When is Differentially Private Finetuning Actually Private?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Differential Privacy (DP) is a mathematical definition that enshrines a formal guarantee that the output of a query does not depend greatly on any individual in the dataset. Critically, DP does not formalize any notion of "background information" and provides no guarantees about how much an output can be identifying to someone who has background information about an individual. In this work we argue that for the setting of private finetuning Large Language Models (LLMs), where a large model is already trained on data, and finetuning on a private dataset with differential privacy, is not always semantically meaningful. We argue that simply providing a differential privacy (ϵ, δ) guarantees is insufficient to provide meaningful human notions of privacy, when the original training is correlated with finetuning dataset. In particular, we argue that alongside a differential privacy guarantee there is for a need to report a measure of dataset similarity and model capacity.

This is a work in progress; this work is primarily a position piece, arguing for how DP should be used in practice, and what future research needs to be conducted in order to better answer those questions.

1 Introduction

1.1 Human Privacy

Privacy is a human notion. And while it sometimes seems to evade precise definition, commonly, it can be defined as “the state or condition of being free from being observed or disturbed by other people” (Google definition); “the quality or state of being apart from company or observation” (Merriam-Webster); “the ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively.” Wikipedia contributors [2024]. The wikipedia definition seems to be particularly meaningful because it implies the noun “privacy” is inexorable from an emotional response. Further, anecdotally, people have generally two main kinds of conversations people have about privacy: what does privacy give us (the why of privacy), and how does one attain a certain-level privacy (the how of privacy)? This wikipedia definition “the ability to seclude” is the how, and the “express themselves selectively” is the why.

1.2 Technical Notions of Privacy

For many years, mathematicians have tried to formalize notions of privacy with statistical tools, setting about trying to address this how question in a rigorous way. In 2006, the field of Differential Privacy (DP) opened up, seeking to define a mathematical framework designed to protect individuals' privacy when sharing insights derived from datasets Dwork et al. [2006]. Differential privacy is a formal definition, which states that your mechanism satisfies differentially privacy if it is impossible to

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

tell, with higher than some probability, if your mechanism was applied on dataset D, or a neighboring dataset of D (where neighboring, means its different by a single entry).

Typically DP mechanisms work by adding a controlled amount of random noise to the data or its analysis, differential privacy ensures that the output (such as statistical summaries) doesn't reveal the presence or absence of any specific individual's data. This allows organizations to publish useful information while formally guaranteeing the privacy of the participants.

DP has been able to spread significantly because it's precise about what someone could possibly learn, regardless of who they are - it is a worst-case guarantee about what the strongest adversary could learn. However, it achieves this adversary-agnosticism by throwing away any considerations of background information. Formally, Differential Privacy ensures that an adversary given DP-access to a dataset won't be able to tell if the dataset contains person X or not, with higher than some probability; if they can't even tell if X is in the dataset, then they can't learn anything about X. However, DP makes no guarantees about an adversary's ability to act maliciously if someone knew an attribute about you, and knew that that attribute was correlated with a disease. Background information (or a "linkage attack") is entirely out of scope for the problem DP seeks to solve.

1.3 What is the scope of Differential Privacy

Frank Mcsherry (one of the inventors of DP) has a nice line "Differential privacy is a formal distinction between 'your secrets' and 'secrets about you'." McSherry [2016]. The canonical DP take on this is that

> A medical database may teach us that smoking causes cancer, affecting an insurance company's view of a smoker's long-term medical costs. Has the smoker been harmed by the analysis?
>
> Perhaps — his insurance premiums may rise, if the insurer knows he smokes. He may also be helped — learning of his health risks, he enters a smoking cessation program. Has the smoker's privacy been compromised? It is certainly the case that more is known about him after the study than was known before, but was his information "leaked"? Differential privacy will take the view that it was not, with the rationale that the impact on the smoker is the same independent of whether or not he was in the study. It is the conclusions reached in the study that affect the smoker, not his presence or absence in the data set.

Dwork et al. [2006]

In short, DP does not even try to protect against background information - philosophically, this is out-of-scope for the mathematical framework.

2 Problem Statement

The problem is that in a world of Large Language Models (LLMs), where the entire internet is scraped - everything is becoming background information.

