Categorical Prior Lock-in: Why In-Context Learning Fails for Structured Data

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used as conditional generators for structured data, relying on in-context learning (ICL) to adapt to new distributions without parameter updates. We investigate the limits of ICL for structured generation under distribution mismatch, using high-cardinality tabular data as a controlled test case, and identify a structural failure mode we term *categorical prior lock-in*: the inability of ICL to update the model's prior over token distributions inherited from pre-training. Across two 7B-parameter open-weight models, ICL improves numerical fidelity with additional examples but exhibits a sharp ceiling on categorical distributions, failing to reproduce rare classes entirely. Parameter-efficient fine-tuning (LoRA) overcomes these limitations but introduces measurable memorization risk and, in some cases, destabilizes structured output generation, highlighting a fundamental trade-off between adaptability and privacy.

1 Introduction

Large language models (LLMs) have demonstrated a remarkable ability to perform new tasks from a small number of examples provided in context, without updating model parameters. This paradigm, known as in-context learning (ICL), is often interpreted as a form of implicit adaptation, enabling models to approximate new input-output mappings through conditioning alone. As a result, ICL has been widely adopted as a lightweight alternative to fine-tuning, particularly in settings where parameter updates are costly or infeasible.

A growing line of work applies this paradigm to structured generation tasks, including the synthesis of tabular data, where the model is prompted with a schema and example records and expected to generate samples matching the target distribution. While recent studies report promising results, these evaluations primarily focus on downstream utility metrics and often overlook the underlying distributional behaviour of the model, leaving open a fundamental question: *to what extent can ICL adapt a pre-trained language model to a new data distribution in structured generation tasks?*

In this work, we answer that question empirically and identify a structural failure mode we term *categorical prior lock-in*: the tendency of a language model to adhere to its pre-training distribution over tokens, even when conditioned on examples drawn from a different target distribution. Using high-cardinality categorical features as a stress test, we show that ICL cannot approximate rare or domain-specific label distributions regardless of prompt design or the number of in-context examples. Unlike numerical features – whose fidelity improves consistently as more examples are provided – categorical distributions exhibit a sharp ceiling, and minority categories are never reproduced. This limitation propagates to inter-feature dependencies and minority class omission, and is not resolved by prompt engineering within the tested range. We analyse this phenomenon and its mechanisms in depth in Section 6.

In contrast, parameter-efficient fine-tuning (LoRA) substantially improves both marginal and joint fidelity by performing the global distributional adjustment that ICL cannot. This improvement, however, comes at a cost: all fine-tuned configurations exhibit proximity to training records below the memorization threshold, raising concerns about data leakage and privacy. Our contributions are as follows: (1) we identify and characterise categorical prior lock-in as a fundamental failure mode of ICL in structured generation; (2) we provide a systematic empirical evaluation across models and prompting strategies; and (3) we analyse the trade-offs between ICL and parameter-efficient fine-tuning in terms of fidelity, structural consistency, and privacy risk.

2 Related Work

The dominant pre-LLM approach relies on Generative Adversarial Networks. CTGAN (Xu et al., 2019) addresses mixed column types and class imbalance through a conditional generator and mode-specific normalization, establishing a widely-used baseline. GAN-based methods, however, require explicit architectural decisions per column type and struggle with high-cardinality categoricals, a limitation LLM-based methods were expected to overcome through pre-trained semantic knowledge.

Brown et al. (Brown et al., 2020) established that LLMs can perform novel tasks from a handful of demonstrations without weight updates, a capacity extended to tabular synthesis by subsequent work. CLLM (Seedat et al., 2024) showed that LLMs prompted with in-context examples, paired with a post-hoc curation mechanism, can augment tabular datasets in low-data regimes without fine-tuning. EPIC (Kim et al., 2025) systematically characterized prompt design elements for ICL-based generation, showing that balanced, grouped examples mitigate class imbalance. Both works, however, focus on frontier models and report ML utility metrics only, leaving open questions about distributional fidelity on high-cardinality categoricals and privacy risk under local deployment.

GReaT (Borisov et al., 2023) first demonstrated that fine-tuning a pre-trained autoregressive model on serialized tabular records produces realistic synthetic data. REaLTabFormer (Solatorio and Dupriez, 2023) extended this to relational datasets, and TabuLa (Zhao et al., 2025) explored efficiency improvements for the same paradigm. None of these works, however, targets organizations under data residency constraints, for whom 7B instruction-tuned models with parameter-efficient adaptation represent the upper limit of on-premise deployment.

