Summary of Revisions

We appreciate all four reviewers for their valuable feedback and supportive comments.

Our Key Strengths Highlighted by Reviewers:

1. **Addressing a Crucial, Relevant Problem**:
   This work has a strong motivation for an important issue that is of broad importance, providing preliminary studies to analyze class distributions, which motivates the proposed approach (Reviewer bNDz, poPD, 86JX).
2. **Interesting LLM-based Methodology**: The proposed LLM-based, prompt-based method for generating synthetic tabular data is recognized as interesting, directly utilizing a group-wise prompting method to generate data in CSV format (Reviewer yJwy).
3. **Comprehensive Experimental Evaluation**: The experimental evaluation is comprehensive, covering a good set of experiments and a broad set of baselines, which includes analyses of the model's ability to address the class imbalance, an extensive ablation study to analyze the contribution of CSV style and group-wise style, LLM generation efficiency, and the similarity between generated data distribution and the original data and additional feature correlation study (Reviewer bNDz, yJwy, poPD, 86JX).
4. **Strong Performance**: The performance looks good, outperforming the proposed method with previous state-of-the-art table generation baselines and demonstrating that the proposed method improves downstream performance and outperforms other baselines (Reviewer poPD, 86JX).
5. **Good Presentation**: The visualization and detailed experimental descriptions help readers understand the results better (Reviewer yJwy).

Incorporating the reviewers' feedback, we highlight key reviews we addressed:

- We have shared our source code and prompt examples to improve transparency and replicability: source code (https://anonymous.4open.science/r/ST-Prompt-29F2/README.md), prompt examples (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_Sick_ST-Prompt_ICML_24.jpg) (Reviewer yJwy, poPD).
- We additionally adopted our method to open-source LLMs, including Llama2 and Mistral. (Reviewer yJwy)
- Four additional baseline methods (now a total of 10 baselines) were added to our comparison, enriching our analysis (Reviewer poPD, 86JX). `General response 1`
- Three additional strong classifiers were used (now a total of four classifiers) in our experiments to validate our method’s effectiveness (Reviewer yJwy, 86JX). `General response 2`
- Significance tests on experimental results were conducted to confirm statistically significant improvements (Reviewer bNDz, yJwy). `General response 3`

- Two additional real-world datasets were included to ensure robustness against LLM pretraining (Reviewer bNDz).
- We further detailed our random word replacement method and conducted an ablation study (Reviwer poPD).
- Notation errors, typos, and missing details will meticulously revised in the main manuscript and appendix.

We trust these responses comprehensively address all reviewers' concerns.
We are enthusiastic about the further impact and insights our approach offers.
We are eager to provide any further information or clarification required.

General Response 1: Comparison with additional baseline models

> Q2. More recent baselines need to be included, such as [1,2]. (Reviewer poPD)
> Q2: Lack of prompt optimization baselines (mentioned in Appendix E) and other tabular data generation baselines (mentioned in [3]). (Reviewer 86JX)

A1. In our paper, we initially compared our method with six models focused on tabular data generation and class imbalance, such as GReaT, CTGAN, CopulaGAN, TVAE, SMOTE, and SMOTENC.
In response to the valueable feedback, we have incorporated four additional baselines, bringing our total comparison to 10 distinct methods.

Specifically, we added CTAB-GAN+ [1] and CuratedLLM [2], as recommended by Reviewer poPD, and TabDDPM [4] and CTAB-GAN [5] model from the synthetic tabular data generation benchmark paper [3] (CTGAN and T-VAE in this paper were already included in our initial analysis), recommended by Reviewer 86JX.

Given the space constraints, we prioritized key metrics that best reflect the model performance: the F1 Score (harmonic mean of sensitivity and precision) and Balanced Accuracy (arithmetic mean of sensitivity and specificity).

We utilized the official open-source code [7] from the TabDDPM GitHub repository to reproduce CTAB-GAN, CTAB-GAN+, and TabDDPM. For the Adult Income and California datasets, the hyperparameters provided were used. For other datasets lacking provided hyperparameters, we applied the hyperparameters from the Wilt dataset, which, according to the corresponding paper, demonstrated the largest performance improvement over the original data.

Due to the absence of source code for CuratedLLM, we could not accurately reimplement their proposed curation mechanism. Instead, we reproduced the tabular data generation method using the LLM prompts proposed by them, employing the gpt-3.5-turbo-0613 model.

Table 1-1. Balanced accuracy for additional baseline comparison.

data	*model*	*Original*	*+ CTABGAN*	*+ CTABGAN+*	*+ TabDDPM*	*+ CuratedLLM*	**+ Ours**
Travel	*GB*	0.723±0.000	0.762±0.000	0.718±0.000	0.711±0.000	0.638±0.000	**0.808**±0.000
	XGB	0.691±0.000	0.638±0.000	0.695±0.000	0.663±0.000	0.691±0.000	**0.797**±0.000
	Cat	0.703±0.000	0.657±0.003	0.675±0.000	0.643±0.011	0.651±0.000	**0.765**±0.000
	LGBM	0.723±0.000	0.726±0.000	0.655±0.000	0.691±0.000	0.658±0.000	**0.760**±0.000
	Mean	0.710±0.014	0.696±0.051	0.686±0.024	0.677±0.027	0.660±0.020	**0.782**±0.021
Sick	*GB*	0.899±0.000	0.900±0.000	0.868±0.000	0.889±0.000	0.933±0.000	**0.934**±0.000
	XGB	0.912±0.000	0.879±0.000	0.869±0.000	0.889±0.000	0.889±0.000	**0.921**±0.000
	Cat	0.915±0.006	0.890±0.000	0.824±0.000	0.882±0.006	0.910±0.000	**0.921**±0.000
	LGBM	0.923±0.000	0.912±0.000	0.891±0.000	0.912±0.000	0.924±0.000	**0.941**±0.000
	Mean	0.912±0.010	0.895±0.012	0.863±0.025	0.893±0.012	0.914±0.017	**0.929**±0.009
HELOC	*GB*	0.734±0.000	0.730±0.000	0.731±0.000	0.729±0.000	0.728±0.000	**0.737**±0.000
	XGB	0.727±0.000	0.727±0.000	0.733±0.000	0.729±0.000	0.725±0.000	**0.739**±0.000
	Cat	0.732±0.001	0.732±0.001	0.731±0.001	0.730±0.001	0.732±0.001	**0.735**±0.001
	LGBM	**0.735**±0.000	0.726±0.000	0.732±0.000	0.727±0.000	0.732±0.000	**0.735**±0.000
	Mean	0.732±0.003	0.729±0.002	0.732±0.001	0.729±0.001	0.729±0.003	**0.737**±0.002

