Qeios

Peer Review

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

**Andrey Makrushin**[1]

1. Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany

The authors propose to apply the differential privacy (DP) concept to pre-trained large language models (LLM) when they are leveraged for data generation (LLM-driven data generation). It is stated that the resulting synthetic data are privacy-friendly and also retain utility for training further classification models. The trade-off between the privacy budget and data utility is estimated empirically.

The paper is generally well written and well structured. It is easy to understand and to follow the argumentation. The language is perfect. The state-of-the-art studies are also well addressed.

My main point of criticism is the missing technical part of the proposed "novel framework," i.e., it is not formalized how noise is added to prompts. In my view, it is not enough to write: "For the Laplace mechanism, we adjust the counts of common words or phrases, altering their likelihood in the synthetic text. Similarly, Gaussian noise is applied with a distribution that helps smooth the impact on the text's features." In a scientific paper, I would expect a more formal mathematical description using equations. In fact, noise addition to prompts is the key idea of the paper. The missing technical description does not allow us to assess both the technical correctness of the proposed concept and its novelty. Hence, the reproducibility of the research results is questionable.

The paper would definitely benefit from publishing the source code implementing the noise addition and the AGNews subsets used in evaluation.

Some minor criticisms:



doi.org/10.32388/V4M4AE

- The importance of conducting privacy-preserving ML/ICL is mentioned too often and reads like engaging in empty talk.

- I see no argumentation as to whether 12k training samples and 4k test samples are enough for statistically significant reasoning.

- It is confusing that the authors write "Laplace/Gaussian noise" throughout the paper, but the experiments seem to include the Gaussian noise only, as I understood from Section 5.5.

Suggestions:

- Move Section 4.2 to Methodology.

- In Section 4.3, visualize a pipeline of preparing training data (original vs. synthetic) for ML/ICL, including both types of features, TF-IDF and GloVe.

- For the sake of correctness, the evaluations in Sections 5.1 and 5.2 require mentioning the privacy budget used.

- Are the results in Sections 5.1, 5.2, and 5.3 reported for Laplace or for Gaussian noise?

- Section 5.4 belongs rather to the Methodology chapter.

All in all, the scientific value of the paper is very limited. The method is not described in enough detail to fully understand its superiority over competing methods. The experimental results with the AGNews dataset do not allow general conclusions about the applicability of the "token frequency manipulation" as a DP mechanism to other types of data.

## Declarations

**Potential competing interests:** No potential competing interests to declare.