3 Experimental Setup

Dataset The dataset is a publicly available credit card transaction log with binary fraud labels. Each record contains seven numerical features,transaction amount (`amt`), cardholder and merchant geographic coordinates (`lat`, `long`, `merch_lat`, `merch_long`), city population (`city_pop`), and Unix timestamp (`unix_time`), and four categorical features: transaction category (`category`), gender (`gender`), occupation (`job`), and US state (`state`). The `job` field contains 494 distinct labels with a highly imbalanced distribution (most common label: 0.75% of records). Geographic coordinates are constrained to continental US territory with strong inter-feature correlations ($r > 0.99$ between cardholder and merchant coordinates). `amt` and `city_pop` follow heavy-tailed distributions (skewness $= 42.3$, kurtosis $= 4545.6$; median $= 2456$ vs. mean $= 88824$). Only 0.58% of transactions are fraudulent, a ratio critical for downstream classifier calibration.

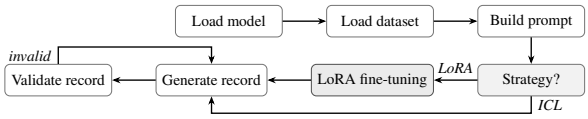


Figure 1: Generation pipeline.

Models **Qwen2.5-7B-Instruct** (Qwen et al., 2025) and **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023) are open-weight, instruction-tuned decoder-only transformers selected for local deployment under data residency constraints. Their differences in training data and methodology allow us to assess whether observed behaviors are model-specific or general to this scale.

Generation Strategies All configurations use a schema-encoding base prompt across three paradigms: **Zero-shot ICL**; **Few-shot ICL** (1, 5, or 10 records); and **LoRA fine-tuning** (Hu et al., 2021) (trained on 10% or 50% of real data, then prompted zero-shot).

Generation Pipeline and Validation As illustrated in Figure 1, all configurations follow a unified pipeline[1].

Records are generated one at a time via temperature sampling ($T = 0.8$, top-$p = 0.95$, repetition penalty $= 1.05$) with a fixed random seed. A record is valid if it parses as well-formed JSON, contains all required fields, and conforms to the expected type signature; invalid records are discarded and regenerated until 10000 valid records are collected.

4 Evaluation Framework

Structural Reliability Structural reliability is measured by the error rate: the fraction of generation attempts that fail the validation pipeline, a key indicator of deployment efficiency as high failure rates impose significant computational overhead.

[1]Code will be made available upon acceptance.

4.1 Distributional Fidelity

Statistical fidelity is assessed using Total Variation Distance (TVD) computed independently for each feature between the synthetic and real marginal distributions:

$$\mathrm{TVD}(P, Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \quad (1)$$

TVD is bounded in $[0, 1]$, where 0 indicates identical distributions and 1 completely disjoint support. Continuous features are discretized into 50 equal-width bins before computation. Note that TVD captures only marginal distributions: a synthetic dataset can match real marginals while failing to preserve inter-feature correlations, making TVD a necessary but not sufficient condition for distributional fidelity.

Privacy and Memorization Risk We quantify memorization risk via the DCR Ratio, which compares the average distance from each synthetic record to its nearest real neighbor against the average nearest-neighbor distance within the real dataset:

$$\text{DCR Ratio} = \frac{\mathbb{E}_{s \sim S}[\min_{r \in R} d(s, r)]}{\mathbb{E}_{r \sim R}\left[\min_{r' \in R \setminus \{r\}} d(r, r')\right]} \quad (2)$$

where S and R are the synthetic and real record sets, and $d(\cdot, \cdot)$ is the Euclidean distance over the six numerical features standardized via z-score normalization; categorical features and `unix_time` are excluded. A ratio above 1.0 indica tes synthetic records are on average more distant from real records than real records are from each other; below 1.0 suggests the model may be reproducing training records too closely. As a mean-based statistic, the DCR Ratio can mask tail behavior; it should therefore be interpreted alongside the exact duplication rate—the fraction of synthetic records whose distance to the nearest real record falls below 0.01 in standardized feature space.

Fraud Class Reproduction Fraud class reproduction is assessed by comparing the prevalence of fraudulent transactions (`is_fraud` $= 1$) in the synthetic corpus against the real base rate of 0.58%. Complete absence of fraudulent records is treated as a categorical failure, as any classifier trained on such data will lack exposure to the minority class entirely.

Inter-Feature Correlation Preservation We assess inter-feature dependency preservation by computing the Pearson correlation matrix over the seven numerical features and the binary fraud label for both the real and synthetic datasets. Two summary statistics are reported: the maximum absolute pairwise correlation error and the total absolute correlation error over all off-diagonal entries, with lower values indicating better preservation of the dependency structure.

