Position: Considerations for Differentially Private Learning with Large-Scale Public Pretraining

Florian Tramèr [1] **Gautam Kamath** [2 3] **Nicholas Carlini** [4]

Abstract

The performance of differentially private machine learning can be boosted significantly by leveraging the transfer learning capabilities of non-private models pretrained on large *public* datasets. We critically review this approach. We primarily question whether the use of large Web-scraped datasets *should* be viewed as differential-privacy-preserving. We further scrutinize whether existing machine learning benchmarks are appropriate for measuring the ability of pretrained models to generalize to sensitive domains. Finally, we observe that reliance on large pretrained models may lose *other* forms of privacy, requiring data to be outsourced to a more compute-powerful third party.

1. Introduction

While machine learning models have made tremendous progress at learning *generalizable concepts* from data at scale, these models also frequently *memorize* parts of their training data (Feldman, 2020; Feldman & Zhang, 2020; Shokri et al., 2017). This poses a threat when the model's training data contains privacy-sensitive information, as deployed models may regurgitate memorized private data (Carlini et al., 2019; 2021; Somepalli et al., 2023).

Differential privacy (DP) (Dwork et al., 2006) offers a formal solution to this problem. Informally, training a model with (user-level) differential privacy offers a *guarantee* that the model will not depend too heavily on the sensitive data contributed by any one individual. Among other threats, this protects against the model *memorizing* training data. But current approaches to differentially private learning scale poorly, and greatly sacrifice the model's useful generalization capabilities in order to provably prevent memorization.

To address this issue, a growing line of work suggests to augment differentially private learning algorithms with access to *public data* (Abadi et al., 2016; Papernot et al., 2019; Tramèr & Boneh, 2021; Yu et al., 2022; Li et al., 2022; Arora & Ré, 2022; De et al., 2022; Mehta et al., 2023; Kurakin et al., 2022; Panda et al., 2022; Nasr et al., 2023b; Tang et al., 2024). The goal is to first use large troves of non-privacy-sensitive data to learn generic features—independent from any data owner's private data—which can then be efficiently *finetuned* with DP on sensitive data.

For example, suppose a company wishes to train a model on a corpus of chat messages from its end users. While the content of these messages is sensitive, the general structure of a chat message (i.e., syntax, grammar, etc.) is not sensitive. Thus, the company may wish to leverage a model that was pretrained on a large public corpus of text (preferably including chat conversations) and then finetune this model on the specific sensitive content of the end users' messages.

Even in the the absence of any privacy concerns this *pretraining* and *transfer learning* approach has become the de-facto strategy for achieving state-of-the-art results across a variety of challenging tasks in computer vision and natural language processing. Here, a generic "foundation model" (Bommasani et al., 2021) is first pretrained on massive and weakly curated data—typically scraped from the public Internet. Thereafter, the model can be efficiently finetuned on various downstream tasks (Radford et al., 2019). Generative models such as large language models even exhibit powerful *in-context learning* abilities, where the model "learns" new tasks at inference time solely on the basis of a small number of examples (Brown et al., 2020).

The impressive performance of foundation models naturally places these models as ideal candidates for private learning. Indeed, as the pretraining data comes from publicly available sources, the pretrained model is fully independent of individuals' privacy-sensitive target data. And since these models learn new tasks extremely (sample)-efficiently, they

[*]Authors listed in reverse alphabetical order. A version of this paper, unconstrained by page limits, is available on arXiv (Tramèr et al., 2022). [1]Department of Computer Science, ETH Zürich, Zürich, Switzerland [2]Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada [3]Vector Institute, Toronto, Ontario, Canada [4]Google DeepMind, Mountain View, USA. Correspondence to: Florian Tramèr <florian.tramer@inf.ethz.ch>, Gautam Kamath <g@csail.mit.edu>, Nicholas Carlini <nicholas@carlini.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

should be able to also learn these tasks *privately* with only a minor impact on performance. Thus, it is unsurprising that a growing body of work investigates the benefits of using Web-scale pretraining for private learning (Li et al., 2022; Yu et al., 2022; Arora & Ré, 2022; De et al., 2022; Mehta et al., 2023), and showcases significant improvements in performance on canonical private learning benchmarks.

For example, on the ImageNet dataset (Deng et al., 2009), in the absence of any pretraining, the approach of Sander et al. (2023) achieves a top-1 accuracy of 39.2%, under a fairly weak provable DP guarantee of $\varepsilon = 8$; more recent work of Tang et al. (2023) improved this slightly to the current state of the art: 39.39%. This represents an almost $6\times$ increase in error rate compared to the best non-private model trained solely on ImageNet (at least 86.7% accuracy) (Tu et al., 2022). In contrast, when leveraging a dataset of 4 billion Web images for public pretraining, Berrada et al. (2023) achieve an accuracy of 86.8% at a much more reasonable privacy budget of $\varepsilon = 1$ (with comparable results obtained by De et al. (2022); Mehta et al. (2023)).