data	*model*	*Original*	*+ CTABGAN*	*+ CTABGAN+*	*+ TabDDPM*	*+ CuratedLLM*	**+ Ours**
Adult Income	*GB*	0.780±0.000	0.778±0.000	0.774±0.000	0.776±0.000	0.780±0.000	**0.799**±0.000
	XGB	0.746±0.000	0.743±0.000	0.751±0.000	0.746±0.000	0.764±0.000	**0.779**±0.000
	Cat	0.755±0.001	0.755±0.000	0.761±0.001	0.759±0.001	0.772±0.001	**0.790**±0.000
	LGBM	0.777±0.000	0.776±0.000	0.771±0.000	0.779±0.000	0.779±0.000	**0.798**±0.000
	Mean	0.764±0.015	0.763±0.015	0.764±0.009	0.765±0.013	0.774±0.007	**0.792**±0.008
Diabetes	*GB*	0.433±0.000	0.426±0.000	0.425±0.000	0.430±0.000	0.428±0.000	**0.435**±0.000
	XGB	0.411±0.000	0.407±0.000	0.407±0.000	0.409±0.000	0.406±0.000	**0.413**±0.000
	Cat	**0.407**±0.000	0.402±0.000	0.404±0.000	0.403±0.000	0.401±0.000	0.405±0.000
	LGBM	0.432±0.000	0.426±0.000	0.424±0.000	0.431±0.000	0.427±0.000	**0.433**±0.000
	Mean	**0.421**±0.012	0.415±0.011	0.415±0.010	0.418±0.013	0.415±0.013	**0.421**±0.013

Table 1-2. F1 score for additional baseline comparison.

data	*model*	*Original*	*+ CTAB − GAN*	*+ CTAB − GAN+*	*+ TabDDPM*	*+ CuratedLLM*	*+ Ours*
Travel	*GB*	0.600±0.000	0.667±0.000	0.596±0.000	0.583±0.000	0.465±0.000	**0.702**±0.000
	XGB	0.553±0.000	0.465±0.000	0.560±0.000	0.511±0.000	0.553±0.000	**0.677**±0.000
	Cat	0.571±0.000	0.502±0.005	0.531±0.000	0.481±0.017	0.489±0.000	**0.644**±0.000
	LGBM	0.600±0.000	0.609±0.000	0.500±0.000	0.553±0.000	0.500±0.000	**0.643**±0.000
	Mean	0.581±0.020	0.561±0.083	0.547±0.036	0.532±0.041	0.502±0.033	**0.667**±0.025
Sick	*GB*	0.841±0.000	0.860±0.000	0.829±0.000	0.847±0.000	0.899±0.000	**0.920**±0.000
	XGB	**0.884**±0.000	0.843±0.000	0.840±0.000	0.847±0.000	0.847±0.000	0.876±0.000
	Cat	**0.881**±0.007	0.857±0.000	0.758±0.004	0.829±0.007	0.866±0.004	0.871±0.005
	LGBM	0.907±0.000	0.884±0.000	0.867±0.000	0.884±0.000	**0.918**±0.000	0.882±0.000
	Mean	0.878±0.025	0.861±0.015	0.823±0.042	0.852±0.021	0.882±0.028	**0.887**±0.020
HELOC	*GB*	0.713±0.000	0.709±0.000	0.710±0.000	0.706±0.000	0.711±0.000	**0.719**±0.000
	XGB	0.702±0.000	0.701±0.000	0.711±0.000	0.706±0.000	0.706±0.000	**0.721**±0.000
	Cat	0.712±0.001	0.710±0.001	0.710±0.001	0.709±0.001	0.717±0.001	**0.718**±0.001
	LGBM	0.713±0.000	0.704±0.000	0.711±0.000	0.705±0.000	0.714±0.000	**0.719**±0.000
	Mean	0.710±0.005	0.706±0.004	0.710±0.000	0.707±0.002	0.712±0.004	**0.719**±0.001

data	*model*	*Original*	*+ CTAB − GAN*	*+ CTAB − GAN+*	*+ TabDDPM*	*+ CuratedLLM*	*+ Ours*
Adult Income	*GB*	0.690±0.000	0.686±0.000	0.677±0.000	0.684±0.000	0.683±0.000	**0.702**±0.000
	XGB	0.642±0.000	0.637±0.000	0.649±0.000	0.643±0.000	0.660±0.000	**0.677**±0.000
	Cat	0.656±0.001	0.654±0.001	0.659±0.001	0.660±0.001	0.669±0.001	**0.688**±0.000
	LGBM	0.688±0.000	0.683±0.000	0.673±0.000	0.687±0.000	0.682±0.000	**0.700**±0.000
	Mean	0.669±0.021	0.665±0.021	0.665±0.011	0.668±0.018	0.673±0.010	**0.692**±0.010
Diabetes	*GB*	0.562±0.000	0.554±0.000	0.554±0.000	0.560±0.000	0.557±0.000	**0.564**±0.000
	XGB	0.537±0.000	0.533±0.000	0.533±0.000	0.536±0.000	0.530±0.000	**0.540**±0.000
	Cat	**0.533**±0.000	0.527±0.001	0.529±0.000	0.528±0.000	0.526±0.000	0.532±0.000
	LGBM	0.561±0.000	0.555±0.000	0.554±0.000	0.561±0.000	0.557±0.000	**0.562**±0.000
	Mean	**0.549**±0.014	0.542±0.013	0.542±0.012	0.546±0.015	0.542±0.015	**0.549**±0.014

The results demonstrate that our proposed method consistently outperforms all baseline models across various datasets. These results bolster the robustness and effectiveness of our approach, demonstrating its superiority against the comprehensive set of latest advancements in the field.

Regarding prompt optimization, following Reviewer 86JX's suggestions, we expanded our baseline comparison to include an additional prompting method, zero-shot-CoT [6], referenced in our related work (Appendix E). While these methods were not initially designed for tabular data generation, their inclusion offers a broader perspective on prompt efficacy.

To apply the zero-shot-CoT approach, we refined our initial prompts by prefixing them with "Q:" to signify questions and appending "A: Let's think step by step." as a prompt for detailed reasoning. With this modification, the LLM suggested Python code for tabular data analysis or analyzing the data columns rather than generating new data based on demonstrations. When we further refined the prompt to "Q: Generate tabular datasets following the patterns in the provided example data.", the LLM produced synthetic tabular data. However, the generated data often deviated from the example patterns, focusing excessively on a single class or adopting markdown formatting instead of the expected CSV format. This inconsistency can lead to an additional cost for analyzing the generated results, as they hinder automatic data parsing.

The results highlight the necessity for a prompting method specifically tailored to tabular data generation. This underscores the importance of our proposed prompting techniques, which are crucial for the successful creation of high-quality synthetic datasets.

Through these additional experiments and analyses, we have significantly strengthened our manuscript, providing a more comprehensive and robust comparison with current methods in both synthetic tabular data generation and prompt optimization.