5 Results

5.1 Structural Reliability

Table 1: Generation Reliability

Model	Strategy	Valid	Errors	Error Rate (%)	Total Attempts
Qwen2.5-7B	ICL Zeroshot	10,000	118	1.17%	10,118
	ICL Fewshot 1r	10,000	3	0.03%	10,003
	ICL Fewshot 5r	10,000	0	0.00%	10,000
	ICL Fewshot 10r	10,000	1	0.01%	10,001
	LoRA FT 10%	10,000	0	0.00%	10,000
	LoRA FT 50%	10,000	0	0.00%	10,000
Mistral-7B	ICL Zeroshot	10,000	372	3.59%	10,372
	ICL Fewshot 1r	10,000	139	1.37%	10,139
	ICL Fewshot 5r	10,000	90	0.89%	10,090
	ICL Fewshot 10r	10,000	128	1.26%	10,128
	LoRA FT 10%	10,000	6,962	41.04%	16,962
	LoRA FT 50%	0	1,000	100.00%	1,000

Qwen2.5-7B behaves consistently across all tested configurations. The 1.17% zero-shot ICL error rate—already low enough for most practical purposes—drops to near-zero with a single in-context example and reaches exactly zero at five. Fine-tuning introduces no degradation, with both LoRA configurations at 0.00%.

Mistral-7B's behavior under ICL is acceptable, if somewhat noisier: error rates fluctuate between 0.89% and 3.59% across the few-shot series. Fine-tuning tells a very different story. At 10% data exposure, the error rate climbs to 41.04%—over ten times the zero-shot ICL baseline. At 50% exposure, generation fails completely, producing no valid records in 1000 attempts. The LoRA 50% configuration is excluded from all subsequent analyses.

The high rejection rates, especially for Mistral-7B, trace back to a range of parsing and structural anomalies we identified during validation. For Qwen2.5-7B, these rare failures are limited to minor syntactic inconsistencies in zero-shot settings, such as the injection of code snippets or arithmetic operators (e.g., the + symbol) within JSON fields, the representation of `cc_num` as a string, and slight schema deviations like duplicating the `zip` field or mislabeling `Unnamed: 0` as `Unnamed`.

Mistral-7B exhibits more severe issues: infinite

Table 2: Per-Feature Numerical TVD

Model	Strategy	amt	city_pop	lat	long	merch_lat	merch_long	unix_time
Qwen2.5-7B	ICL Zeroshot	0.0506	0.4144	0.5183	0.6499	0.4565	0.6415	0.6931
	ICL Fewshot 1r	0.0021	0.1747	0.2642	0.3730	0.2733	0.4048	0.5884
	ICL Fewshot 5r	0.0127	0.0417	0.2179	0.2990	0.2175	0.3162	0.4829
	ICL Fewshot 10r	0.1004	0.0453	0.2196	0.2937	0.1785	0.2937	0.3745
	LoRA FT 10%	0.0073	0.0175	0.0432	0.0429	0.0403	0.0403	0.0767
	LoRA FT 50%	0.0064	0.0152	0.0309	0.0427	0.0283	0.0320	0.0680
Mistral-7B	ICL Zeroshot	0.0090	0.7146	0.5391	0.6018	0.3823	0.5931	0.0005
	ICL Fewshot 1r	0.0090	0.0623	0.2080	0.2992	0.2160	0.3017	0.0001
	ICL Fewshot 5r	0.0090	0.0692	0.1691	0.1691	0.1628	0.1710	0.1986
	ICL Fewshot 10r	0.0090	0.0563	0.1281	0.1450	0.1564	0.1708	0.2777
	LoRA FT 10%	0.0084	0.0700	0.2170	0.2104	0.1163	0.1493	0.3884

sequences in `trans_num`, integers parsed from numerical strings, spurious quotation marks, and hallucinated parameters—all causing immediate validation rejection.

In our runs, errors were noticeably less frequent when two conditions held:

- **Contextual Alignment:** The use of few-shot examples or LoRA fine-tuning provides the necessary grounding to align the models with the target schema.
- **Explicit Constraints:** The inclusion of explicit `constraints` within the prompt serves as a critical safeguard against formatting drift.

5.2 Distributional Fidelity

Numerical Distributional Fidelity Transaction amount (`amt`) is well-reproduced across both models and all strategies, with TVD below 0.1 in nearly all configurations. The right-skewed shape of consumer spending amounts is recognizable enough that models can approximate it from schema context alone, without requiring example records.