In a similar vein, Li et al. (2022) and Yu et al. (2022) finetune large pretrained language models such as RoBERTa (Liu et al., 2019) and GPT-2 (Radford et al., 2019) to achieve strong performance on downstream tasks with differential privacy. Arora & Ré (2022) further argue that LLMs can be personalized to each individual's personal and sensitive data while incurring *no privacy cost* (i.e., $\varepsilon = 0$) by leveraging the pretrained model's "zero-shot" abilities. This line of work suggests we are getting close to "solving" private learning. Indeed, as Web-scraped datasets grow larger and larger, the ability of pretrained models to privately adapt ("for free") to new tasks will only get better.

This position paper challenges this view, and critiques the *public-pretraining and private-finetuning* paradigm. We raise two (orthogonal) concerns that models trained in this manner may fail to be a) *private*, or b) *useful*. We thus question the validity of current findings in this area for informing real-world deployments of differential privacy.

Our primary criticism challenges the notion that pretraining on publicly-available Web data should be viewed as neutral (i.e., non-sensitive) from the perspective of user privacy:

> *Pretraining data scraped from the Web may be sensitive itself; because a "privacy-preserving" finetuned model can still memorize its pretraining data, this causes direct harm and dilutes the meaning of "private learning".*

Specifically, our critique raises issue with the privacy semantics when finetuning data is sensitive, but the pretraining data is considered to be public: as we explore, the latter assumption is mismatched with norms and expectations of what it colloquially means for a model to be private.

Beyond this core concern with the appropriateness of using publicly available data for privacy-preserving learning, we further posit that this paradigm might not be as useful as existing research suggests, and that it could even lead to a net *loss* of privacy at training or deployment time:

- Current private learning benchmarks likely overestimate the value of public pretraining by fixating on settings with highly overlapping public and private data distributions.
- Public pretraining performs best with massively large models that cannot be run on end-user devices, thereby trading off one form of privacy (DP for the sensitive finetuning data) for another (the model's users have to outsource private data to a third party).

Each of the three issues we raise are largely orthogonal to one another, and solving any one of them need not affect the others. For example, even if we were to develop benchmarks that accurately reflect private workloads, the core issue of the potential sensitivity of pretraining data would remain.

This work is a *position paper*, which takes a critical view of the current state of the field and highlights several aspects we find problematic. We thus put forward a call for solutions from the community – while we offer some broad suggestions on potential ways to address our concerns, we (intentionally) stop short of technically exploring solutions, as each of these challenges deserves significant attention beyond the scope of this article.

1.1. Paper Overview

In the reminder of this introduction, we provide a broad overview of all three issues above, and outline some open questions and paths forward for the field. For the interested reader, Sections 2, 3 (and, due to space restrictions, Appendix A) then delve into more details to support our main arguments. Finally, Section 4 provides some concluding remarks and future outlooks.

1. The Web contains privacy-sensitive data. Training data scraped from the Web is indeed *publicly accessible*, but this does not imply that using this data in machine learning applications poses no privacy risks.

Individuals may put some data on the Internet with a specific context-of-use in mind. For example, someone may post their contact information along with a research publication with the intent that it is used to contact that person about details of the publication. Sensitive data about individuals could also be uploaded to the Internet unintentionally (or by third parties privy to this information). As a result, people

often underestimate how much information about them is accessible on the Web (Ellis & Thomas, 2022), and might not consent to their "publicly accessible" personal data being used for training machine learning models.

The question then is whether finetuning (with DP) on top of such publicly pretrained models should really be publicized as *privacy-preserving*. It is entirely possible that models might memorize a large fraction of their (public, yet sensitive) pretraining dataset. Then if the model ever leaks this pretraining data, this still harms the privacy of the data subjects. Such situations could run the risk of eroding affected individuals' trust in differential privacy to appropriately protect their data in other settings (e.g., for collecting census data (Abowd et al., 2022)).

The guarantees of differential privacy are complicated enough to understand when no public data is involved (Cummings et al., 2021), even for researchers (McSherry, 2016a;b). Asking data owners to distinguish between their "public" and "private" data further muddies the waters—especially because this distinction may not always be evident to data owners, or even knowable for some data.

Going forward, we argue that researchers should make the privacy ramifications of using certain public data sources clearer. Indeed, not all public data is created equal. Some public sources (e.g., Wikipedia) are highly curated and may pose low risks of containing sensitive information. Alternatively, some data sources might consist of public data that carries explicit consent to be used. It is an important open research question to understand if pretraining models solely on such stringently curated data can provide similar benefits for downstream tasks, while mitigating privacy risks.

2. Benchmarks conflate private and public distributions. Even if we were to solve the core privacy issue above—for example by pretraining very powerful models solely on non-sensitive datasets—it remains unclear if these models will actually be *useful* for privacy-sensitive downstream tasks.

We argue that the usefulness of the public-pretraining paradigm on private tasks is currently hard to assess, because existing benchmarks study "private" datasets that are not actually any more "sensitive" than the "public" dataset that is used for pretraining. In fact, the two are often drawn from the same (or from a similar) underlying distribution.