_

[1] Zhao, Zilong, et al. "Ctab-gan+: Enhancing tabular data synthesis." Frontiers in big Data 6 (2023).
[2] Seedat, Nabeel, et al. "Curated LLM: Synergy of LLMs and Data Curation for tabular augmentation in ultra low-data regimes." arXiv preprint arXiv:2312.12112 (2023).
[3] Hansen, Lasse, et al. "Reimagining synthetic tabular data generation through data-centric AI: A comprehensive benchmark." Advances in Neural Information Processing Systems (2023).
[4] Kotelnikov, Akim, et al. "Tabddpm: Modelling tabular data with diffusion models." International Conference on Machine Learning (2023).
[5] Zhao, Zilong, et al. "Ctab-gan: Effective table data synthesizing." Asian Conference on Machine Learning (2021).
[6] Kojima, Takeshi, et al. "Large language models are zero-shot reasoners." Advances in neural information processing systems (2022).
[7] https://github.com/yandex-research/tab-ddpm (https://github.com/yandex-research/tab-ddpm)

General Response 2: Experimental results on additional machine learning models

> Q3. The use of a relatively weak classifier. Considering stronger classifiers, such as XGBoost, as the backbone could potentially improve the results (Reviewer 86JX).
> Q2. The experiments only test gradient boosting gradient classifier for classification and random forest regression for regression. Moreover, the classifier is tested on five out of all six datasets. The random forest is tested on two out of all six datasets. Compared to the previous paper, GREAT, this experiment is not very complete. (Reviewer yJwy)

A2. To address the feedback regarding the range of machine learning models used in our experiments, we have broadened our experimental scope. We now include three additional, widely recognized machine learning models: XGBoost [1], as recommended by 86JX, CatBoost [2], and LightGBM [3], known for its robust performance.

Regarding hyperparameter settings: For the gradient boosting classifier (GB), we continued to use the default hyperparameters provided by scikit-learn. For the newly added machine learning models, inlcluding XGBoost, CatBoost, and LightGBM, we conducted 5-fold cross-validation on the training set of the *original data*, optimizing two key hyperparameters: learning rate and max_depth. These optimized settings were then consistently applied across all experiments, aligning the model's performance closely with the characteristics of the original data.

Due to the lack of space, we focus on reporting on two crucial metrics, the F1 Score (harmonic mean of sensitivity and precision) and Balanced Accuracy (arithmetic mean of sensitivity and specificity).

Table 2-1. Balanced Accuracy for additional machine learning model comparison.

data	*model*	*Original*	*+TVAE*	*+CopulaGAN*	*+CTGAN*	*+GReaT*	*+Ours*
Travel	*GB*	0.723±0.000	0.708±0.000	0.545±0.000	0.557±0.000	0.718±0.000	**0.808**±0.000
	XGBoost	0.691±0.000	0.763±0.000	0.560±0.000	0.585±0.000	0.751±0.000	**0.797**±0.000
	CatBoost	0.703±0.000	0.675±0.000	0.556±0.011	0.593±0.016	0.707±0.009	**0.765**±0.000
	LightGBM	0.723±0.000	0.748±0.000	0.560±0.000	0.577±0.000	0.738±0.000	**0.760**±0.000
	Mean	0.710±0.014	0.723±0.035	0.555±0.008	0.578±0.016	0.729±0.018	**0.782**±0.021
Sick	*GB*	0.899±0.000	0.899±0.000	0.869±0.000	0.900±0.000	0.899±0.000	**0.934**±0.000
	XGBoost	0.912±0.000	0.909±0.000	0.870±0.000	0.890±0.000	0.912±0.000	**0.921**±0.000
	CatBoost	0.915±0.006	0.916±0.006	0.869±0.000	0.902±0.000	**0.922**±0.011	0.921±0.000
	LightGBM	0.923±0.000	0.934±0.000	0.857±0.000	0.902±0.000	0.900±0.000	**0.941**±0.000
	Mean	0.912±0.010	0.915±0.013	0.866±0.005	0.899±0.005	0.908±0.011	**0.929**±0.009
HELOC	*GB*	0.734±0.000	0.734±0.000	0.736±0.000	0.732±0.000	0.732±0.000	**0.737**±0.000
	XGBoost	0.727±0.000	0.731±0.000	0.735±0.000	0.734±0.000	0.726±0.000	**0.739**±0.000
	CatBoost	0.732±0.001	0.728±0.001	0.730±0.001	0.731±0.001	0.731±0.001	**0.735**±0.001
	LightGBM	0.735±0.000	**0.737**±0.000	0.732±0.000	0.726±0.000	0.730±0.000	0.735±0.000
	Mean	0.732±0.003	0.732±0.003	0.733±0.003	0.731±0.003	0.730±0.002	**0.737**±0.002

data	*model*	*Original*	*+ TVAE*	*+ CopulaGAN*	*+ CTGAN*	*+ GReaT*	*+ Ours*
Adult	*GB*	0.780±0.000	0.779±0.000	0.780±0.000	0.769±0.000	0.784±0.000	**0.799**±0.000
Income	*XGBoost*	0.746±0.000	0.755±0.000	0.750±0.000	0.754±0.000	0.762±0.000	**0.779**±0.000
	CatBoost	0.755±0.001	0.760±0.000	0.760±0.001	0.758±0.001	0.768±0.000	**0.790**±0.000
	LightGBM	0.777±0.000	0.778±0.000	0.780±0.000	0.769±0.000	0.786±0.000	**0.798**±0.000
	Mean	0.764±0.015	0.768±0.011	0.767±0.013	0.763±0.007	0.775±0.011	**0.792**±0.008
Diabetes	*GB*	0.433±0.000	0.433±0.000	0.430±0.000	0.429±0.000	0.431±0.000	**0.435**±0.000
	XGBoost	0.411±0.000	0.410±0.000	0.405±0.000	0.411±0.000	0.411±0.000	**0.413**±0.000
	CatBoost	0.407±0.000	0.405±0.000	0.402±0.000	**0.407**±0.000	0.406±0.000	0.405±0.000
	LightGBM	0.432±0.000	0.430±0.000	0.427±0.000	0.429±0.000	0.431±0.000	**0.433**±0.000
	Mean	**0.421**±0.012	0.420±0.012	0.416±0.013	0.419±0.010	0.420±0.012	**0.421**±0.013

Table 2-2. F1 score for additional machine learning model comparison.

data	*model*	*Original*	*+ TVAE*	*+ CopulaGAN*	*+ CTGAN*	*+ GReaT*	*+ Ours*
Travel	*GB*	0.600±0.000	0.577±0.000	0.200±0.000	0.250±0.000	0.596±0.000	**0.702**±0.000
	XGBoost	0.553±0.000	0.654±0.000	0.214±0.000	0.312±0.000	0.640±0.000	**0.677**±0.000
	CatBoost	0.571±0.000	0.531±0.000	0.242±0.027	0.328±0.043	0.577±0.013	**0.644**±0.000
	LightGBM	0.600±0.000	0.630±0.000	0.214±0.000	0.303±0.000	0.625±0.000	**0.643**±0.000
	Mean	0.581±0.020	0.598±0.049	0.218±0.020	0.298±0.036	0.610±0.026	**0.667**±0.025
Sick	*GB*	0.841±0.000	0.851±0.000	0.840±0.000	0.860±0.000	0.851±0.000	**0.920**±0.000
	XGBoost	**0.884**±0.000	0.854±0.000	0.850±0.000	0.857±0.000	**0.884**±0.000	0.876±0.000
	CatBoost	0.881±0.007	0.887±0.006	0.840±0.000	0.892±0.000	**0.894**±0.011	0.871±0.005
	LightGBM	0.907±0.000	**0.920**±0.000	0.815±0.000	0.892±0.000	0.860±0.000	0.882±0.000
	Mean	0.878±0.025	0.878±0.029	0.836±0.013	0.875±0.017	0.872±0.019	**0.887**±0.020
HELOC	*GB*	0.713±0.000	0.713±0.000	0.716±0.000	0.710±0.000	0.706±0.000	**0.719**±0.000
	XGBoost	0.702±0.000	0.710±0.000	0.713±0.000	0.710±0.000	0.699±0.000	**0.721**±0.000
	CatBoost	0.712±0.001	0.707±0.001	0.709±0.001	0.708±0.001	0.707±0.001	**0.718**±0.001
	LightGBM	0.713±0.000	0.715±0.000	0.712±0.000	0.705±0.000	0.703±0.000	**0.719**±0.000
	Mean	0.710±0.005	0.711±0.003	0.712±0.003	0.708±0.002	0.704±0.003	**0.719**±0.001