Geographic features (`lat`, `long`, `merch_lat`, `merch_long`) are where models consistently struggle to maintain numerical accuracy via ICL. Even at 10 examples, Qwen2.5-7B's longitude TVD remains at 0.2937; Mistral-7B reaches 0.1450 at the same example count. The problem is partly structural: US latitude–longitude pairs are bounded by land borders, and a handful of examples is nowhere near enough to convey that to the model. Only LoRA fine-tuning on Qwen2.5-7B brings geographic features within the 0.10 threshold.

The `unix_time` feature reveals an interesting asymmetry between the two models. Mistral-7B records TVD of 0.0005 at zero-shot ICL and 0.0001 at one-shot, while Qwen2.5-7B records 0.6931 at zero-shot ICL. However, the near-zero TVD for

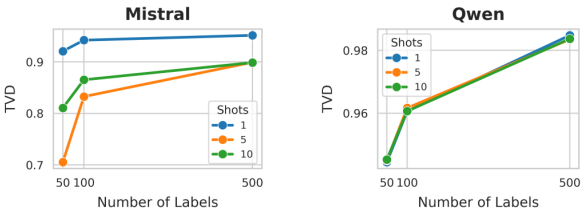
Table 3: Per-Feature Categorical TVD

Model	Strategy	category	gender	job	state
Qwen2.5-7B	ICL Zeroshot	0.6694	0.1250	1.0000	0.7268
	ICL Fewshot 1r	0.6962	0.4308	0.9927	0.4516
	ICL Fewshot 5r	0.5324	0.4526	0.9840	0.3855
	ICL Fewshot 10r	0.6396	0.4406	0.9837	0.3794
	LoRA FT 10%	0.0486	0.0012	0.1551	0.0535
	LoRA FT 50%	0.0474	0.0059	0.1430	0.0477
Mistral-7B	ICL Zeroshot	0.5528	0.0016	0.9999	0.7352
	ICL Fewshot 1r	0.5394	0.3488	0.9806	0.3515
	ICL Fewshot 5r	0.4293	0.2760	0.9593	0.3128
	ICL Fewshot 10r	0.4870	0.3048	0.9542	0.3015
	LoRA FT 10%	0.0687	0.1204	0.5943	0.2504

Mistral-7B most likely reflects the model defaulting to timestamps within a narrow range that happens to coincide with the real distribution's density peak, rather than any genuine modeling of the temporal distribution. If the model were genuinely learning the temporal distribution from examples, TVD should keep decreasing as more are added. Instead, it jumps from 0.0001 at one shot to 0.1986 at five, which suggests the additional examples pull the model away from a default that happened to work rather than actually improving its grasp of the temporal structure.

Qwen2.5-7B LoRA fine-tuning achieves TVD below 0.10 for all seven numerical features at both data exposure levels, with the 50% configuration showing modest but consistent improvement over 10%.

Categorical Distributional Fidelity The `job` field produces TVD near 1.0 under every ICL configuration for both models. In practice, this means the generated occupation labels have essentially no distributional overlap with the real data. Both models draw from a small vocabulary of common occupations that are well-represented in pre-training text but poorly representative of the actual dataset's distribution, which spans several hundred distinct occupations with a frequency structure that cannot be approximated from ten or fewer examples. Without weight updates (Brown et al., 2020), ICL can



(a) Mistral-7B-Instruct. (b) Qwen2.5-7B-Instruct.

Figure 2: TVD vs. cardinality. Zipf $\alpha = 0.5$

Table 4: Aggregate Mean TVD

Model	Strategy	Numerical TVD	Categorical TVD	Total TVD
Qwen2.5-7B	ICL Zeroshot	0.4892	0.6303	0.5405
	ICL Fewshot 1r	0.2972	0.6428	0.4229
	ICL Fewshot 5r	0.2268	0.5886	0.3584
	ICL Fewshot 10r	0.2151	0.6108	0.3590
	LoRA FT 10%	0.0383	0.0646	0.0479
	LoRA FT 50%	0.0319	0.0610	0.0425
Mistral-7B	ICL Zeroshot	0.4058	0.5724	0.4664
	ICL Fewshot 1r	0.1566	0.5551	0.3015
	ICL Fewshot 5r	0.1355	0.4944	0.2660
	ICL Fewshot 10r	0.1348	0.5119	0.2719
	LoRA FT 10%	0.1657	0.2585	0.1994

bias generation toward in-context values but cannot reshape the prior over the full label vocabulary.

