Qeios

Peer Review

Review of: "SafeSynthDP: Leveraging Large Language Models for Privacy-Preserving Synthetic Data Generation Using Differential Privacy"

Karima Makhlouf[1]

1. Ecole Polytechnique, France

Review of: SafeSynthDP: Leveraging LargeLanguage Models for Privacy-PreservingSynthetic Data Generation UsingDifferential Privacy

The paper presents an innovative approach to privacy-preserving synthetic data generation using LLMs and Differential Privacy, demonstrating the feasibility of balancing privacy with data utility. More specifically, the proposed framework, SafeSynthDP, integrates Laplace and Gaussian noise mechanisms into the data generation process within the LLM pipeline to enhance privacy while preserving data utility. This represents a relatively novel approach, as most prior work has focused on DP in model training rather than data generation.

One of the key strengths of this work is its experimental evaluation across multiple ML architectures, including traditional models (MNB, SVM) and deep learning models (GRU, LSTM), making the findings more generalizable. Additionally, the study's exploration of LLM-driven classification using synthetic data is an interesting contribution, particularly in the context of zero-shot and few-shot learning.

The authors effectively acknowledge the limitations of their work, particularly the lack of evaluation beyond AGNews—the framework has not yet been tested with real-world datasets. Additionally, the study lacks a comparative analysis with other DP-based synthetic data techniques, such as DP-SGD (Differentially Private Stochastic Gradient Descent) or DP-based GANs, which would help contextualize SafeSynthDP's performance relative to existing methods.

doi.org/10.32388/609SBX

Furthermore, the Related Work section could better situate SafeSynthDP within existing research. While extensive, this section does not clearly differentiate SafeSynthDP from prior approaches in LLM-driven DP synthetic data generation. Providing a more structured comparison with previous methodologies would strengthen the paper's contribution.

Some recommendations for improvement:

1- Enhance the visual representation of results by adding plots that illustrate the privacy-utility tradeoff across ICL approaches. This would provide a clearer pattern and behavior analysis of the tradeoff.

2- Maintain consistency with acronyms (e.g., ML for machine learning) throughout the paper to improve readability.

3- Minor Correction: Page 10, last line of Section 4.1 → Remove the "s" from "*trainings*" (should be "training").

Overall, this paper provides a strong foundation for integrating Differential Privacy within LLM-driven synthetic data generation, but addressing the above points would further strengthen its impact and clarity.

Declarations

Potential competing interests: No potential competing interests to declare.