T-REST: A WATERMARK FOR GPT TABULAR MODEL

Author:
Minh Nguyen
NguyenHoangMinh@student.tudelft.nl

Supervisors:
Prof. L. Chen - Y.Chen-10@tudelft.nl
Chaoyi Zhu - c.zhu-2@tudelft.nl
Jeroen Galja - j.m.galjaard@tudelft.nl

Affiliations:
Delft University of Technology

1 Introduction

Tabular data is one of the most common forms of data in the industry and science. Recent research on synthetic data generation employs auto-regressive generative language models (LMs) to create highly realistic tabular data samples. With the increasing use of LLMs, there is a need to govern the data generated by these models, including watermarking the model output. While the state-of-the-art Soft Red List watermarking framework has shown impressive results on standard GPT-like models, it can not be seamlessly applied to models fine-tuned for generating tabular data due to i) column permutation and ii) the task's nature of generating low entropy sequences.

2 Objective

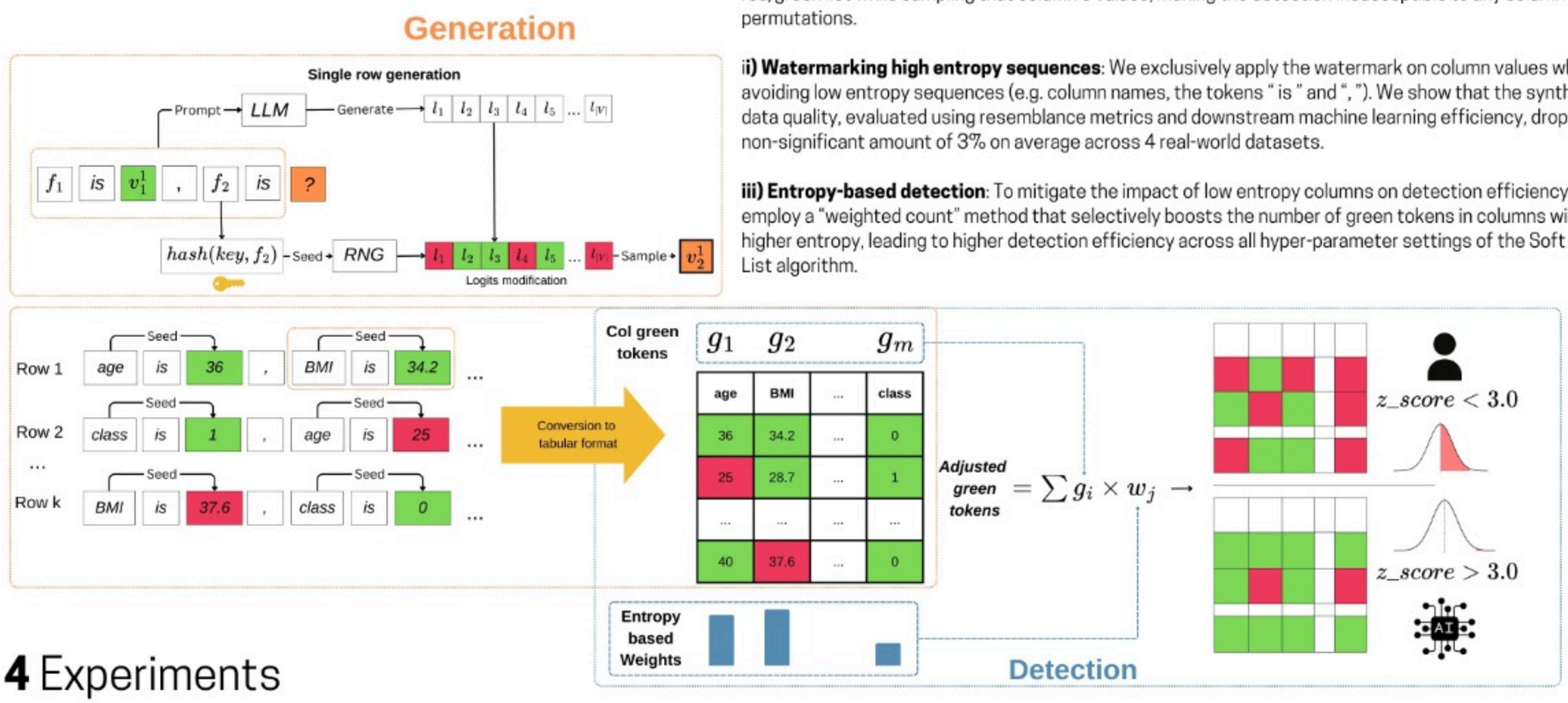
We propose Tabular Red GrEen LiST (T-REST), an adaptation of the Soft Red List watermarking algorithm on tabular LLMs that is agnostic to column permutation and improves detection efficiency by employing a weighted count method that favors columns with higher entropy

References

[1] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators, 2023

[2] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models, 2023.