LoRA fine-tuning on Qwen2.5-7B reduces `job` TVD to 0.1551 at 10% exposure and 0.1430 at 50%, still above the 0.10 threshold, but a substantial improvement over any ICL result. For `category`, `gender`, and `state`, both LoRA configurations achieve TVD below 0.06. Mistral-7B LoRA 10% partially recovers `category` and `gender` but cannot close the gap on `job` (0.5943) or `state` (0.2504), suggesting that the 10% training sample is insufficient to update the model's representation of high-cardinality and geographically structured features.

To determine whether the failures on `job` reflect a general ICL limitation, we evaluate TVD on synthetic Zipf-distributed features ($\alpha = 0.5$) across cardinalities of 50, 100, and 500, averaged over 30 runs. Most exemplary outcomes from the experimental results are presented in Figure 2. One might notice in particular how in Figure 2b all curves are virtually identical.

Aggregate Fidelity For Qwen2.5-7B, numerical TVD decreases consistently across the few-shot ICL series, from 0.489 at zero-shot to 0.215 at 10 examples. Categorical TVD shows no comparable trend—it oscillates between 0.589 and 0.643 with no directional improvement as example count increases. Adding more examples improves the numerical features but has essentially no effect on the categorical ones. Switching to LoRA fine-tuning is not simply the next step in the same trend—the drop is much sharper: mean numerical TVD goes from 0.215 to 0.038, and mean categorical TVD from 0.611 to 0.065.

For Mistral-7B, numerical TVD improves steadily through the few-shot series, reaching 0.135 at 10 examples—slightly better than Qwen2.5-7B at the same point. Fine-tuning, however, makes things worse: LoRA 10% yields a numerical TVD of 0.166, which is actually higher than the best ICL result. On the categorical side, LoRA 10% improves substantially over ICL (0.259 vs. 0.494–0.572), but remains far from the values Qwen2.5-7B achieves under fine-tuning.

Table 5: Privacy Metrics

Model	Strategy	DCR Ratio	Exact Copies (%)
Qwen2.5-7B	ICL Zeroshot	9.3597	0.00%
	ICL Fewshot 1r	4.4597	0.00%
	ICL Fewshot 5r	4.1291	0.00%
	ICL Fewshot 10r	20.4266	0.10%
	LoRA FT 10%	0.8655	0.10%
	LoRA FT 50%	0.8609	0.05%
Mistral-7B	ICL Zeroshot	37.2999	0.00%
	ICL Fewshot 1r	3.6881	0.00%
	ICL Fewshot 5r	1.1806	0.00%
	ICL Fewshot 10r	2.1070	0.00%
	LoRA FT 10%	0.9512	0.00%

5.3 Privacy and Memorization Risk

All ICL configurations yield DCR Ratios well above 1.0, ensuring synthetic records remain distant from the training data. The sole exception is Mistral-7B Fewshot 5r (1.1806), which approaches the memorization boundary. Conversely, all LoRA configurations fall below the 1.0 threshold; for Qwen2.5-7B, the proximity remains nearly identical at 10% (0.8655) and 50% (0.8609) exposure. This suggests that memorization signals saturate early rather than scaling with training volume. Notably, the exact copy rate for Qwen2.5-7B actually halves (from 0.10% to 0.05%) as exposure increases, confirming that distributional proximity and literal duplication are distinct phenomena. Mistral-7B LoRA 10% similarly dips below the threshold (0.9512) without producing exact copies.

5.4 Fraud Class Reproduction

Neither model produces a single fraudulent transaction under any ICL configuration—a clean failure on minority-class reproduction. LoRA fine-tuning partially restores the fraud class: Qwen2.5-7B produces 12 fraudulent records at 10% exposure and 14 at 50% (0.12% and 0.14% respectively), while Mistral-7B LoRA 10% produces only 5 (0.05%).

Table 6: Inter-Feature Correlation Errors (Qwen2.5-7B)

Strategy	Max Error	Total Error	Worst Pair
ICL Zeroshot	0.40	4.34	lat/long, merch_lat/merch_long
ICL Fewshot 10r	0.12	2.26	lat/long, merch_lat/merch_long
LoRA FT 50%	0.15	0.82	amt / is_fraud

All three fall well short of the real distribution's 0.58% base rate, meaning the class imbalance is substantially underrepresented even after fine-tuning.

5.5 Inter-Feature Correlation Preservation

Since TVD captures only marginal distributions, we assess dependency structures by computing the correlation matrix across the seven numerical features and the binary fraud label. We report two summary statistics in Table 6: the maximum absolute pairwise correlation error (the most discrepant pair) and the total absolute correlation error (the sum of all off-diagonal absolute errors in the 8×8 matrix). These metrics, evaluated for Qwen2.5-7B across zero-shot, few-shot (10r), and LoRA (50%) configurations, reveal whether synthetic datasets preserve the real data's underlying feature interactions.