data	*model*	*Original*	*+ TVAE*	*+ CopulaGAN*	*+ CTGAN*	*+ GReaT*	*+ Ours*
Adult	*GB*	0.690±0.000	0.683±0.000	0.684±0.000	0.670±0.000	0.692±0.000	**0.702**±0.000
Income	*XGBoost*	0.642±0.000	0.653±0.000	0.646±0.000	0.652±0.000	0.664±0.000	**0.677**±0.000
	CatBoost	0.656±0.001	0.660±0.001	0.656±0.001	0.656±0.001	0.669±0.001	**0.688**±0.000
	LightGBM	0.688±0.000	0.682±0.000	0.683±0.000	0.670±0.000	0.693±0.000	**0.700**±0.000
	Mean	0.669±0.021	0.670±0.014	0.667±0.017	0.662±0.008	0.679±0.014	**0.692**±0.010
Diabetes	*GB*	0.562±0.000	0.563±0.000	0.558±0.000	0.558±0.000	0.560±0.000	**0.564**±0.000
	XGBoost	0.537±0.000	0.537±0.000	0.529±0.000	0.538±0.000	0.537±0.000	**0.540**±0.000
	CatBoost	0.533±0.000	0.532±0.000	0.528±0.001	**0.535**±0.000	0.533±0.000	0.532±0.000
	LightGBM	0.561±0.000	0.560±0.000	0.556±0.000	0.559±0.000	0.561±0.000	**0.562**±0.000
	Mean	**0.549**±0.014	0.548±0.014	0.543±0.015	0.547±0.011	0.548±0.013	**0.549**±0.014

Our extended analysis further substantiates the robustness and effectiveness of our method, demonstrating the best average balanced accuracy and F1 score across various datasets, affirming its superiority and versatility. It reveals that our approach enhances performance across these diverse machine learning models, consistently outperforming the original baselines.

[1] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (2016).
[2] Prokhorenkova, Liudmila, et al. "CatBoost: unbiased boosting with categorical features." Advances in neural information processing systems (2018).
[3] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems (2017).

General Response 3: Statistical significance of the proposed method

> Q1-1. In addition the F1/MSE gains are rather small and it is not clear if they are statistically significant. In addition, the performance degrades with more samples (see Table 4). This shows the inherent variance in the process and statistical significance tests are important (Reviewer bNDz).
> Q2. The gains of some metrics are really very small. The paper probably needs to add a significance test for those data pairs (e.g., 0.271 vs 0.274) (Reviewer yJwy).

A3. To address the concerns, we conducted independent two-sample t-tests on the results presented in Tables 2 and 4 of our manuscript.

We would like to clarify regarding '0.271 vs 0.274' as mentioned by Reviewer yJwy that both values are results of our method under different hyperparameter settings for group numbers in the regression task.

Our statistical tests were conducted over five separate runs, each with a unique random seed, to ensure robustness and adherence to the original experimental settings.
As a result, for the F1 metric, all p-values were found to be less than 0.001 when comparing our method with each of the baseline models. Similarly, for the balanced accuracy metric, all p-values were less than 0.01 across comparisons.
For the mean absolute error and mean squared error metrics, all p-values were smaller than 0.001 when comparing our method with baselines for generating 30K data points.

The consistently low p-values across key metrics further validate the quality and reliability of the data generated by our method. These results strongly confirm that our proposed method achieves statistically significant improvements over existing baseline models.

General Response 4: Novelty

> Q3. The proposed method is kind of incremental. The few-shot method and in-context learning have been widely applied to LLM.(Reviewer yJwy)
> Q1. The contribution is limited, as the paper mainly combines different tricks in prompt engineering. (Reviewer poPD)
> Q1. The novelty of the proposed method is limited. (Reviewer 86JX)

A4. We understand the concerns regarding the novelty of our approach, given the use of few-shot methods and in-context learning with LLMs. However, our approach stands out in its innovative application and significant improvements over existing baselines, particularly in tabular data generation, a domain less explored with LLMs.

Our method capitalizes on the inherent pattern recognition capability of Large Language Models (LLMs), trained on massive datasets, to recognize and predict patterns within the input. This capability allows our LLM to create high-quality examples by understanding contrasting patterns across different classes in tabular data.

We achieve this by proposing a simple yet effective method, ST-prompt.
We are the first to propose a strategic grouping method with CSV-style format, specifically designed to generate more accurate and representative data, while optimizing token efficiency and preserving inter-feature correlations in the generated data. This is complemented by our novel Random Word Replacement Strategy, enhancing the LLM's accuracy and representativeness in handling monotonous categorical values in tabular data.
While the concept may seem straightforward, its application in this context is novel and it led to the significant performance improvement in the important problem in the field.

LLMs are known to be highly sensitive to prompt method, well-acknowledged challenge in the field, as evidenced by the considerable body of research on this topic, including multiple papers presented at ICML [1,2,3,4,5]. We also contributes to this evolving field by demonstrating a novel application of prompt design specifically tailored for tabular data generation.

Our experiments demonstrate that simple prompting method without our tailored approach fails to generate good quality of data, leading to loss of feature correlations and lower generation efficacy. Our method overcomes these limitations, as proven by our results (Tables 5, 6, and Figure 4 in our manuscript). Moreover, our extensive evaluations demonstrate the robustness and effectiveness of our method, covering a variety of models and datasets. Our method consistently outperforms 10 different baselines and validated using four machine learning models across eight distinct datasets. Also, we have shown that our method effecitvely work on two open source LLMs, Llama2 and Mistral, additional to GPT 3.5.

Furthermore, while existing baselines in this field often showcase standalone experimental results, our approach not only demonstrates standalone capabilities (Table 7 and 8 in our manuscript) but also significantly enhances performance when our generated data is added to existing datasets (Table 2, 3, and 4 in our manuscript). This is particularly crucial in scenarios of class imbalance. Given the prevalence of imbalanced data in real-world scenarios, our research provides valuable insights into effectively using LLMs for such datasets, a topic of high relevance and interest in the field.

The rigorous process of designing and refining these prompts, as well as the significant performance improvements observed in our experiments, underscore the originality and impact of our contribution. In doing so, our research adds a valuable dimension to the ongoing discourse on LLMs and their applications, aligning with the pioneering spirit of ICML's research community.

