to capture complex correlations within in-context examples. We leave a deeper study of this phenomenon as future work.

5. Adversarial In-Context Bias Injection

In the previous section, we treat prompt skew as an *unintended* consequence of in-context learning for synthetic tabular data generation. Here, we adopt a security perspective and ask the following question:

> *Can a malicious user craft a small number of in-context demonstrations that deliberately manipulate the statistical properties of the synthetic data for a targeted subgroup?*

5.1. Threat Model

We consider a collaborative data generation pipeline where multiple users or agents individually contribute few-shot in-context examples to a shared LLM service. These individual contributions are combined into a single prompt, which is then used to generate synthetic data. An adversary controls a fraction π of the k demonstrations in the final prompt, but has no access to the model parameters or the in-context examples provided by honest users. However, we assume that the attacker knows the downstream task and the names of the most influential attributes in the downstream model. The adversary's objective is to influence the label rate for a protected group in the downstream model (e.g., the proportion of high-income labels assigned to Black individuals) while maintaining high overall utility to avoid detection.

Feature-Aligned Examples. We consider an adversary that creates *feature-aligned* examples by embedding correlations between the protected group and features predictive of the downstream task. Instead of only manipulating labels, which proves ineffective in scenarios where the protected attribute is ignored by the downstream model, the adversary creates a protected-group profile statistically indicative of the target label, and biases generation toward producing protected-group examples with patterns likely to be learned by the downstream model.

5.2. Experimental Design

For each value of π we generate 1000 synthetic examples with Granite-8B, Mixtral-22B, and Llama-70B. Specifically, we consider $k = 80$ in-context examples from which we randomly bias πk examples, with the rest of the demonstrations corresponding to real examples i.i.d from the real dataset. This set-up simulates a realistic collaborative scenario in which multiple parties contribute in-context examples, while an attacker injects πk malicious demonstrations.

In the ADULT dataset, we define high-income indicators as a proxy for the positive label (i.e., income $\geq$ 50K). Specifically, we consider samples with a college education, non-

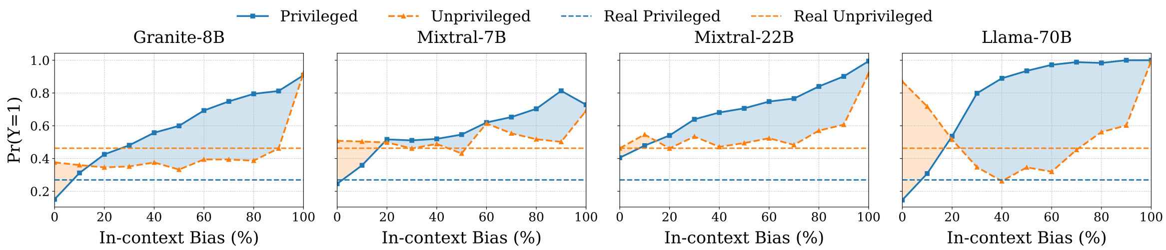

Figure 4. **Adversarial bias in synthetic data.** (*Adult, 80 ICL samples*) Positive label probability of the synthetic data for each racial subgroup, across different values of π. As π increases, the targeted group's positive-label probability also increases, while the complementary group's probability remains constant except for larger values of π.

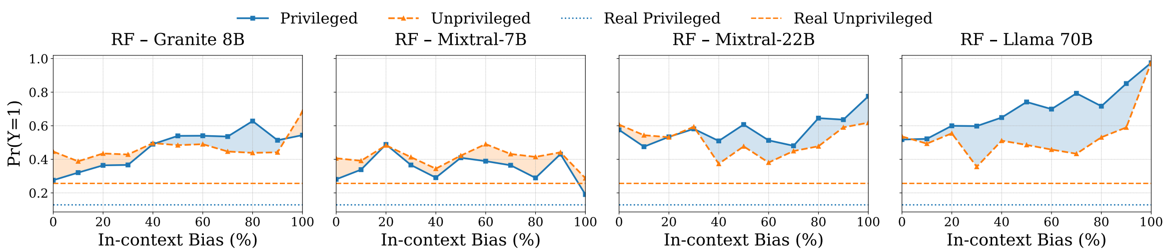

Figure 5. **Adversarial bias in downstream predictions.** (*Adult, 80 ICL samples*) Positive label probability of the downstream model for each racial subgroup across different values of π. The targeted racial group increasingly receives more positive predictions compared to the complementary group, reflecting the propagation of bias to the downstream model.

zero capital gain, and at least 45 hours of work per week as indicative of high income. The attacker constructs biased prompt examples by enforcing these conditions, setting the race attribute to Black, and assigning the target label to 1. The goal is to manipulate the data generation process so that the synthetic distribution over-represents Black individuals with high income without inadvertently decreasing the quality of the generated synthetic data.