For example, when we transfer from ImageNet to CIFAR-10 (e.g., as by Tramèr & Boneh (2021); De et al. (2022)), *every single* class contained in the CIFAR-10 dataset has an identical class label in the ImageNet dataset! So can we say that any "private learning" actually happened? After all, training on ImageNet alone has already taught the model how to recognize a cat, or an airplane, or a dog—thus, (privately) finetuning on CIFAR-10 is merely performing a loose form of domain adaptation to classify low-resolution images more accurately. Despite this critique, this is a standard evaluation metric for private ML with public data.

Of course, the aforementioned papers do not actually care about privately classifying CIFAR-10, *per se*. Rather, they aim to provide and evaluate a general *framework* for combining public and private datasets. But as a result, we argue it is not clear that measuring progress of "private" learning on any of these benchmarks is at all meaningful. Specifically, are these benchmarks actually measuring progress in private learning? Or are they just a direct proxy for progress on non-private representation learning?

The answer to this question likely depends on whether there exists an overlap between "public" and "private" data in *real* privacy-sensitive applications. We posit that this will not be the case in many applications, i.e., the privacy sensitive data to be finetuned on will come from a data distribution that is only poorly represented in the public pretraining data. For example, machine learning on medical data is a canonical motivation for private ML, but the data distributions may not resemble those which are publicly accessible.

Unfortunately, it has already been shown that if the overlap between the pretraining and target distributions is small, then current methods for large scale pretraining may be less effective. For example, in the challenging (but perfectly-privacy) zero-shot setting, foundation models tend to perform poorly on medical tasks. To illustrate, the authors of BASIC-L (a representation learning method) write:

> *PCam [a skin lesion dataset] is perhaps the most sensitive dataset where BASIC-L performs poorly. For such an important task, the top-1 accuracy of BASIC-L (59.6%) [is] far below the bars for practical deployments, [...] [and] just slightly above random guessing. [...] As our training data are weakly crawled and automatically curated from the internet, without any emphasis on medical images, our BASIC-L model cannot learn enough to perform well on PCam. [...] despite the benefits of open-vocabulary image classification models, they are not ready to be deployed to tasks that require in-domain expertise (Pham et al., 2023)*

There is evidence that some deficiencies of zero-shot learning in these settings may be overcome by (non-private) *finetuning* (Radford et al., 2021). But this need not always be the case. For example, while large language models pretrained on Internet text achieve impressive performance on a variety of downstream tasks (Brown et al., 2020), they still achieve poor performance when finetuned on (potentially sensitive) tasks that are only weakly represented online.

Understanding the efficacy and limitations of transfer learning has been a major research direction in the non-private

machine learning community, particularly through the lens of distribution shifts (Koh et al., 2021). Yet, it is not always clear whether conclusions from the non-private setting are valid when we introduce privacy constraints, as data that is very privacy sensitive may be poorly represented in the distribution of publicly available pretraining data.

We thus call for privacy researchers to consider (or create) new benchmarks that more closely match envisioned deployments of private learning. Such benchmarks could for instance leverage sensitive datasets that were publicly released for research purposes, such as e.g., MIMIC (Johnson et al., 2016) or the dataset from the infamous Netflix Prize (Bennett & Lanning, 2007).

3. Large private models require trusting cloud services. When we train a model with DP, this guarantees that anyone who can access the trained model cannot learn much about any individual training sample. However this is orthogonal to any confidentiality considerations about who sees the data during training and inference.

Ideally, when dealing with personal data (e.g., private chat messages), the sensitive data would not leave the individual's device. This is usually possible: the differentially private training could be decentralized (e.g., as in Federated Learning (McMahan et al., 2017)[1]), and the trained model could be shipped to people's devices for inference.

Unfortunately, unlocking the full power of large-scale public pretraining currently requires drastically scaling model sizes. With current techniques, most foundation models are impossible to train or serve on end-user devices. For example, MobileBERT (Sun et al., 2020)—a language model optimized for on-device inference—has about 25M parameters; this is between two and four orders of magnitude smaller than state-of-the-art language models considered in recent works on private finetuning of language models (Li et al., 2022; Arora & Ré, 2022; Yu et al., 2022).

We thus encourage researchers in private machine learning to also take into consideration the scale of these models, and their privacy implications. An important direction for future work is to develop techniques for *distilling* (Hinton et al., 2015) large foundation models into smaller, more efficient models that are tuned for a specific (private) task.[2]

Are these issues not also present in *non-private* learning? While our paper focuses on the shortcomings of large-scale public pretraining for *private* workloads, many of our criticisms may seem to apply more broadly to any application of pretrained models. However we believe that these issues are especially important in privacy-sensitive applications.

First, the act of labeling the whole Web as "public" for machine learning purposes is particularly egregious when these models are explicitly touted as "privacy preserving", as this dilutes the meaning of "privacy" and may downplay the benefits of other uses of privacy enhancing technologies.

Second, a large overlap between pretraining data and common benchmarks may not be a concern if the goal is to measure *absolute progress on the considered task*, rather than *progress of generic learning techniques*. Many standard benchmarks are useful in the former sense (e.g., ImageNet measures the ability to classify 1000 types of every-day objects). But since these tasks are not privacy relevant, their use as benchmarks in the privacy literature solely serves the latter goal: to evaluate the progress of generic (private) learning *techniques*. In this case, a large overlap between pretraining and finetuning tasks is problematic.