Under zero-shot ICL, the most severe failures concentrate in the geographic block. The correlation between cardholder and merchant coordinates ($r > 0.99$ in real data) collapses to near-zero in synthetic output, reaching a maximum pairwise error of 0.40. As shown in Fig. 3, zero-shot models generate coordinates independently, lacking the spatial co-occurrence typical of legitimate transactions. Providing ten in-context examples halves the total correlation error (from 4.34 to 2.26) and reduces the geographic error to 0.12, suggesting partial inference of spatial coupling. However, Fig. 4 and 3 reveal that few-shot generation still fails to replicate co-location structure, instead clustering around urban centers. These issues, consistent with the high sensitivity of geographic TVD noted in Sec. 5.2, are resolved only via LoRA fine-tuning. LoRA FT (50% exposure) reduces total correlation error to 0.82—five times lower than zero-shot—and eliminates geographic discrepancies (Fig. 5). The worst-case error shifts to the `amt/is_fraud` pair (0.15); while the model learns to produce fraudulent records, it lacks the full calibration between high amounts and fraud present in real data. This aligns with the fraud underrepresentation in Sec. 5.4, as a deficit in minority-class samples inherently weakens the captured correlation. Consequently, correlation fidelity serves as a vital diagnostic for structural dependencies that marginal TVD statistics overlook.

6 Discussion

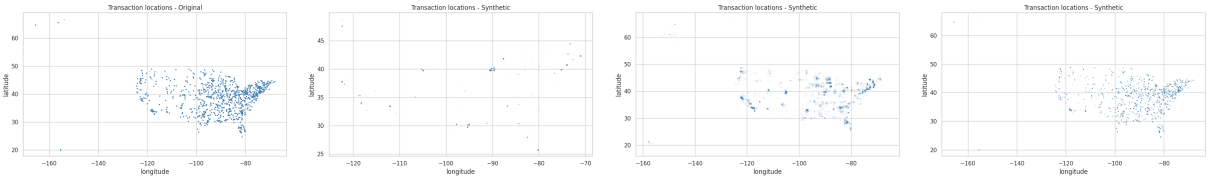
Categorical Prior Lock-in in In-Context Learning The central finding of this work is that in-context learning (ICL) exhibits a structural limitation when applied to structured generation tasks requiring adaptation to new categorical distributions. Across all ICL configurations and both models, high-cardinality categorical features remain poorly approximated, with total variation distance (TVD) approaching 1.0 regardless of the number of in-context examples. Increasing the number of examples improves numerical fidelity but does not produce any consistent improvement in categorical distributions, and minority categories are never reproduced.

As already pointed out in Section 1, we refer to the tendency of a language model to adhere to its pre-training distribution over tokens (even when conditioned on examples drawn from a different target distribution) as *categorical prior lock-in*. In the case of the `job` feature, both models repeatedly generate a small set of common occupations that are likely overrepresented in pre-training corpora, while failing to approximate the long-tailed empirical distribution present in the dataset.

This behavior reflects a fundamental property of ICL. While ICL conditions generation on a finite set of examples, it does not update the model's parameters and therefore cannot globally reshape the underlying token probability distribution. As a result, ICL can bias generation toward values observed in the prompt, but cannot reweight the full support of a high-cardinality categorical variable. When the target distribution differs significantly from the model's pre-training prior—as is typical for domain-specific structured data—this leads to systematic mismatch.

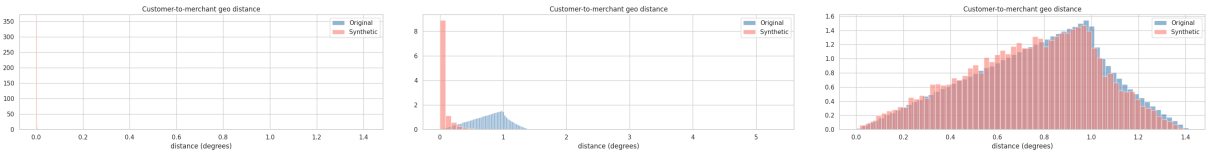
This limitation is not resolved by prompt design or example scaling within the tested range. Instead, it appears to be inherent to the mechanism of conditioning without learning: ICL operates locally in context, while the task of approximating a high-cardinality distribution requires global adjustment of token probabilities. Our results suggest that this gap cannot be closed without parameter updates.