In the COMPAS dataset, the target feature for prediction is recidivism i.e., a binary variable indicating if a convicted criminal will re-offend. The influential indicators for recidivism are defined as: individuals who are young; have a large number of prior offenses; and face a felony charge. In this scenario, the attacker constructs examples with African-American race that satisfy these conditions and assigns a target label of 1, indicating recidivism.

5.3. Experimental Results

Fairness-impact. Figure 4 illustrates the probability that synthetic samples receive a positive target label across varying in-context prompt bias rates π for the ADULT dataset. For each model evaluated we observe that increasing the proportion of biased in-context examples leads to a rise in the positive label probability of Black group synthetic examples. However, the positive label probability for the non-Black group remains constant and low except for larger π values. Moreover, we can observe that when real samples are used as in-context examples, the probability of assigning a positive label to the Black group remains very low, indicating a strong prior bias in the underlying model toward assigning $y = 0$ (low income) to this subgroup. This prior bias may interfere with or mask the propagation of injected prompt-level biases, a phenomenon we also examine in Appendix B using the COMPAS dataset.

Figure 5 further shows that in-context bias in the prompt propagates into downstream models trained on the synthetic data. In particular, the downstream classifier exhibits a growing disparity in positive label predictions between racial subgroups as π increases, with the Black group consistently receiving higher positive label rates. This trend, observed across all tested model families, reveals a systemic vulnerability in LLM-based data generation pipelines, in-context prompt biases not only influence the synthetic data distribution but also propagates to downstream models.

Utility trade-off. Finally, we investigate the relation between in-context bias and data utility in Table 1. For the ADULT dataset, we observe a substantial drop in downstream performance across models when in-context examples are fully biased. However, a moderate in-context bias rate (e.g., 40%) preserves utility and fidelity of the synthetic data while significantly affecting fairness for a specific subgroup. This suggests that only a small fraction of biased examples is sufficient for an attacker to influence subgroup-

Table 1. Quality, fidelity and fairness metrics for different levels of in-context bias (π). For both datasets, a moderate number of biased in-context samples ($\pi = 0.4$) maintains the utility (Acc, F1) and the fidelity (TVC, JSD) to the real distribution, while simultaneously introducing a fairness violation (e.g., $|\Delta\text{SP}| > 0.1$).