We believe these points demonstrate that our work goes beyond combining existing techniques and contributes novel insights and methods to the machine learning community, particularly in the context of tabular data generation and handling class imbalance. This contribution is not only novel but also practical and applicable, warranting attention and discussion in venues like ICML. Our work contributes new insights to the academic discussion around LLMs and synthetic data generation, potentially inspiring future research in this significant field.
We believe our unique application of these methods to the domain of tabular data generation, coupled with the notable performance improvements over existing techniques, represents a significant advancement in the field.
The application of our method is broad, demonstrating its utility across multiple datasets and scenarios. This versatility further underscores the innovative nature of our approach.

[1] Shao, Zhihong, et al. "Synthetic prompting: Generating chain-of-thought demonstrations for large language models." International Conference on Machine Learning (2023).
[2] Shrivastava, Disha, Hugo Larochelle, and Daniel Tarlow. "Repository-level prompt generation for large language models of code." International Conference on Machine Learning (2023).
[3] Allingham, James Urquhart, et al. "A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models." International Conference on Machine Learning. (PMLR, 2023).
[4] Sun, Tianxiang, et al. "Black-box tuning for language-model-as-a-service." International

Conference on Machine Learning (2022).
[5] Zhang, Biao, Barry Haddow, and Alexandra Birch. "Prompting large language model for machine translation: A case study." International Conference on Machine Learning (2023).

Response to Reviewer bNDz

We sincerely appreciate the time and effort you dedicated to reviewing our paper, and we are grateful for the valuable feedback you have provided.

Q1. In addition the F1/MSE gains are rather small and it is not clear if they are statistically significant. In addition, the performance degrades with more samples (see Table 4). This shows the inherent variance in the process and statistical significance tests are important.

A1. Please see the general response, titled `General Response 3: Statistical significance of the proposed method`.

Q2. I believe the costs of using LLMs are much higher than using other models. Hence cost-benefit analysis is required.

A2. We analyze the training and inference cost and the classification performance using the Income dataset, comparing our method with other baseline models.

The analysis was performed on a single NVIDIA GeForce RTX 3090 GPU, measuring time cost for 100 epochs of training and sampling 1000 data points, following the cost analysis protocol of the GReaT original paper.
For the F1 score and balanced accuracy metrics, we report the average values for the four distinct classifiers discussed in `General Response 2`.

Table 2. Cost-benefit analysis on tabular data generation methods on the adult income dataset.

	$TVAE$	$CopulaGAN$	$CTGAN$	$GReaT$	$Ours$
Training time	$00h:01m:27s$	$00h:03m:01s$	$00h:03m:19s$	$07h:10m:23s$	$-$
Inference time	$00h:00m:0.1s$	$00h:00m:0.1s$	$00h:00m:0.1s$	$00h:00m:58s$	$00h:02m:20s$
Sum	$00h:01m:27s$	$00h:03m:01s$	$00h:03m:19s$	$07h:11m:21s$	$00h:02m:20s$
F1 score	$0.670_{\pm 0.014}$	$0.667_{\pm 0.017}$	$0.662_{\pm 0.008}$	$\underline{0.679}_{\pm 0.014}$	$\mathbf{0.692}_{\pm 0.010}$
BAL ACC	$0.768_{\pm 0.011}$	$0.767_{\pm 0.013}$	$0.763_{\pm 0.007}$	$\underline{0.775}_{\pm 0.011}$	$\mathbf{0.792}_{\pm 0.008}$

Unlike other methods that require time-consuming training processes, our method involves no additional training time, which is a substantial advantage. This is in stark contrast to methods like GReaT, which demand extensive training, exceeding 7 hours in our experiments. This difference becomes even more crucial in scenarios involving large datasets or the need for frequent retraining due to evolving data.

Our approach's independence from retraining offers substantial benefits in scalability and resource efficiency. This is particularly relevant when multiple hyperparameter tunings are necessary, a process that can linearly increase training time in traditional methods. In contrast, our method remains efficient and scalable.

For the total time required for training and inference to generate 1000 samples, our method is on par with CTGAN and CopulaGAN, taking approximately 2 minutes and 20 seconds.

Importantly, the data generated by our method enhanced downstream task performance, achieving the highest F1 score and balanced accuracy compared to other methods. This underscores the practical effectiveness and high quality of our synthetic data.

Beyond GPT-3.5, our method has proven effective with open-source LLMs, such as Mistral and Llama2 (Please refer to Q1&A1 in `Response to Reviewer yJwy`). This flexibility extends its applicability, making it accessible for a broader range of users and use cases.

In conclusion, our method offers substantial savings in trainig time and resources compared to other methods, making it a practical, economically viable solution for generating high-quality tabular data.

Q3. The datasets used are in public domain and it is very likely that the complete underlying data has been seen by the LLM (GPT 3.5). Experiments on dataset not available publicly (yet) will be helpful.

A3. We expanded our evaluation to include datasets released after the training cut-off date for the GPT-3.5-turbo [1]. This ensures that our method's efficacy is tested on completely unseen data, thereby providing a more rigorous validation of our approach.

Based on that the model was trained on data up to September 2021 [1], we selected two datasets released in 2022, namely the Thyroid [2] and Salary [3] datasets, for classification and regression, respectively.

Table 3-1. Balanced accuracy and mean absolute error

data	*model*	*Original*	*+TVAE*	*+CopulaGAN*	*+CTGAN*	*+GReaT*	*+Ours*
Thyroid	*GB*	$\underline{0.955}_{\pm 0.000}$	$0.922_{\pm 0.000}$	$0.877_{\pm 0.000}$	$0.886_{\pm 0.000}$	$0.913_{\pm 0.012}$	$\mathbf{0.967}_{\pm 0.000}$
	XGBoost	$\underline{0.967}_{\pm 0.000}$	$0.945_{\pm 0.000}$	$0.932_{\pm 0.000}$	$0.773_{\pm 0.000}$	$0.932_{\pm 0.000}$	$\mathbf{0.967}_{\pm 0.000}$
	CatBoost	$0.927_{\pm 0.010}$	$0.899_{\pm 0.000}$	$0.886_{\pm 0.000}$	$0.791_{\pm 0.010}$	$\underline{0.932}_{\pm 0.000}$	$\mathbf{0.963}_{\pm 0.010}$
	LightGBM	$\underline{0.955}_{\pm 0.000}$	$0.922_{\pm 0.000}$	$0.854_{\pm 0.000}$	$0.795_{\pm 0.000}$	$0.922_{\pm 0.000}$	$\mathbf{0.958}_{\pm 0.000}$
	Mean	$\underline{0.951}_{\pm 0.016}$	$0.922_{\pm 0.016}$	$0.887_{\pm 0.029}$	$0.811_{\pm 0.046}$	$0.925_{\pm 0.010}$	$\mathbf{0.964}_{\pm 0.006}$
salary	*RF*	$3.797_{\pm 0.012}$	$3.805_{\pm 0.011}$	$3.851_{\pm 0.013}$	$3.789_{\pm 0.011}$	$\underline{3.737}_{\pm 0.017}$	$\mathbf{3.712}_{\pm 0.017}$
$\times 1e-4$	*XGBoost*	$3.721_{\pm 0.000}$	$\underline{3.689}_{\pm 0.000}$	$3.813_{\pm 0.000}$	$3.777_{\pm 0.000}$	$3.690_{\pm 0.000}$	$\mathbf{3.685}_{\pm 0.000}$
	CatBoost	$3.800_{\pm 0.002}$	$3.771_{\pm 0.002}$	$3.882_{\pm 0.002}$	$3.869_{\pm 0.002}$	$\underline{3.758}_{\pm 0.002}$	$\mathbf{3.734}_{\pm 0.003}$
	LightGBM	$3.724_{\pm 0.000}$	$3.736_{\pm 0.000}$	$3.767_{\pm 0.000}$	$3.746_{\pm 0.000}$	$\underline{3.671}_{\pm 0.000}$	$\mathbf{3.613}_{\pm 0.000}$
	Mean	$3.761_{\pm 0.040}$	$3.750_{\pm 0.045}$	$3.828_{\pm 0.044}$	$3.795_{\pm 0.047}$	$\underline{3.714}_{\pm 0.037}$	$\mathbf{3.686}_{\pm 0.047}$