Beyond Marginals: Structural Failures of ICL The effects of categorical prior lock-in extend be-



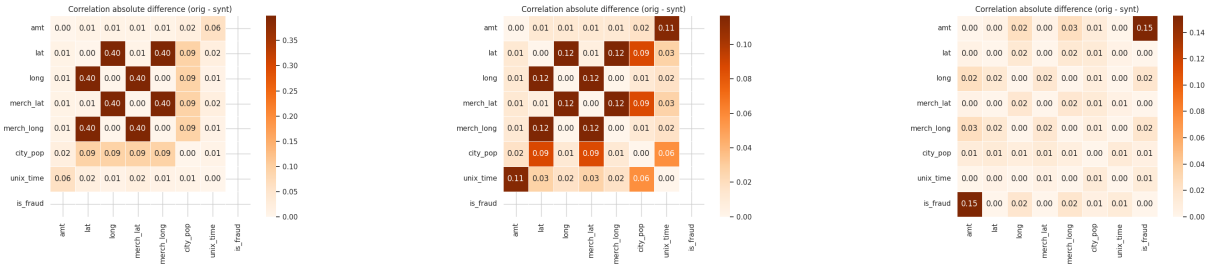
(a) Original (b) Zero-shot (c) Few-shot 10r (d) LoRA 50%

Figure 3: Geographic distribution across generation strategies (Qwen2.5-7B)



(a) Zero-shot (b) Few-shot 10r (c) LoRA 50%

Figure 4: Geographic distance across generation strategies (Qwen2.5-7B)



(a) Zero-shot (b) Few-shot 10r (c) LoRA 50%

Figure 5: Correlation heatmap across generation strategies (Qwen2.5-7B)

yond marginal distributions. Under zero-shot ICL, geographic features are generated independently, collapsing strong correlations present in the real data; few-shot prompting partially reduces these errors but does not recover the underlying dependency structure. Because categorical variables often serve as anchors for inter-feature dependencies, failure to model their distribution propagates to the broader joint structure of the data. These results suggest that ICL is insufficient for structured generation tasks requiring both accurate marginals and coherent joint distributions under distribution shift.

Fine-Tuning as Distributional Adaptation In contrast to ICL, parameter-efficient fine-tuning (LoRA) substantially improves both categorical and numerical fidelity. For Qwen2.5-7B, fine-tuning reduces categorical TVD by an order of magnitude and restores inter-feature correlations that are absent under ICL. These results support the interpretation that weight updates enable the model to adapt its internal representation of token distributions to match the target dataset.

From this perspective, fine-tuning can be understood as performing the global distributional adjustment that ICL cannot. By modifying model parameters, fine-tuning reshapes the probability mass assigned to tokens across the full vocabulary, allowing the model to represent high-cardinality categorical distributions more accurately.

However, this improvement comes at a cost. All fine-tuned configurations exhibit DCR ratios below 1.0, indicating that synthetic records are, on average, closer to real records than real records are to each other. This suggests the emergence of memorization behavior, even at relatively low data exposure levels. The trade-off is therefore clear: while fine-tuning enables distributional adaptation, it introduces risks associated with data leakage and reduced privacy.

Instability of Fine-Tuning in Small Instruction-Tuned Models A key finding of our experiments is that fine-tuning can degrade instruction-following in small models. The behaviour of Mistral-7B under LoRA fine-tuning illustrates this

risk concretely: unlike Qwen2.5-7B, which maintains stable structured output generation after adaptation, Mistral-7B exhibits severe degradation in output validity, culminating in complete generation failure at higher data exposure levels—producing no valid records whatsoever in 1 000 attempts at 50% data exposure. This is not a marginal degradation but a catastrophic collapse of the model's ability to follow the structured output schema it had previously adhered to under ICL. This instability aligns with prior evidence that fine-tuning small models can be harmful, potentially because specialized adaptation saturates limited model capacity (Wei et al., 2022) or induces catastrophic forgetting of core instruction-following abilities during domain specialization (Fatemi et al., 2024).

One possible interpretation is that the additive updates introduced by LoRA interfere with the latent representations responsible for schema adherence. While the base weights remain frozen, the injected low-rank signal may disrupt the pre-existing instruction-following logic. Small models with less representational redundancy, such as Mistral-7B, appear particularly vulnerable to this low-rank interference, which effectively overrides the model's structural priors. These results indicate that output stability depends on the compatibility between the adapter's signal and the base model's architecture, a factor that must be evaluated alongside standard performance metrics.