In the remainder of this paper we study each of the three challenges in more detail, provide further evidence for our claims, and discuss conclusions of our work.

2. Is publicly accessible data public?

When a model is reported to be "trained with differential privacy," it should mean something. And if a model is trained with DP from scratch, it means something very precise: no data specific to any individual training record will be memorized by the final model.[3] In the common pretrain-publicly-then-finetune-privately paradigm, the privacy semantics are slightly different. The finetuning dataset enjoys the privacy guarantees bestowed by DP, but there is **absolutely no privacy** afforded to data in the pretraining dataset. Our main argument in this section is that these privacy semantics, while rigorous and precise, fall short of satisfying several privacy norms in the manner they are generally used. As such, we consider it detrimental to label the resulting models as "privacy-preserving," as their guarantees are at odds with how most individuals would interpret such a claim.

The issue comes from the fact that such a "privately-trained" model will still leak details of individuals whose data were contained in the pretraining dataset. And if a data subject notices this and asks "if this model is private, why was my data leaked?" the only possible answer to give is that this data was not part of the dataset that was considered worth

[1] We remind that federated learning, even *with* differential privacy, comes with its own associated privacy risks (Zhu et al., 2019; Boenisch et al., 2023)

[2] Recently, subsequent to the original appearance of this paper, there has been significant interest in the development of powerful small models that facilitate on-device computation (e.g., Gunasekar et al., 2023; Li et al., 2023; Gemini Team, 2023).

[3] Informally speaking, if the training procedure satisfies (ε, δ)-DP, then with probability at least $1 - \delta$ the inclusion of one individual training record changes the probability of observing any outcome by at most a factor e^{ε}. Most DP models are trained with Rényi DP (Abadi et al., 2016; Mironov, 2017), which is translated into (ε, δ)-DP to give more interpretable guarantees.

protecting. Such an explanation would likely not be very satisfactory, as individuals may still view some publicly accessible data as sensitive—especially if, as we make the case here, it was not intentionally made public.[4]

This issue could be mitigated by pretraining models solely on data that is entirely non-sensitive. Alternatively, we could ask data owners to provide explicit consent for their data to be used for machine learning. But then we have to ensure that the resulting privacy risks are very clearly communicated first. For example, if some user application were to ask "please share your data to help improve this product," users may expect that their data will be shared with the application developer, but *not* potentially with all *other* application users.

Unfortunately, the value in pretraining currently seems to arise mostly from the fact that we are able to train on massive uncurated datasets. As a consequence of the size of these datasets, much of the collected content will inevitably come from uncertain origins with no explicit user consent, and requesting consent becomes challenging.

Such issues arise even for extremely well-studied and strongly supervised datasets such as ImageNet. Despite its ubiquitous use, the dataset contains sensitive content of individuals (e.g., images of children, nudity, etc.) (Quach, 2019). Some datasets have even been completely retracted on privacy grounds: TinyImages (Torralba et al., 2008) is a dataset of 80M images scraped from the Web which was later subsampled to create the CIFAR-10 dataset (Krizhevsky, 2009). This dataset has since been deprecated due to the discovery of offensive and derogatory images (Birhane & Prabhu, 2021). Larger-scale datasets used for natural language processing are possibly even more challenging to curate. These datasets are often hundreds of gigabytes (Gao et al., 2020) to terabytes (Hoffmann et al., 2022) in size, gathered mostly by scraping the Internet for any available text data, with minimal content filtering or curation.

While such datasets contain, by definition, only data that is *public* (in the sense of "publicly accessible" on the Internet), their use in ML still presents significant privacy risks, as illustrated by the following two (real) examples.

2.1. Two Motivating Examples

Intentionally shared data, for use in a particular context: Consider again the case of a company that trains a language model on the text messages of its end users. The company initializes their model with the publicly available GPT-2 model, and then finetunes it with DP on its own corpus of private chat messages. The company then deploys the model and promises users that this model is privacy preserving!

Peter W. uses the model and types: "`The phone number of Peter W. is:`" and the model autocompletes his correct phone number (and also helpfully supplies his fax number, physical address, and email address). Peter W. claims this model violates his privacy. The company assures Peter that it does not, since their implementation satisfies a state-of-the-art ($\varepsilon = 0.1, \delta = 10^{-12}$) level of differential privacy (with respect to the data used for finetuning). His privacy was actually compromised when he posted his phone number along side some technical documents in a report to the government several years ago.

Peter might not be fully satisfied with such an answer. We argue that many people might react like Peter if personal information about them were ever output by a "privacy preserving" machine learning model. In fact, this example is not hypothetical: Peter W. is a real person, and the GPT-2 language model does know his phone number for exactly the reason described above (Carlini et al., 2021). And he is not alone: even state-of-the-art production models like ChatGPT know the phone numbers of many people who placed their phone number online for one purpose only for it to be used during model training (Nasr et al., 2023a).

