Evaluating the Efficacy of Differentially Private Stochastic Gradient Descent implementations for Pre-training Large Language Models

April 27, 2025

1 Introduction

1.1 background

Large Language Models (LLMs), built upon the transformer network architecture [11], have gained immense popularity in recent years, holding promise of greatly revolutionizing both academia and industry, increasing efficiency, replacing menial tasks, even supplementing the innovation and ideation traditionally limited to the most creative and ingenious human intellect. The immense applicability and generalizability of LLMs can be accredited to its vast swaths of training data. In fact, Chinchilla Scaling laws even quantify the fact that there exist predictable, measureable improvement in model generations as training data quantity increases [4]. So as to gather and utilize immense corpuses of text-based data for training models, companies turn towards scraping the internet for human-generated text. However, as the quantity of text increases, it quickly becomes intractable to fully screen and vet all text used to train the LLMs.

Among data scraped from the internet, there are undoutedly samples problematic statements, and of potentially sensitive information. The former has been combatted with fine-tunning; after a model is pre-trained on vast swathes of internet text, specific frameworks are introduced to fine-tune the LLM, tweaking model weights to remove unfavorable, incoherent, and problematic generations from the LLMs lexicon [9]. However, fine-tuning only serves to suppress such problematic generations, and cannot guarantee removal, with prompt hacking methodologies demonstrating that it is trivially feasible to coerce language models into generating outputs which are intentionally suppresed by its developers [10]. Such methodolgies could be equally applied to exposing private and sensitive information inadvertently incorperated into the LLMs training data.

Unlearning–the process of removing the influence of undesirable information from an LLM without impacting unrelated information–has been explored as a potential solution to this dilemma [5]. However, unlearning primarily targets known, identified instances of undesirable information. Sensitive information that is not discovered and unlearned can remain indefinitely. This, in conjunction with the prevalence of LLMs memorizing and regurgitating text found in training data verbatim [3] thus poses a severe security risk and privacy violation [7].

As such, it becomes critical to identifying methodologies by which it becomes possible to restrict the influence of undesirable information at training time, without external classification or detection of undesirability. Thus enters differential privacy, a field of study which not only proposes rigorous and quantitative definitions of privacy, but also provable guarantees for privacy preservation, making it a widely accepted gold standard for the purposes of statistical analysis and release of datasets [2]. Treating the training of an LLM as a form of data aggregation and release then, it becomes possible to utilize differential privacy to design and deliver guarantees which bound the extent to which any particular instance of undesirable information can influence generations elicited from the model.

1.2 Differentially Private Stochastic Gradient Descent

In the context of LLMs, model parameters are adjusted based on training data to minimze a loss function which represents conformity to the training data. The most commonly used mechanism of minimizing the loss function is Stochastic Gradient Descent (SGD). As such, SGD can be treated as the process by which data and private information is used to influence the model parameters, which are ultimately used in inference and the generation of output, making it a primary target for differentially private frameworks, which would enable developers and researchers to bound the influence of any particular training example on the overall model weights. Thus is differentially private SGD (DP-SGD), first proposed by Abadi et al. in 2016 [1].

INSERT MATHS HERE

Thus, DP-SGD, while a powerful technique capable of creating provable and quantifiable privacy gurantees on the influence of training data on machine learning models, is not without drawbacks. In particular, the process of gradient clipping, and the addition of random noise not only increases computational overhead in an already computationally demanding workflow, but also degrades the quality of the resulting model, if given the same hyperparameters [8]. In particular, previous research has shown that to mitigate the degredation of generations caused by DP-SGD, and reach the same level of performance, it becomes necessary to train the model for more iterations, resulting in an increase in computational costs [6].

In this work, we aim to quantify the extent to which DP-SGD impacts model training and performance, by examining and comparing various implementations of DP-SGD with a non-DP SGD optimizer. So as to examine the impact of DP-SGD in a controlled environment, we opt to train a simple transformer model implementation from scratch, utilizing toy training data examples to evaluate the model across different circumstances.

2 Methodology

TODO

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[2] Ferdinando Fioretto, Pascal Van Hentenryck, and Juba Ziani. Differential privacy overview and fundamental techniques, 2024.

[3] Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. Sok: Memorization in general-purpose large language models, 2023.

[4] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[5] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models, 2024.

[6] Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

[7] Tarun Ram Menta, Susmit Agrawal, and Chirag Agarwal. Analyzing memorization in large language models through the lens of model attribution, 2025.

[8] Nicolas Papernot, Martin Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

[9] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities, 2024.

[10] Baha Rababah, Shang, Wu, Matthew Kwiatkowski, Carson Leung, and Cuneyt Gurcan Akcora. Sok: Prompt hacking of large language models, 2024.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.