Data	Generator	π	Acc↑	F1↑	SPD↓	DI↓	TVC↑	JSD↓
Adult	Real oracle	–	0.857 ± 0.001	0.852 ± 0.001	0.131 ± 0.006	2.084 ± 0.088	–	–
	Granite - 8B	0.0	0.750 ± 0.000	0.766 ± 0.000	0.226 ± 0.064	2.756 ± 1.176	0.843 ± 0.009	0.205 ± 0.027
	Granite - 8B	0.4	0.631 ± 0.001	0.658 ± 0.001	-0.182 ± 0.084	0.682 ± 0.134	0.794 ± 0.014	0.304 ± 0.015
	Granite - 8B	1.0	0.290 ± 0.000	0.202 ± 0.001	0.012 ± 0.036	1.013 ± 0.040	0.695 ± 0.007	0.397 ± 0.009
	Mixtral - 22B	0.0	0.491 ± 0.001	0.520 ± 0.001	0.058 ± 0.073	1.157 ± 0.187	0.747 ± 0.011	0.303 ± 0.011
	Mixtral - 22B	0.4	0.437 ± 0.003	0.465 ± 0.004	-0.209 ± 0.061	0.694 ± 0.074	0.720 ± 0.010	0.394 ± 0.024
	Mixtral - 22B	1.0	0.234 ± 0.000	0.089 ± 0.000	-0.080 ± 0.027	0.920 ± 0.027	0.662 ± 0.010	0.734 ± 0.019
	Llama - 70B	0.0	0.491 ± 0.002	0.520 ± 0.002	0.723 ± 0.021	6.024 ± 0.819	0.696 ± 0.008	0.766 ± 0.011
	Llama - 70B	0.4	0.437 ± 0.003	0.465 ± 0.004	-0.628 ± 0.054	0.294 ± 0.052	0.677 ± 0.007	0.754 ± 0.014
	Llama - 70B	1.0	0.234 ± 0.000	0.089 ± 0.000	-0.005 ± 0.005	0.995 ± 0.005	0.669 ± 0.008	0.717 ± 0.014
COMPAS	Real oracle	–	0.621 ± 0.000	0.620 ± 0.000	0.132 ± 0.030	0.749 ± 0.049	–	–
	Granite - 8B	0.0	0.559 ± 0.000	0.556 ± 0.000	0.194 ± 0.066	0.660 ± 0.010	0.819 ± 0.013	0.396 ± 0.011
	Granite - 8B	0.4	0.557 ± 0.000	0.557 ± 0.000	0.488 ± 0.069	0.304 ± 0.048	0.872 ± 0.008	0.464 ± 0.010
	Granite - 8B	1.0	0.542 ± 0.000	0.202 ± 0.001	0.816 ± 0.039	0.168 ± 0.031	0.804 ± 0.011	0.496 ± 0.002
	Mixtral - 22B	0.0	0.598 ± 0.000	0.598 ± 0.000	0.506 ± 0.094	0.364 ± 0.082	0.863 ± 0.011	0.400 ± 0.014
	Mixtral - 22B	0.4	0.541 ± 0.000	0.526 ± 0.000	0.819 ± 0.045	0.118 ± 0.044	0.849 ± 0.013	0.448 ± 0.010
	Mixtral - 22B	1.0	0.514 ± 0.000	0.473 ± 0.000	0.020 ± 0.012	0.980 ± 0.012	0.686 ± 0.012	0.525 ± 0.004
	Llama - 70B	0.0	0.580 ± 0.001	0.561 ± 0.000	0.692 ± 0.031	0.156 ± 0.038	0.860 ± 0.010	0.427 ± 0.002
	Llama - 70B	0.4	0.543 ± 0.000	0.538 ± 0.000	0.996 ± 0.009	0.004 ± 0.009	0.801 ± 0.008	0.493 ± 0.013
	Llama - 70B	1.0	0.534 ± 0.000	0.525 ± 0.000	0.020 ± 0.019	0.980 ± 0.019	0.737 ± 0.012	0.557 ± 0.004

specific outcomes, all while staying within acceptable fidelity and utility ranges. Following Baracaldo et al. (2022), we consider a successful attack when SPD is greater than 0.1 compared to the SPD under $\pi = 0$. Under this threshold, we observe that for a $\pi = 0.4$ bias rate, the SPD fluctuates more than 0.2 in all tested models and datasets. Despite this fairness degradation, downstream accuracy and F1 scores remain deceptively stable for intermediate values of π. This trade-off reveals an attack surface in which fairness attacks may go undetected if only utility metrics are used to assess synthetic data pipelines.

6. Discussion

Recent work demonstrated that language models exhibit strong few-shot data generation capabilities without requiring further fine-tuning (Kim et al., 2024; Seedat et al., 2023). This black-box approach effectively improves downstream performance while avoiding the computational cost of model training. While prior work mostly relies on recent GPT models (Achiam et al., 2023), we explore the propagation of in-context statistical patterns to synthetic data on three families of user-accessible language models. Our goal is not to compare model quality, given the different pretraining and model alignments, but rather to highlight that in-context prompt injection vulnerabilities persist across different model architectures and sizes. We leave as future work a more in-depth analysis of how in-context examples affect recent complex data generation pipelines, as well as how other prompt components, such as prompt phrasing or data types, affect the propagation of bias.