Table 3-2. F1 score and mean absolute percentage error

data	*model*	*Original*	*+ TVAE*	*+ CopulaGAN*	*+ CTGAN*	*+ GReaT*	*+ Ours*
Thyroid	*GB*	$\underline{0.952}_{\pm 0.000}$	$0.905_{\pm 0.000}$	$0.850_{\pm 0.000}$	$0.872_{\pm 0.000}$	$0.894_{\pm 0.015}$	$\mathbf{0.955}_{\pm 0.000}$
	XGBoost	$\mathbf{0.955}_{\pm 0.000}$	$0.930_{\pm 0.000}$	$0.927_{\pm 0.000}$	$0.706_{\pm 0.000}$	$0.927_{\pm 0.000}$	$\mathbf{0.955}_{\pm 0.000}$
	CatBoost	$0.910_{\pm 0.011}$	$0.878_{\pm 0.000}$	$0.872_{\pm 0.000}$	$0.735_{\pm 0.017}$	$\underline{0.927}_{\pm 0.000}$	$\mathbf{0.950}_{\pm 0.011}$
	LightGBM	$\mathbf{0.952}_{\pm 0.000}$	$0.905_{\pm 0.000}$	$0.821_{\pm 0.000}$	$0.743_{\pm 0.000}$	$0.905_{\pm 0.000}$	$\underline{0.933}_{\pm 0.000}$
	Mean	$\underline{0.942}_{\pm 0.020}$	$0.904_{\pm 0.019}$	$0.867_{\pm 0.040}$	$0.764_{\pm 0.066}$	$0.913_{\pm 0.016}$	$\mathbf{0.948}_{\pm 0.010}$
salary	*RF*	$0.391_{\pm 0.003}$	$0.381_{\pm 0.004}$	$0.414_{\pm 0.001}$	$\underline{0.379}_{\pm 0.004}$	$0.384_{\pm 0.004}$	$\mathbf{0.369}_{\pm 0.003}$
	XGBoost	$0.412_{\pm 0.000}$	$\mathbf{0.396}_{\pm 0.000}$	$0.425_{\pm 0.000}$	$0.408_{\pm 0.000}$	$0.407_{\pm 0.000}$	$\underline{0.402}_{\pm 0.000}$
	CatBoost	$0.433_{\pm 0.001}$	$\mathbf{0.414}_{\pm 0.001}$	$0.439_{\pm 0.001}$	$0.431_{\pm 0.000}$	$0.426_{\pm 0.000}$	$\mathbf{0.414}_{\pm 0.000}$
	LightGBM	$0.393_{\pm 0.000}$	$0.388_{\pm 0.000}$	$0.407_{\pm 0.000}$	$0.400_{\pm 0.000}$	$0.391_{\pm 0.000}$	$\mathbf{0.373}_{\pm 0.000}$
	Mean	$0.407_{\pm 0.017}$	$\underline{0.395}_{\pm 0.013}$	$0.421_{\pm 0.013}$	$0.404_{\pm 0.019}$	$0.402_{\pm 0.017}$	$\mathbf{0.390}_{\pm 0.020}$

As shown in the above tables, our method significantly improves the classification and regression performance on both of the additional datasets. Notably, as Table 3-1 reveals, our method achieved the highest balanced accuracy compared to the baselines with a large margins.

[1] https://community.openai.com/t/gpt-3-5-4-model-documentation-nov-7-2023-still-has-inaccuracies-about-snapshots/462172 (https://community.openai.com/t/gpt-3-5-4-model-documentation-nov-7-2023-still-has-inaccuracies-about-snapshots/462172)

[2] https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence (https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence)

[3] https://huggingface.co/datasets/Einstellung/demo-salaries (https://huggingface.co/datasets/Einstellung/demo-salaries)

Response to Reviewer yJwy

We express our deep gratitude for the dedicated time and effort you spent on reviewing our paper. We would like to respond to your comments and hope that we address all your concerns below:

Q1. The paper proposes a model that relies on GPT3.5 and fails to provide any code. The proposed methods might make it hard for researchers to reproduce results. The paper might need to include some open source LLM such as LLAMA2 and Mistral as base model instead.

A1. Addressing your concerns regarding reproducibility, we are pleased to inform that the source code has been made publicly available on an anonymous GitHub page: source code (https://anonymous.4open.science/r/ST-Prompt-29F2/README.md).

This code facilitates the generation of synthetic tabular data as outlined in our methodology, and provides a framework for conducting experiments with various Machine Learning (ML) predictors.

Further, in line with your suggestion, we have extended our research to include experiments on three distinct datasets utilizing open-source Large Language Models (LLMs), specifically Llama2 [1] and Mistral [2].

Table 1. Experimental results using open-source LLMs

Dataset	*Method*	*Model*	N_{syn}	*F1score*	*Sensitivity*	*Specificity*	*BALACC*	*ACC*	*AUC*
Travel	*Original*	–	–	60.00±0.00	60.00±0.00	84.62±0.00	72.31±0.00	77.78±0.00	**89.35**±0.00
	+ *GReaT*	–	+1*K*	59.57±0.00	56.00±0.00	**87.69**±0.00	71.85±0.00	78.89±0.00	86.12±0.00
	+ *TabDDPM*	–	+1*K*	58.33±0.00	56.00±0.00	86.15±0.00	71.08±0.00	77.78±0.00	88.09±0.00
	+Ours	*Mistral*	+1*K*	66.67±0.00	76.00±0.00	80.00±0.00	78.00±0.00	78.89±0.00	88.74±0.11
	+Ours	*Llama2*	+1*K*	67.80±0.00	**80.00**±0.00	78.46±0.00	79.23±0.00	78.89±0.00	87.94±0.00
	+Ours	*GPT3.5*	+1*K*	**70.18**±0.00	**80.00**±0.00	81.54±0.00	**80.77**±0.00	**81.11**±0.00	87.38±0.00