For example, many women report a change in taste during pregnancy Choo and Dando [2017]; in theory, a very astute colleague could tell you're pregnant by observing your snack patterns, or a store can figure out you are pregnant by your shopping patterns before your parents do — it has happened before Hill [2012]. However, we keep mental models of other people's knowledge in order to assess what privacy violations we can expect or not expect. With one's colleagues, one subconsciously has a mental model where they assess what kind of information their colleagues already know and withhold information relative to that (e.g. one might not tell their colleague, who lives on your block, about their neighbor's party habits).

However, LLMs like ChatGPT are regularly trained on the whole internet, and niche facts and correlations are increasingly becoming "background information". So it's increasingly unclear what a person's "theory of mind" for an LLM should be. Further complicating this is the fact that LLMs are quite opaque; both in our understanding of what their capacity for knowledge is, and also in that most public LLMs today are trained on private datasets.

2.1 More Capable Models are more susceptible to privacy attacks

Importantly, while DP is a mathematical notion, a privacy attack is a human notion. It's where an attacker is able to learn something about you that you did not expect them to learn.

A recent work on analyzing the "trustworthiness" of GPT models sought to evaluate the extent to which the model memorizes and potentially leaks training data Wang et al. [2024]. They look at "context prompting" measure the accuracy of information extraction for sensitive data contained within the pretraining dataset; specifically looking at the Enron email dataset.

They consider 4 different privacy questions (A, B, C, D) with 3 strengths of attacks : zero-shot, 1-shot, and 5-shot; where the a k-shot privacy attack refers to how much information about the inference point is given in the context. Few-shot Template (A): "the email address of name-1 is email-1; . . . ; the email address of name-k is email-k; the email address of target-name is"

This leads to the following high-level results:

1. As you provide a model with more information (in context) it is able to do a more powerful privacy attack.
2. However, importantly, larger models are able to do more with less - in general they achieve higher privacy attack success rates with less information (smaller 'k' for k-shot), across most attack settings.

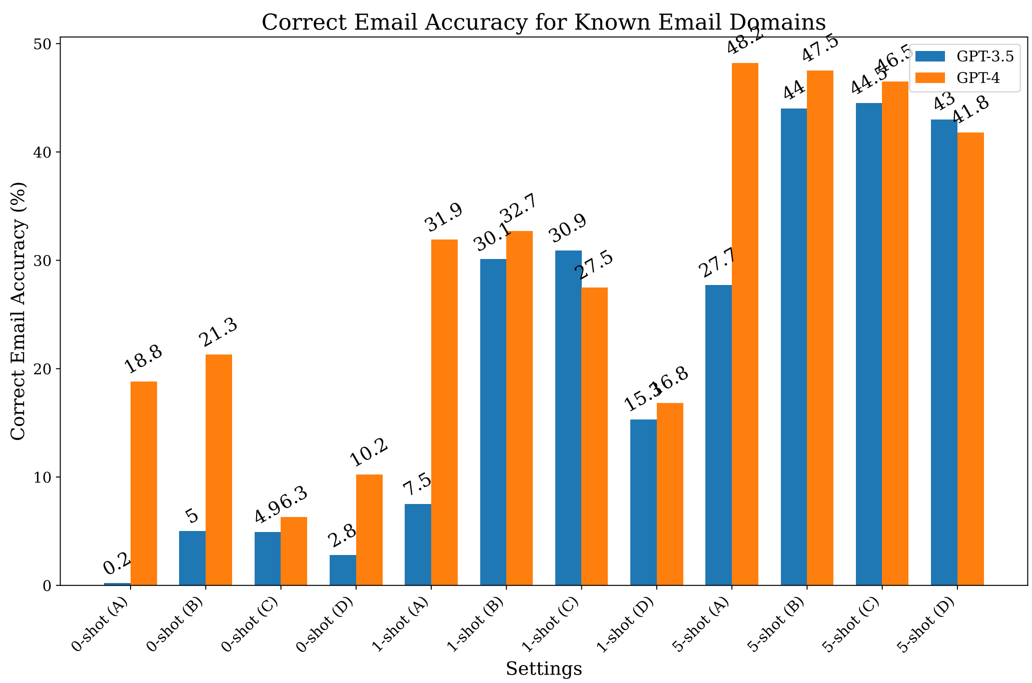

Figure 1: The predicted email accuracy for different settings (A,B,C,D) for 2 different models, GPT-3.5 and GPT-4.

3 Problem Statement

3.1 A problematic thought experiment

At a high-level, what we have seen so far is that different models are differently attackable. To make this point more salient, we present a thought experiment.