Similar issues could arise for other modalities, for example for photos that individuals post online. These may be posted for a particular purpose, e.g., on an individual's homepage for professional purposes, or on a social media site for sharing memories with friends. This does not imply that the subjects consent to all possible downstream uses—as one extreme example, consider a generative model trained on publicly accessible photos from the Web, that is then abused for deepfake pornography. This constitutes a clear privacy violation (Paris & Donovan, 2019), regardless of whether the model's training data was public or not.

Privacy violations could also arise if machine learning models create new ways of searching and linking data that was posted online anonymously or pseudonymously. Attacks of this nature have always been possible (e.g., using existing image search engines), but cutting-edge advances in ML make (and will continue to make) them easier and easier to mount. For example, a model similar to CLIP (Radford et al., 2021) that was trained on the entire Internet might enable problematic forms of image search (e.g., "find all online images that match this photo").

Not intentionally (or knowingly) shared data: Not all content available on the Internet is posted intentionally. In some cases, the original owner might not even know their

[4]We comment that, even *without* any public data, and *with* DP training correctly implemented, there are still risks that could lead to catastrophic privacy violations, generally with very low probability of occurrence. As one extreme example (which occurs with exceptionally low probability), consider a run of DPSGD in which all added noise happens to be negligibly small, and thus has similar privacy risks to an unprotected model. Such failures of DP are possible, albeit unlikely, and outside the scope of our paper.

information has been posted without their consent.

We begin with another example from GPT-2 where data was posted without the knowledge of the original author, and was then used to train a large language model. The GPT-2 language model was trained on many webpages that were linked to by the social media website Reddit. One of these articles was a transcript of an IRC conversation (of several thousand messages) between several individuals discussing sensitive political and societal topics. These individuals were likely not aware their private conversations were recorded—let alone published for anyone to see. However GPT-2 trained on this data as if it was intentional.

Another example involves surveillance camera footage. While surveillance cameras are ubiquitous, the public generally expects recordings to be available exclusively to security personnel. Nonetheless, many surveillance camera configurations employ minimal security, leading to livestreams of their feeds being publicly available (Xu et al., 2018).

Data can also be published unintentionally. Reportedly,[5] in at least one example, a GitHub user unintentionally uploaded information about their cryptocurrency wallet to a public Git repository. When Copilot (a coding assistant language model) (Chen et al., 2021) was trained on this repository, it memorized this wallet's private key and allowed another user to withdraw money from the account. In all likelihood, this user did not post the key to their cryptocurrency wallet on their GitHub intentionally.

In all of these examples, the data is publicly accessible on the Internet, but particular uses of this data constitutes a significant violation of privacy norms, particularly when used in combination with machine learning. One could argue that it is not the act of *training* the machine learning model that caused a privacy leak—but rather the original act of *publishing* this data on the Internet.

We disagree! ML models (and foundation models in particular) have the capacity to *amplify* this leakage by disseminating this information in a much broader context, e.g., in new applications built on top of these pretrained models. As an analogy, a malicious person who *doxes* another by "releasing someone's personal details onto the Internet in an easily accessible form" is still causing harm even though "these details may already be publicly available, but in difficult to access forms or distributed across various sources that obscure them from casual discovery" (Douglas, 2016).[6]

The recent work of Brown et al. (2022) raises some related concerns, for the specific case of language models. Their core argument is that the privacy categorization of text data is inherently *contextual* (see also, the more recent works of Mireshghallah et al. (2024); Hartmann et al. (2023); Neel & Chang (2023)). Thus, collecting text data from various contexts on the Web and aggregating it into a single public-facing language model may violate users' privacy expectations for this data. We expand on this argument here by: (1) considering other publicly available data modalities than just text; and (2) by discussing the potential erosion of trust in privacy technologies that could arise when conflating "public" and "private" data sources (see below).

2.2. Privacy Expectations

Given that not all content available online has the expectation of being used to pretrain large models (either because it was posted online for one particular purpose, or because it was not even posted intentionally), this raises reasonable privacy considerations for training on this "public" data.

By publicizing such models as being "privacy preserving," the leakage of people's so-called "public" data could erode their trust in technologies such as differential privacy (which is not technically at fault here). This trust erosion could then also (mistakenly) carry over to settings where differential privacy *is* applied properly to all collected data—e.g., the collection of census data in the US.

A counterargument may be that the issue here is merely one of properly *educating* the public about DP and its guarantees (in the face of public pretraining). But it is already challenging for users to understand how the semantic guarantees of DP align with their own notions of "privacy", even when no pretraining is involved. By training non-privately we now compound this issue by introducing two "tiers" of data that are presumed to have very different privacy expectations attached to them. But if peoples' own privacy expectations do not match with this assumption, we run the risk that "privacy-preserving" models will cause real privacy harms.

3. Are we still measuring progress on private learning?

The purpose of a benchmark is to measure progress on a particular task of interest. The ImageNet benchmark, for example, measures the ability of classifiers to perform image classification across a range of everyday objects.