Additionally, our findings suggest that model scale interacts with bias propagation in non-trivial ways. For example, larger models within the same family (e.g., Mixtral-8x22B) appear more sensitive to cross-feature correlations, potentially increasing their vulnerability to subtle but effective adversarial manipulations. Additionally, language models inherit prior biases from their pretraining data, which can interact with in-context biases in complex ways. We also leave for future work a systematic study of how model capacity affects the correlations learned from in-context examples, and how the inherent language model prior biases can be leveraged to design defenses against such vulnerabilities.

7. Conclusion

In this paper, we studied how statistical biases present in in-context examples propagate to synthetic data during LLM-based tabular data generation. We introduced a novel attack surface in which a malicious actor can manipulate the output distribution for a targeted subgroup by injecting a small number of biased examples. Our results show that this manipulation can further propagate to the downstream model, potentially compromising the fairness and integrity of generated data in sensitive domains.

Acknowledgements

This work has been partially financed by grant agreement EU-HORIZON GA.101095717 and by grant agreement EU-HORIZON MSCA GA.101086248. Also, it has been partially financed by Generalitat de Catalunya (AGAUR) under grant agreement 2021-SGR-00478, by Severo Ochoa Center of Excellence CEX-2021-001148-S-20-3, and by the Spanish Ministry of Science (MICINN), the Research State Agency (AEI) and European Regional Development Funds (ERDF/FEDER) under grant agreement PID2021-126248OB-I00, MCIN/AEI/10.13039/501100011033/ FEDER, UE. Kieran Fraser and Anisa Halimi were partly supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No. 101172997 – SEARCH.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Angwin, J., Larson, J., Kirchner, L., and Mattu, S. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016. ProPublica.

Asuncion, A. and Newman, D. Uci machine learning repository. http://archive.ics.uci.edu/ml, 2007. University of California, Irvine, School of Information and Computer Sciences.

Baracaldo, N., Eykholt, K., Ahmed, F., Zhou, Y., Priya, S., Lee, T., Kadhe, S., Tan, Y., Polavaram, S., Suggs, S., et al. Benchmarking the effect of poisoning defenses on the security and bias of the final model. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Cherepanova, V., Lee, C.-J., Akpinar, N.-J., Fogliato, R., Bertran, M. A., Kearns, M., and Zou, J. Improving llm group fairness on tabular data via in-context learning. *arXiv preprint arXiv:2412.04642*, 2024.

Kim, J., Kim, T., and Choo, J. Epic: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models. *Advances in Neural Information Processing Systems*, 37:31504–31542, 2024.

Liu, Y., Gautam, S., Ma, J., and Lakkaraju, H. Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications. *arXiv preprint arXiv:2310.14607*, 2023.

Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., Zhang, S., Fu, H., Hu, Q., and Wu, B. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155, 2023.

Moreno-Muñoz, P., Garcia Recasens, P., and Hauberg, S. On masked pre-training and the marginal likelihood. *Advances in Neural Information Processing Systems*, 36: 79781–79791, 2023.

Nanda, V., Dooley, S., Singla, S., Feizi, S., and Dickerson, J. P. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 466–477, 2021.

Seedat, N., Huynh, N., van Breugel, B., and van der Schaar, M. Curated llm: Synergy of llms and data curation for tabular augmentation in ultra low-data regimes. 2023.

Tornqvist, M., Zucker, J.-D., Fauvel, T., Lambert, N., Berthelot, M., and Movschin, A. A text-to-tabular approach to generate synthetic patient data using llms. *arXiv preprint arXiv:2412.05153*, 2024.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Zhou, H., Feng, Z., Zhu, Z., Qian, J., and Mao, K. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. *arXiv preprint arXiv:2405.20612*, 2024.

Appendix

This supplementary work complements the main manuscript with the following points.

- Appendix A shows the complementary results of the analysis for the COMPAS dataset.
- Appendix B defines the fairness and quality metrics used to evaluate the synthetic datasets.
- Appendix C shows each prompt structure used in the two sections.

A. Additional analysis

A.1. COMPAS dataset

In this section, we present a complementary analysis with the COMPAS dataset. Similar to studied ADULT scenario, the attacker's objective is to influence the synthetic data such that the African-American subgroup receives a higher proportion of positive labels through in-context sample manipulation. We evaluate the impact of manipulating a proportion π of the in-context examples, with the remaining samples drawn from the original training distribution.