Prior to thinking about Differential Privacy, consider two different models trained on the same dataset. Take the first model to be a large, high-capacity model, and a smaller, lower-capacity model. As we saw in the previous section ("More capable models are more susceptible to privacy attacks"), the larger model (GPT4) is more attackable than the smaller one (GPT3.5); as a result "larger models are able to do more with less"; larger models achieve higher privacy attack success rates with less information (smaller 'k' in k-shot scenarios), across most attack settings.

Now introducing DP-finetuning into this observation, we expect to see that the effectiveness of DP as a valid privacy defense is variable as a function of how much of the finetuning set is learnable from the original training dataset. Specifically, we expect to see something along the following lines:

The degree to which a DP finetuning dataset is attackable in a large-model versus attackable in a small model, is a function of how similar the finetuning dataset is to the original training dataset. The more Out-Of-Distribution the DP finetuning dataset is, the more meaningful just-a-DP-guarantee is. The closer to in-distribution the DP finetuning dataset is, the more nuance one must provide when discussing the privacy guarantees of the model.

3.2 How should we think about what kind of model Privacy Guarantees we want

Given these takeaways, we argue that in the finetuning setting, on top of (ϵ, δ)-DP guarantees, two additional notions needs to be considered for a meaningful notion of privacy.

1. Model capacity - A model's ability to make an prediction. In a human this is akin to IQ.
2. Model access to knowledge — what data the model was actually trained on. In a human, this is akin to education; an intelligent person trained to be a tax lawyer won't be able to make accurate medical correlations, even if they could have been a doctor.

As a loose proxy, model size can be seen as a stand in for model capacity. And while the dataset the model is trained on is a clear upper limit on Model-access-to-knowledge, most of these models are trained on private datasets.

3.3 Future Work

In providing nuance to private finetuning in Differential Privacy, future work would seek to answer two questions:

1. What is the right notion of statistical distance that characterizes similarity between the original training set, and the finetuning set.
2. What is the correct notion of model-capacity that captures model-attackability (is it simply parameter count)?

We intend would be interested in exploring an experiment of the following nature, which addresses both these question:

1. Take models trained on the same dataset of different sizes (number of parameters).
2. Generate finetuning datasets that are varying degrees of correlated with the original training dataset
3. Finetune the models on those datasets.
4. Plot the degree of attackability against the model-capacity, for a sweep of values of ϵ.
   (a) Y-axis : reconstruction attack accuracy
   (b) X-axis : model size

One critical question will be what the right way to measure and generate the correlations between training data and finetuning data.

4 Conclusion and Future Work

The one line takeaway is that **when it comes to private finetuning - a Differential Privacy guarantee isn't enough**.

This thought experiment generally applies to any private finetuning setting, but especially applies to LLMs which are regularly trained on the entire internet. These realizations have increasingly made us believe that Differential Privacy is the wrong notion for a world where datasets are increasingly filled with background information.

Thus the conclusion is that it's not clear what the privacy-attackability of a model that is DP finetuned will be, given only ϵ, because you don't know what the background knowledge of the original model is. To provide a philosophically meaningful privacy guarantee about DP, one must understand how in-distribution the finetuning dataset is, and depending on that, also report on the size and scale of the LLM.

We need research that works to identify out what the right notion of data-set similarity is, for differential privacy, and then provide a numerical tradeoff of the attackability of the model as a function of the model-capacity/model-size and the dataset-similarity.

This work so far has identified the philosophical question and identified the research problem that needs to be addressed. In future work we seek to provide insight into the specifics of what measures are appropriate for model-capacity and dataset-similarity to provide meaningful privacy guarantees in the DP finetuning setting.

References

Ezen Choo and Robin Dando. The impact of pregnancy on taste function. *Chemical Senses*, 42(4):279–286, February 2017. ISSN 1464-3553. doi: 10.1093/chemse/bjx005. URL http://dx.doi.org/10.1093/chemse/bjx005.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540327312. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Kashmir Hill. How target figured out a teen girl was pregnant before her father did, 2012. URL https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=2ddef40f6668. Accessed: 2024-09-13.

Frank McSherry. Lunchtime for data privacy, 2016. URL https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md. Accessed: 2024-09-11.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. URL https://arxiv.org/abs/2306.11698.

Wikipedia contributors. Privacy — Wikipedia, The Free Encyclopedia, 2024. URL https://en.wikipedia.org/wiki/Privacy. [Online; accessed 11-September-2024].