It is important to use the right benchmark—one where progress serves as an appropriate proxy for progress on the true task of interest. For example, while researchers had historically used the MNIST dataset of hand-written digits to evaluate the performance of neural networks, today this dataset is not seen as a reliable measure of progress. This is both because it has become "too easy", and also because

[5] https://www.theregister.com/2022/05/03/openai_copilot_cryptocurrency/

[6] We emphasize that we are making a (subjective) *moral* argument that the model trainer bears some culpability for propagating this information. Legally, their liability may differ based on jurisdiction, and we omit further discussion for simplicity.

lessons learned from squeezing the last 0.1% test accuracy out of the dataset often do not generalize to more interesting datasets (e.g., Goodfellow et al., 2014).

We now make the case that current benchmarks used to evaluate privacy-preserving machine learning are similarly insufficient, and that we should instead study tasks that are more directly indicative of performance on real-world privacy-sensitive tasks.

3.1. Intra-domain versus Cross-domain Finetuning

Existing research that follows the public-pretraining-and-private-finetuning paradigm discussed earlier has so far focused mostly on the following tasks:

- Pretrain on CIFAR-100 or ImageNet and finetune on CIFAR-10 (Abadi et al., 2016; Papernot et al., 2019; Tramèr & Boneh, 2021; Panda et al., 2022)
- Pretrain on Places365 and finetune on ImageNet (Kurakin et al., 2022)
- Pretrain on JFT or LAION-5B and finetune on ImageNet (De et al., 2022; Mehta et al., 2023)
- Pretrain on text data scraped from the Web, and finetune on other text data scraped from the Web (Yu et al., 2021b; Li et al., 2022; Yu et al., 2022)

The issue is that many of the above settings have the property that the "private" finetuning data distribution is essentially a subset of the "public" data distribution. For example, the data distribution from which CIFAR-10 is drawn is a *strict subset* of the data distribution from which ImageNet is drawn. CIFAR-10 is drawn from (heavily downsampled) images from the Internet representing one of 10 objects: cats, horses, airplanes, etc. ImageNet is similarly drawn from images from the Internet representing one of 1000 objects, *including each of the CIFAR-10 classes*. So when we pretrain on ImageNet and privately finetune on CIFAR-10, is any "private learning" actually happening or are we merely performing a loose form of intra-domain transfer from high-resolution to low-resolution images? The latter is a worthy goal, but the former is better aligned with what researchers try to understand in these settings: how to adapt to novel concepts which are only well-represented in sensitive data. We emphasize that the issue here is an overlap property of the data *distributions*, which occurs even when the *datasets* themselves are entirely disjoint.

The reliance on public pretraining with significant overlaps between "public" and "private" data distributions has become more prevalent in recent research. Abadi et al. (2016) were, to our knowledge, the first to present private learning results with public pretraining. While their "private" dataset (CIFAR-10) and "public" dataset (CIFAR-100) share close similarities, the authors argued that "the examples and the image classes [of CIFAR-100] are different from those of CIFAR-10." Follow-up papers then moved on to using larger pretraining sets (e.g, ImageNet (Tramèr & Boneh, 2021; De et al., 2022)) but omitted the concern about class overlap between the private and public datasets.

The situation is analogous when benchmarking large language models. For example, the GPT-4 (OpenAI, 2023) and Gemini (Team et al., 2023) papers present detailed analyses of how common evaluation benchmarks in NLP might overlap with the model's training data, and the CLIP paper (Radford et al., 2021) analyzes how many ImageNet test images might be contained in their CLIP training dataset.

To highlight an extreme example of the questionable use of "public pretraining," two recent works in this area (De et al., 2022; Mehta et al., 2023) have pretrained on Google's JFT dataset (Zhai et al., 2022; Sun et al., 2017) to achieve high accuracy (privately) on ImageNet. However while ImageNet (Deng et al., 2009) is a public dataset that any researcher is allowed to download, JFT is a proprietary dataset of 4 *billion* Web images collected and labeled by Google that has not been made public. Thus, in this setup the "private" dataset is actually more accessible than the "public" dataset![7] On top of this, the datasets are similar enough that parts of ImageNet are directly contained in JFT. The papers above do account for this by removing images from JFT that are near-duplicates of images from ImageNet. But the fact remains that here the private and public datasets are essentially identically distributed—and thus likely not representative of many real private learning scenarios.

Of course, the above papers merely adopted these benchmarks as illustrative examples of a public-to-private transfer setup. Yet, we caution against such benchmarks becoming the standard for assessing progress in private learning *techniques*. This is because **these benchmarks make it hard to disentangle generic progress in unsupervised representation learning, from algorithmic improvements for private learning**. This issue is compounded by the fact that there is no consensus on what public pretraining data to use, and so different papers use a wide variety of incomparable public sources. For instance, prior work has presented results for private learning on CIFAR-10 while leveraging the

[7]The machine learning community has created large-scale, open-source datasets which can serve as an alternative to proprietary datasets like JFT. Some notable such datasets include LAION-5B (Schuhmann et al., 2022) and the Pile (Gao et al., 2020). However, these particular datasets have been taken down, by the authors due to unintentional indexing of child sexual abuse material, and by a DMCA request, respectively. These developments leave nebulous the future of large open-source datasets, and raise further question about the nature of the contents of proprietary datasets.

following public data sources: CIFAR-100 (Abadi et al., 2016; Asadian et al., 2022); *unlabeled* CIFAR-100 (Asadian et al., 2022); ImageNet (De et al., 2022); *unlabeled* ImageNet (Tramèr & Boneh, 2021); 2,000 random ImageNet samples (Yu et al., 2021a); a *single* 600×225 image engineered for pretraining (Asadian et al., 2022), etc.