Dataset	*Method*	*Model*	N_{syn}	*F1score*	*Sensitivity*	*Specificity*	*BALACC*	*ACC*	*AUC*
Sick	*Original*	–	–	84.09±0.00	80.43±0.00	99.28±0.00	89.86±0.00	98.12±0.00	99.64±0.00
	+ *GReaT*	–	+1*K*	85.06±0.00	80.43±0.00	99.43±0.00	89.93±0.00	98.25±0.00	99.66±0.00
	+ *TabDDPM*	–	+1*K*	84.71±0.00	78.26±0.00	99.57±0.00	88.92±0.00	98.25±0.00	99.60±0.00
	+Ours	*Mistral*	+1*K*	88.42±0.00	**91.30**±0.00	99.00±0.00	**95.15**±0.00	98.52±0.00	99.68±0.00
	+Ours	*Llama2*	+1*K*	85.53±0.95	87.39±0.97	98.88±0.06	93.14±0.52	98.17±0.12	99.22±0.03
	+Ours	*GPT3.5*	+1*K*	**91.95**±0.00	86.96±0.00	**99.86**±0.00	93.41±0.00	**99.06**±0.00	**99.71**±0.00

Dataset	*Method*	*Model*	N_{syn}	*F1score*	*Sensitivity*	*Specificity*	*BALACC*	*ACC*	*AUC*
HELOC	*Original*	–	–	71.32±0.04	68.47±0.06	78.31±0.00	73.39±0.03	73.59±0.03	80.61±0.01
	+ *GReaT*	–	+1*K*	70.59±0.03	66.50±0.05	**79.81**±0.05	73.16±0.02	73.43±0.02	80.79±0.01
	+ *TabDDPM*	–	+1*K*	70.65±0.00	67.48±0.00	78.31±0.00	72.89±0.00	73.11±0.00	80.64±0.00
	+Ours	*Mistral*	+1*K*	71.48±0.00	68.00±0.00	79.47±0.00	**73.74**±0.00	**73.97**±0.00	**80.83**±0.01
	+Ours	*Llama2*	+1*K*	70.77±0.00	67.37±0.00	78.79±0.00	73.08±0.00	73.32±0.00	80.74±0.00
	+Ours	*GPT3.5*	+1*K*	**71.93**±0.02	**69.90**±0.00	77.45±0.04	73.68±0.02	73.83±0.02	80.72±0.02

Our results demonstrate that our method exhibits robust performance when applied with open-source LLMs across diverse datasets. Notably, in the Sick dataset, the Mistral model significantly outperforms all other methods inlcuding our method with GPT-3.5 in terms of balanced accuracy (BAL ACC). These results validate the broader applicability and general effectiveness of our methodology when applied with open-source LLMs.
We have successfully tested and validated our approach using open-source alternatives such as Mistral and Llama2. This demonstrates the flexibility of our method, ensuring it remains accessible and viable even for users without access to high-end cloud services.

[1] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
[2] Jiang, Albert Q., et al. "Mistral 7B." arXiv preprint arXiv:2310.06825 (2023).

Q2-1. The experiments only test gradient boosting gradient classifier for classification and random forest regression for regression. Moreover, the classifier is tested on five out of all six datasets. The random forest is tested on two out of all six datasets. Compared to the previous paper, GREAT, this experiment is not very complete.

A2-1. Please see the general response, titled `General Response 3: Statistical significance of the proposed method`.

Q 2-2. The gains of some metrics are really very small. The paper probably needs to add a significance test for those data pairs (e.g., 0.271 vs 0.274).

A2-2. Please see the general response, titled `General Response 3: Statistical significance of the proposed method`.

Q3. The proposed method is kind of incremental. The few-shot method and in-context learning have been widely applied to LLM.

A3. Please see the general response, titled `General Response 4: Novelty`.

Q4. Why the standard deviation for Table 2 is zero?

A4. Thank you for pointing out the oversight regarding the standard deviation in Table 2. We acknowledge that this was an error on our part. The actual standard deviations for the metrics reported in Table 2 fall within the range of 0.01 to 0.06. We have updated the table to accurately reflect these values. The complete results, including the corrected standard deviations, can be found here: Complete Results with Standard Deviation Values (https://anonymous.4open.science/r/ST-Prompt-29F2/Table_complete_results_standard_deviation_ST-Prompt_ICML24.pdf). We truly appreciate your attention to detail and thank you for helping us improve the accuracy of our paper.

Response to Reviewer poPD

We are truly grateful for your thorough analysis and constructive comments, and we hope that we address all your concerns below:

Q1. The contribution is limited, as the paper mainly combines different tricks in prompt engineering.

A1. Please see the general response, titled `General Response 4: Novelty`.

Q2. More recent baselines need to be included, such as [1,2].

A2. Please see the general response, titled `General Response 1: Comparison with aditional baseline models`.

Q3. There's no example for the final prompt design.

A3. We have included prompt examples for the datasets we used in our manuscript:

Prompt example for the Travel dataset (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_TRAVEL_ST-Prompt_ICML24.jpg)
Prompt example for the Sick dataset (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_Sick_ST-Prompt_ICML_24.jpg)
Prompt example for the HELOC dataset (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_HELOC_ST-Prompt_ICML24.jpg)
Prompt example for the Adult Income dataset (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_Adult_Income_ST-Prompt_ICML24_.jpg)
Prompt example for the Diabetes dataset (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_Diabetes_ST-Prompt_ICML24.jpg)
Prompt example for the California housing dataset (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_California_housing_ST-Prompt_ICML24.jpg)

Additionally, we have made our code publicly available, and this repository includes prompt example as well: source code (http://anonymous.4open.science/r/ST-Prompt-29F2).

Q4. The writing could be improved. Some key aspects are not well-explained. For example, in the random-word replacement strategy, from what is described in the paper, the characters in "t,t,t,t" might be replaced with any random characters, and might lead to a result such as "t,a,b,t". Then in this case, while "t" means "true", "a" and "b" do not have any real meaning, which I think might confuse LLM. Its ability to lead to a "more accurate generation" is questionable without real examples and deeper analysis.

A4. We would like to clarify our approach with the random word replacement strategy first. Essentially, this strategy is applied to categorical variables, where each unique value of a categorical variable is replaced with a unique alphanumeric string. This ensures uniform substitution across the dataset, maintaining the integrity of the data structure while introducing variation, as illustrated in this Figure (https://anonymous.4open.science/r/ST-Prompt-29F2/figure/fig_randomword_ST-Prompt_ICML24.jpg).

Table 4-1. Random word mapper

$Variable$	org	new
$V1$	t	$JY0$
	f	GGN
$V2$	t	$E3R$
	f	$L2J$
$V3$	t	CLC
	f	ZWI

To illustrate, consider variable V1 in our dataset, which has two unique values, 't' and 'f'. In our strategy, these values are replaced by distinct 3-character alphanumeric strings, such as 'JY0' for 't' and 'GGN' for 'f', as detailed in Table 4-1. This mapping is consistent across the dataset, meaning every instance of 't' in V1 is replaced with 'JY0', and every 'f' with 'GGN'. Similar mappings are applied to other categorical variables, as demonstrated in the Table 4-1. Even though the new alphanumeric values like 'JY0' or 'GGN' don't carry inherent meaning, they effectively represent categorical distinctions as symbols, e.x., `1 = t = true = yes = y = JY0` and `0 = f = false = no = n = GGN`.