Implications for LLM-Based Structured Generation The findings of this work point to a fundamental trade-off in the use of LLMs for structured data generation. In-context learning offers a lightweight and privacy-preserving mechanism for adapting models to new tasks, but is fundamentally limited in its ability to approximate distributions that diverge from pre-training priors. Fine-tuning overcomes these limitations by enabling global adaptation, but introduces risks related to memorization and, in some cases, model instability.

More broadly, these results suggest that ICL should not be interpreted as a general-purpose substitute for learning in settings that require distributional adaptation. Instead, its strengths lie in tasks that align closely with pre-training distributions or require only local adjustments. For structured generation tasks involving high-cardinality categorical variables or strong distribution shift, parameter updates appear to be necessary.

This perspective reframes the role of ICL within the broader landscape of LLM adaptation methods: rather than a universal mechanism for few-shot learning, it is a form of conditional generation with inherent limitations determined by the model's prior. Understanding these limitations is essential for deploying LLMs in real-world settings where distribution mismatch is the norm rather than the exception.

7 Conclusion

We investigated the viability of 7B-scale language models for synthetic tabular data generation, comparing in-context learning (ICL) and LoRA fine-tuning across two models (Qwen2.5-7B and Mistral-7B) on a fraud detection dataset. Our results show that the generation paradigm—rather than model choice—is the primary determinant of data quality at this scale. ICL proves insufficient for high-cardinality categorical features: marginal fidelity on *job* remains near-random regardless of shot count or model, and neither model generates fraudulent transactions under any ICL configuration, rendering ICL-produced data unsuitable for fraud detection tasks. Numerical fidelity and inter-feature correlation structure improve with more in-context examples but plateau well below the quality achievable through fine-tuning. LoRA fine-tuning on Qwen2.5-7B substantially closes these gaps, recovering both marginal categorical distributions and geographic correlation structure. However, DCR analysis reveals a memorization signal that persists across data exposure levels, raising concerns about the privacy guarantees of the generated data. Mistral-7B proves unstable under LoRA adaptation at the configurations tested, failing to produce valid outputs at higher data exposure.

Future work will focus on validating these insights across a broader range of tabular domains and incorporating formal downstream utility evaluations. Additionally, we aim to integrate differentially private fine-tuning mechanisms (e.g., DP-LoRA) into the pipeline to formally mitigate the identified memorization risks while preserving structural and distributional fidelity.

8 Limitations

While our work systematically characterizes the trade-offs between in-context learning and parameter-efficient fine-tuning for structured tabular generation, we acknowledge two primary limitations:

First, our empirical evaluation is restricted to a single dataset: credit card transaction log optimized for fraud detection. Although this dataset provides a highly controlled environment to test extreme class imbalance and high-cardinality attributes (such as the *job* feature with 494 distinct labels), our findings regarding categorical prior lock-in and adapter instability may not automatically generalize to other structural paradigms. Tabular datasets from various fields, such as healthcare, census data, or financial portfolios, often exhibit entirely different density characteristics, semantic constraints, and relational structures that require independent validation.

Second, the evaluation framework relies entirely on statistical proximity metrics, namely marginal Total Variation Distance (TVD), absolute correlation errors, and the Distance to Closest Record (DCR) ratio. While these metrics provide rigorous mathematical bounds on distributional fidelity and memorization risks, they serve as approximations and do not replace a complete downstream machine learning utility evaluation. A comprehensive assessment–involving the training of downstream predictive classifiers (e.g., gradient-boosted trees or random forests) on the synthetic corpora to evaluate their generalization performance on a real holdout test set–remains a necessary step to fully determine the operational utility of the generated data.

References

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023. Language models are realistic tabular data generators. *Preprint*, arXiv:2210.06280.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Sorouralsadat Fatemi, Yuheng Hu, and Maryam Mousavi. 2024. A comparative analysis of instruction fine-tuning llms for financial text classification. *Preprint*, arXiv:2411.02476.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jinhee Kim, Taesung Kim, and Jaegul Choo. 2025. Epic: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models. *Preprint*, arXiv:2404.12404.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. 2024. Curated llm: Synergy of llms and data curation for tabular augmentation in low-data regimes. *Preprint*, arXiv:2312.12112.

Aivin V. Solatorio and Olivier Dupriez. 2023. Realtabformer: Generating realistic relational and tabular data using transformers. *Preprint*, arXiv:2302.02041.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Preprint*, arXiv:1907.00503.

Zilong Zhao, Robert Birke, and Lydia Chen. 2025. Tabula: Harnessing language models for tabular data synthesis. *Preprint*, arXiv:2310.12746.