Compared to the ADULT case, even small values of π result in a significant shift in the synthetic data distribution. As shown in Figure 6, the positive label probability for the unprivileged group remains nearly constant across most π values, except at very high values of π. In contrast, the positive label probability for the privileged subgroup increases with π. Interestingly, for larger models such as Mixtral-22B and Llama-70B, the privileged group's positive label probability reaches 100% with only minimal manipulation (i.e., very low π).

We argue that, in the case of the COMPAS dataset, the model exhibits a strong prior bias to predict a higher likelihood of recidivism for the Black racial group, which can be observed when $\pi = 0$ in Figure 6. This inherent bias facilitates the drift in positive label probability of the target group, even under very limited in-context bias.

In Figure 7 we observe the positive label probability of the downstream model trained over the synthetic data across different values of π. Compared to the COMPAS dataset, here the strong prior leads to even a larger gap between the positive label probability between the two subgroups.

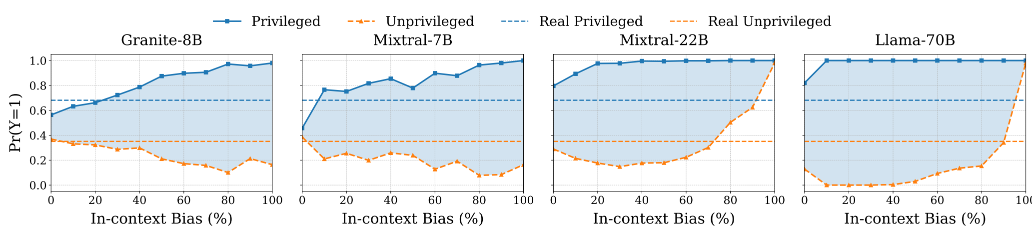

Figure 6. **Adversarial bias in synthetic data.** (*Compas, 80 ICL examples*) Probability of a positive label in the generated samples for each racial subgroup, across different values of π. While smaller models show an increase of the positive label probability for the target group across π, larger models show a sharper bias influence, with minimal manipulation leading to maximum positive label probability.

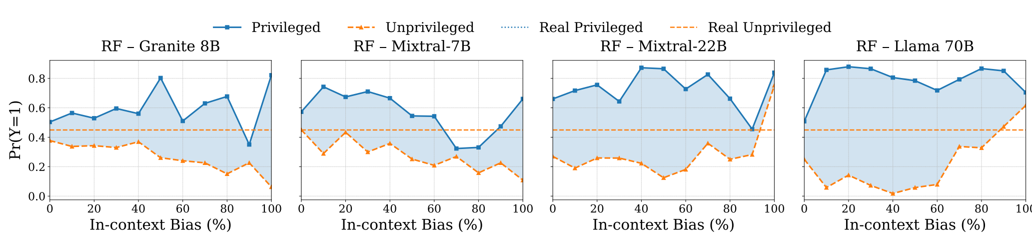

Figure 7. **Adversarial bias in downstream predictions.** (*Compas, 80 ICL examples*) Probability of the downstream model predicting a positive label for each racial subgroup across different values of π. The targeted racial group increasingly receives more positive predictions compared to the complementary group, reflecting the propagation of bias to the downstream model.

B. Evaluation metrics

In this complementary section, we detail the metrics used to evaluate the fairness and fidelity of the synthetic tabular data.

B.1. Fairness metrics

We measure the **statistical parity difference** (SPD) between subgroups of a protected attribute, which represents the difference in the rate of favorable outcomes between those subgroups. In the main manuscript, we consider Gender as the protected attribute and Race $=$ African-American as favorable outcome for Section 4, and Race as the protected attribute and $y = 1$ as the favorable outcome in Section 5. Formally, the statistical parity difference is defined as

$$\mathcal{P}_G(t = 1|a = a_U) - \mathcal{P}_G(t = 1|a = a_P), \tag{3}$$

where $a \in A$ is the protected attribute, a_U denotes the unprivileged subgroup, and a_P to the privileged subgroup. A fair outcome corresponds to a SPD value close to 0. Similarly, we define the **disparate impact** (DI) between subgroup, as the ratio between

$$\frac{\mathcal{P}_G(t = 1|a = a_U)}{\mathcal{P}_G(t = 1|a = a_P)}, \tag{4}$$

where values close to 1 indicate statistical parity between groups. Both metrics are computed over the generated distribution $\mathcal{P}_G$, and are commonly used to assess group fairness in data distributions.