Table 4-2. Before. Original tabular data.

idx	$V1$	$V2$	$V3$
1	t	t	f
2	t	t	t
3	f	f	f

Table 4-3. After. The tabular Data transformed by the random word replacement strategy.

idx	$V1$	$V2$	$V3$
1	$JY0$	$E3R$	ZWI
2	$JY0$	$E3R$	CLC
3	GGN	$L2J$	ZWI

This is applied to all data points in the training dataset (Table 4-2). The transformed dataset (Table 4-3), is then used as in-context learning demonstrations.

The benefit of using random alphanumeric strings over simple characters like 't' or 'f' becomes evident in datasets with monotonos values. In such cases, different variables often share the same symbols, as shown in Table 4-2, leading to potential confusion for the LLM. This is particularly problematic when the dataset includes a large number of repetitive variables.

The use of distinct alphanumeric strings for each categorical value helps to mitigate this issue. For example, replacing 't' with 'JY0' and 'f' with 'GGN' ensures that each variable has a unique representation, reducing confusion and improving the model's ability to generate accurate predictions. Even though the specific alphanumeric strings do not hold intrinsic meaning, LLMs remains effective due to its nature as a pattern recognition machine [1] to differentiate between categorical values. The LLM excels in identifying and utilizing the correlations among features within the data.

Table 4-4. Ablation study on the sick dataset.

$Data format$	$Group-wise$	$Randword$	$Input samples$	$Input tokens(\downarrow)$	$Output samples(\uparrow)$	$Success rate(\uparrow)$
$Sentence$	☐	☐	20	3879.90	0.52	52%
$CSV style$	☐	☐	20	1226.31	3.05	48%
$CSV style$	☑	☐	20	1439.89	6.93	99%
$CSV style$	☑	☑	20	2060.75	17.68	95%

This is corroborated by the results in Table 4-4, where we observed an increase in the number of samples generated per inference from 6.93 to 17.68 when using the random word strategy. Moreover, as shown in Table 5 in our main manuscript, this approach also contributed to an increase in the F1 score from 81.55 to 91.95.

In conclusion, our random word replacement approach significantly improves the LLM's performance by providing clear, distinct signals for each category. This approach not only enhances the LLM's ability to generate a greater number of accurate samples but also positively impacts the overall predictive performance, as evidenced by our empirical results.

[1] Mirchandani, Suvir, et al. "Large Language Models as General Pattern Machines." Conference on Robot Learning. PMLR, 2023.

Q5. The paper does not discuss the ratio of the selected characters for the random-word replacement strategy.

A5. It appears there may be a misunderstanding regarding the methodology of our random-word replacement strategy. This strategy doesn't involve selecting characters at a specific ratio. Essentially, this strategy is applied to every categorical variables in a dataset, where each unique value of a categorical variable is replaced with a unique alphanumeric string.

We have applied this approach to both the Sick and Diabetes datasets, with detailed application results and prompt examples are provided here (https://anonymous.4open.science/r/ST-Prompt-29F2/prompt_example/Table_prompt_Sick_ST-Prompt_ICML_24.jpg).

If this explanation does not align with your understanding of the question, please feel free to clarify so we can address any specific concerns you may have about the methodology.

Q6. The authors claim minimal preprocessing is needed for the frameworks. However, if values are directly input into the LLM, how does the model identify the numerical value range and potential categories for categorical values, especially if demonstrations cannot cover all potential values?

A6. While preprocessing can be applied to the demonstrations if desired, our framework allows minimal preprocessing, significantly reducing the need for the complex data manipulations often required in traditional approaches. This approach not only lessens the labor-intensive preprocessing work but also helps in maintaining the integrity of the raw data, similar to the approach used in GReaT [1].

We achieve this through sufficient number of random sampling of data samples for each prompt, as explained in Section 2.3 in our manuscript. Although each individual generation may be based on the limited distribution of the given subset, cumulatively, the iterations enable the LLM to explore the entire range of the training data. This method ensures that the final output generated by the LLM reflects a comprehensive range of numerical and categorical values, without requiring extensive preprocessing.

[1] Borisov, et al. "Language Models are Realistic Tabular Data Generators." The International Conference on Learning Representations (2022).

Q7. The framework enforces the LLM to generate examples from different classes simultaneously to ensure a balanced dataset. How would the framework handle intentionally imbalanced datasets if needed?

A7. Our method is flexible and can be adapted to handle intentionally imbalanced datasets by selective utilization of the generated classes.

When multiple classes are present, our framework aims to prevent overfitting to any single class by providing a balanced set of examples from all classes as few-shot examples. However, if an imbalanced synthetic dataset is required, users can simply choose to use a portion of the generated output, focusing on the desired class or classes in a predetermined ratio.

For instance, as we have demonstrated in Table 3 of our manuscript, we have already implemented a scenario where synthetic data was intentionally used for one class only. This example demonstrates our framework's ability to adapt to specific requirements.
This flexibility allows our framework to be tailored to various data requirements, enabling the generation of both balanced and intentionally imbalanced datasets according to specific research or application needs.

If there are any other scenarios or considerations that we may have not considered, we would greatly appreciate your response.

Q8. How are the few-shot demonstrations selected in each prompt?

A8. For every prompt, a new set of examples is randomly chosen from each class, as stated in Section 2.3 in our manuscript. Uniform random sampling mitigates the risk of introducing bias that might occur if a certain subset of data is repeatedly used. Also, this approach ensures that over multiple iterations, a wide and representative variety of examples from each class is used to generate synthetic data, encompassing the breadth and variability inherent in each class.

Response to Reviewer 86JX

We appreciate the thorough review and the positive comments regarding our paper. We would like to respond to your comments and hope that we address all your concerns below:

Q1. The novelty of the proposed method is limited. Essentially, there is no genuine "learning" process involved, as it primarily relies on some additional prompting to generate the features. As such, ICML may not be the most suitable venue for this work, and some more application-oriented conferences such as ACL and COLM might be more appropriate targets.

A1. Please see the general response, titled `General Response 4: Novelty`.

Q2. Lack of prompt optimization baselines (mentioned in Appendix E) and other tabular data generation baselines (mentioned in [1]).

A2. Please see the general response, titled `General Response 1: Comparison with additional baseline models`.

Q3. The use of a relatively weak classifier. Considering stronger classifiers, such as XGBoost, as the backbone could potentially improve the results.

A3. Please see the general response, titled `General Response 2: Experimental results on additional machine learning models`.

Q4. It is not sure how this group-wise prompting method may increase the bias of the classifier or not, given LLM generated synthetic data may contain bias.

A4. It is true that using synthetic data generated by LLMs can increase the bias of the classifier. However, our anlaysis of the confusion matrix (Figs 1 and 7 in our main manuscript and appendix, respectively) indicates that machine learning classifiers trained with data generated using our approach exhibit equitable performance across different classes. This suggests a reduction in bias compared to datasets generated by other baseline methods. This is achieved by our group-wise prompting method, where we carefully ensure that the groups from which prompts are sampled are balanced.