A potential explanation for some of these "esoteric" choices of pretraining datasets is that without any restriction on what can be considered as "public" data, some benchmarks (such as CIFAR-10) become uninteresting as the private task can essentially already be solved with *perfect privacy* (Arora & Ré, 2022). For example, OpenAI's pretrained CLIP model gets 96.2% *zero-shot* accuracy on the CIFAR-10 dataset (without any finetuning), just a few percentage points shy of the $\sim 99\%$ state-of-the-art using this dataset alone—despite the fact that CLIP never saw *any* CIFAR-10 training data! Thus, by definition, CLIP achieves 96.2% accuracy at $(\varepsilon, \delta) = (0, 0)$-DP. Similarly, Pham et al. (2023) achieve zero-shot 85.7% top-1 accuracy on ImageNet, thus achieving perfect privacy and state-of-the-art accuracy even compared to models with much larger values of ε (Mehta et al., 2023; De et al., 2022).

Why do we believe this is a problem? After all, if it is possible to reach 96% accuracy on CIFAR-10 without even inspecting the training dataset, is this not actually private? Again, our concern comes down to the fact that the underlying data distribution for both CIFAR-10 and CLIP's training dataset are the same: images scraped from the Internet.[8]

As a result, unlike traditional forms of *cross-domain transfer* where we must take a classifier trained in one setting (e.g., pictures taken from the Internet) and transfer them to a completely new setting (e.g., classifying tumors), we have no such need here. Instead, it is sufficient to adapt from one sampling from a distribution (images from the Internet) to another sampling of the same distribution—albeit a sample with slightly different preprocessing. Thus, we argue that benchmarks such as private learning on CIFAR-10 (with pretraining) are a bit like MNIST for general computer vision: there is little performance left to be squeezed out ($\approx 2\%$), and this marginal progress might not carry over to real privacy-sensitive settings where a close overlap between public and private data sources may not exist.

3.2. Towards Better Benchmarks

Ultimately, the question we want to answer is whether or not we (as a community) are making progress towards effective private learning. It is possible that we live in a world—which we will call `Transfermania`—where all sensitive tasks we care about (e.g., medical classification tasks) are well represented by data that is publicly available, e.g., on the Internet. Alternatively, we might live in a different world—`Privacyland`—where these sensitive tasks are *not well* represented and mostly disjoint from any public data.

Knowing which of these two worlds we are in is important! If we are in `Transfermania`, then everything is great: for a sensitive task of interest, simply use a publicly pretrained foundation model and solve the task with either zero-shot learning (Arora & Ré, 2022) or minimal finetuning (barring the issues raised above about the pretraining data actually also being sensitive). In contrast, if we are in `Privacyland`, then foundation models pre-trained on Internet data might be of little help in many privacy-sensitive settings, as exemplified by the minor benefit of transfer learning reported by Raghu et al. (2019); Pham et al. (2023) on some medical tasks.

Crucially, we cannot know in which world we are in without *collecting data that resembles privacy-sensitive tasks that we actually care about* (i.e., not CIFAR-10 or ImageNet). We thus believe it is necessary for the private learning community to begin considering and curating new benchmarks—to properly disentangle advances in non-private representation learning from advances in privacy-preserving learning. Such benchmarks could include existing sensitive datasets that have been released for research purposes, e.g., medical datasets (Johnson et al., 2016; Irvin et al., 2019; Wang et al., 2017; Bejnordi et al., 2017), email corpora (Klimt & Yang, 2004), user reviews (Bennett & Lanning, 2007), etc.

Of course, it is possible we do live in `Transfermania`, and foundation models will perform well when tuned privately on sensitive tasks. This would be a very promising signal that private learning can be achieved in many real-world deployments. Nevertheless, concerns with the sensitive nature of pretraining sets (Section 2), and the necessity to outsource large models (Section A) could remain.

As a result, regardless of which world we are in, we encourage researchers to continue studying ways to make differentially private learning (without pretraining) better, as progress on this problem can also inform real-world deployments (even if these use some form of pretraining).

Inspired by this paper's first appearance, some subsequent works have studied the efficacy of cross-domain transfer in the differentially private setting. We draw attention to the work of Berrada et al. (2023), which performs public pretraining on large-scale public datasets such as ImageNet-21K and JFT, and private fine-tuning on a variety of datasets, including medical datasets like CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., 2019). They show that this strategy is able to achieve reasonably high utility (i.e.,

[8] Specifically, CIFAR-10 was collected as a subset of TinyImages, which itself is a dataset of 80 million images collected from the public Internet. Similarly, CLIP's training dataset is also a dataset of 400 million images collected by downloading images from the public Internet.

close to the non-private SOTA) even for datasets belonging to such specialized domains, providing evidence towards being in `Transfermania`. We note that the paper lacks a baseline of private training from scratch, so it is tough to evaluate precisely how much the public pretraining helped. As this work only offers results only for a few dataset pairs, it is important for the field to more broadly understand the efficacy of public pretraining for private ML before reaching any general conclusions.