B.2. Fidelity metrics

We evaluate the fidelity of the generated synthetic distribution $\mathcal{P}_G$ respect to the real data distribution $\mathcal{P}_R$. First, for categorical variables with support $\mathcal{V}$, we define the Total Variation Distance as

$$\mathrm{TVD}(\mathcal{P}_R, \mathcal{P}_G) \;=\; \frac{1}{2} \sum_{v \in \mathcal{V}} \big| \mathcal{P}_R(a = v) - \mathcal{P}_G(a = v) \big|. \tag{5}$$

Intuitively, this metric defines the distance between the frequency of each categorical value between the two distributions $\mathcal{P}_G$ and $\mathcal{P}_R$. For evaluating fidelity of categorical attributes we use the **Total Variation Complement**, defined as $1 - \mathrm{TVD}$, with values close to 1 representing high fidelity between the two distributions.

For numerical variables, we use the **Jensen–Shannon divergence** (JSD), which is a smoothed version of the Kullback-Leibler divergence, to quantify overall distributional similarity,

$$\mathrm{JSD}(\mathcal{P}_R \,\|\, \mathcal{P}_G) = \frac{1}{2}\,\mathrm{KL}\big(\mathcal{P}_R \,\big\|\, M\big) + \frac{1}{2}\,\mathrm{KL}\big(\mathcal{P}_G \,\big\|\, M\big), \quad M = \tfrac{1}{2}\big(\mathcal{P}_R + \mathcal{P}_G\big),$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence. Lower JSD indicates higher fidelity.

C. Prompt structure

Balanced Prompt – Adult - Gender

```
System role:
You are a tabular synthetic data generator. You produce realistic yet diverse synthetic
     samples that mirror the causal structure and feature label distributions of the
    provided examples.

User instructions:
Using your knowledge of adult income data, generate exactly two realistic samples, one
    with "gender": "Male" and one with "gender": "Female".
You will first receive a block of real examples as in-context demonstrations.

Example data:
{icl_examples}

Produce your output as a single JSON array containing two objects. Do not include any
    explanatory text, commentary, or extra fields.

Each object must have exactly these keys in this order:

{
    "age": "int",
    "workclass": "string",
    "fnlwgt": "int",
    "education": "string",
    "educational-num": "int",
    "marital-status": "string",
    "occupation": "string",
    "relationship": "string",
    "race": "string",
    "gender": "string",
    "capital-gain": "int",
    "capital-loss": "int",
    "hours-per-week": "int",
    "native-country": "string",
    "income": "string", // "<=50K", ">50K"
}

Output example:
[
  {
    {example1}
  },
  {
    {example2}
  }
]

Strictly return only the JSON array of two objects corresponding to one male and one
    female - no prose, no reordered or extra keys.
```

Balanced Prompt – Adult - Race

```
System role:
You are a tabular synthetic data generator. You produce realistic yet diverse synthetic
     samples that mirror the causal structure and feature label distributions of the
    provided examples.

User instructions:
Using your knowledge of adult income data, generate exactly two realistic samples, one
    with "race": "African-American" and one with "race": "Caucasian".
    You will first receive a block of real examples as in-context demonstrations:

Example data:
{icl_examples}

Produce your output as a single JSON array containing two objects. Do not include any
    explanatory text, commentary, or extra fields.

Each object must have exactly these keys in this order:

{
    "age": "int",
    "workclass": "string",
    "fnlwgt": "int",
    "education": "string",
    "educational-num": "int",
    "marital-status": "string",
    "occupation": "string",
    "relationship": "string",
    "race": "string",
    "gender": "string",
    "capital-gain": "int",
    "capital-loss": "int",
    "hours-per-week": "int",
    "native-country": "string",
    "income": "string", // "<=50K", ">50K"
}

Output example:
[
  {
    {example1}
  },
  {
    {example2}
  }
]

Strictly return only the JSON array of two objects corresponding to one African-
    American and one Caucasian - no prose, no reordered or extra keys.
```