3 T-REST watermark



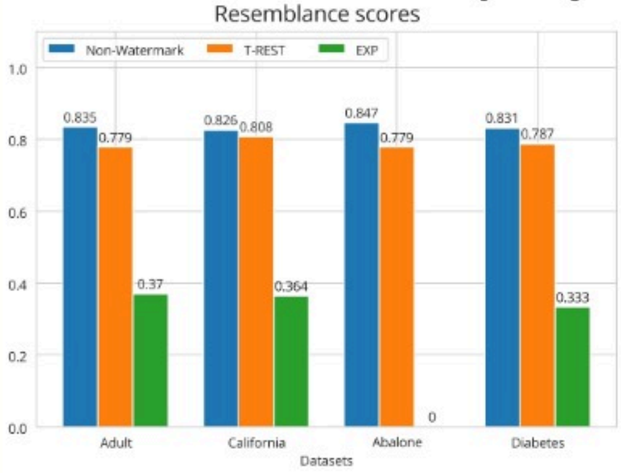
i) Column-based seeding: We use column name and a secret key as the seed for partitioning the red/green list while sampling that column's values, making the detection insusceptible to any column permutations.

ii) Watermarking high entropy sequences: We exclusively apply the watermark on column values while avoiding low entropy sequences (e.g. column names, the tokens " is " and ", "). We show that the synthetic data quality, evaluated using resemblance metrics and downstream machine learning efficiency, drops by a non-significant amount of 3% on average across 4 real-world datasets.

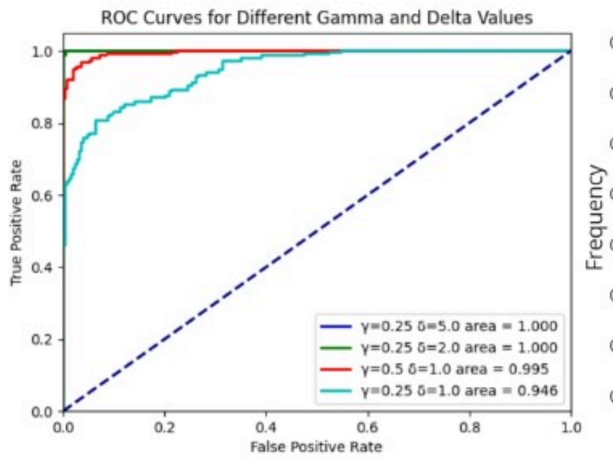
iii) Entropy-based detection: To mitigate the impact of low entropy columns on detection efficiency, we employ a "weighted count" method that selectively boosts the number of green tokens in columns with higher entropy, leading to higher detection efficiency across all hyper-parameter settings of the Soft Red List algorithm.

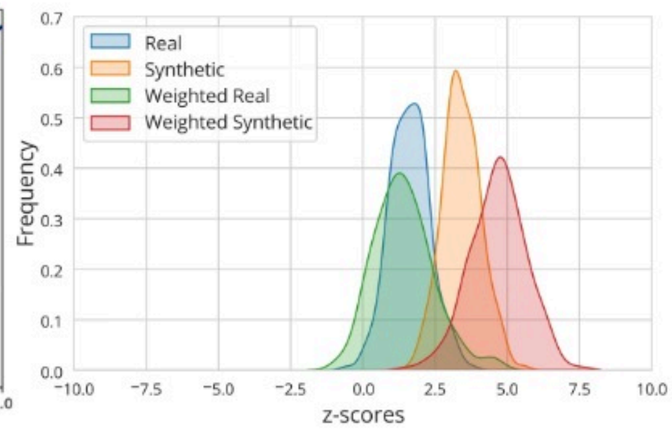
4 Experiments

Minimal distortion to data quality



Improved detection efficiency with entropy-based detection





6 Limitations and Future work

T-REST watermark shows limited robustness against post-editing attacks on numerical columns due to the fact that a reasonably small change to a numerical value can significantly modify its textual representation, resulting in drastically different tokens. We suggest mitigating this issue by employing a non-token based watermark method for numerical columns. Further research is needed to explore the applicability of W-SRL on other tabular LLMs than GReaT

7 Conclusion

In this paper, we identify two major challenges when adapting the Soft Red List watermark framework on GPT-like tabular models, namely column permutation and low entropy sections and columns. We propose Tabular Red GrEen LiST (T-REST), an adaption of the Soft Red List that is agnostic to any column permutations by using column names as seed for partitioning the Green/Red lists. To mitigate the impact of low entropy sections and columns, we exclusively embed the watermark on column values and employ an entropy-based detection method that favors green tokens from columns with relatively higher entropy values.