4. Where do we go from here?

Public pretraining for private learning might not be the panacea that prior work has made it out to be, and we hope future work will carefully consider the use of public data when performing private training. We conclude by outlining open questions and possible directions for future work:

- **Articulate granular privacy considerations for Web data.** The private learning literature often falls back on a simplified dichotomy where all data is either "public" or "private." Yet, individual expectations about privacy are rarely so binary (Nissenbaum, 2004).

  We thus encourage privacy researchers to advocate for a more responsible and granular approach to privacy when it comes to collecting training datasets—and especially datasets collected from the Internet. This could include developing techniques and procedures for establishing *consent* for using Internet data as training data, for auditing existing datasets (including proprietary ones) for sensitive content, and encouraging appropriate disclosure of any privacy concerns (for example, in an accompanying datasheet (Gebru et al., 2021)). This goal is part of a broader research direction focused on responsible dataset curation in machine learning (Bender et al., 2021; Mitchell et al., 2022).

- **Construct privacy-friendly pretrained models.** It is an open problem whether one could train a (useful) foundation model that does not carry the burden of increased privacy risks. One avenue could be to curate and pretrain models on large Internet datasets that do not contain any privacy-sensitive data. This would require careful consideration pertaining to which data should and should not be considered sensitive.[9] Another approach would be to obtain explicit consent-of-use from data owners. Finally, it may be possible to pretrain a foundation model itself with differential privacy (Anil et al., 2022; Ponomareva et al., 2022). A core open problem here is how to set the right granularity for DP when training on data aggregated from various sources (Brown et al., 2022). For example, while current works aim to pretrain language models with privacy at the level of individual *sentences* (Anil et al., 2022; Ponomareva et al., 2022), such privacy guarantees are insufficient unless all references to a piece of sensitive information have first been rigorously eliminated or deduplicated (Lee et al., 2022).

- **Design better benchmarks to measure progress in private learning.** Unfortunately, no good benchmark for *private* learning currently exists. By this, we mean a benchmark that correlates well with the true task we care about, namely, private learning in sensitive domains. Privacy-sensitive data is likely to have many characteristics that standard ML datasets lack. Thus, by re-purposing existing ML benchmarks for private learning we run the risk of promoting progress metrics that are only weakly correlated (or not at all correlated) with progress on real privacy-sensitive tasks. This issue is exacerbated with the avenue of public pretraining, since many canonical ML benchmarks can now be solved privately "for free." We thus encourage the community to explore alternative benchmarks that more closely align with privacy-sensitive tasks of interest.[10]

- **Promote a holistic view on ML privacy.** The issues discussed in this paper fall under the broader concern that the current ML privacy literature is predominantly too "model centric." That is, most research focuses on the narrow (but important) problem of training a model (once) with DP. This line of research largely ignores broader privacy considerations around data collection (what data is collected, from what sources, and why?), data lifetimes (for how long is data kept, and how many models are trained on it?), model lifetimes, etc. We encourage further research on these important topics.

Finally, while the overall tone of our article is critical, we recognize and highlight that many recent works employing public data have played an important role in showing that differential privacy *can* be preserved for certain complex machine learning problems, without suffering devastating impacts on utility. This is an important step forward for the field. We focused our attention on what we believe to be some of the most important considerations in this area, in an effort to steer the community towards making the next important steps advancing private machine learning.

[9] Consequently, we intentionally refrain from making any broad prescriptions on this front, as what is and is not private depends significantly on *context*, with considerations including (but not limited to) the types of data, the application, and relevant privacy norms.

[10] We believe that choosing the right benchmarks for private machine learning is a consequential task, deserving of significant exploration and justification. Such investigation is beyond the scope of the present position paper. As a result, we explicitly abstain from prescribing any specific benchmarks, and leave this for future work.

Acknowledgments

We would like to thank Davis Blalock, Ivan Habernal, Tatsunori Hashimoto, Janardhan Kulkarni, Alexey Kurakin, Katherine Lee, Xuechen Li, Percy Liang, Ashwinee Panda, Thomas Steinke, Ivy Vecna, Sergey Yekhanin, and anonymous reviewers for their valuable comments on previous versions of this work.

GK is supported by an NSERC Discovery Grant, a Canada CIFAR AI Chair, and an unrestricted gift from Google.

Impact Statement

As most of this position paper is already focused on ethical aspects and future societal consequences related to privacy in machine learning, we do not believe further discussion is warranted.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. ACM.

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review*, Special Issue 2, 2022.

Anil, R., Ghazi, B., Gupta, V., Kumar, R., and Manurangsi, P. Large-scale differentially private BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Findings of EMNLP '22, pp. 6481–6491. Morgan Kaufmann Publishers Inc., 2022.

Arora, S. and Ré, C. Can foundation models help us achieve perfect secrecy? *arXiv preprint arXiv:2205.13722*, 2022.

Asadian, A., Weidner, E., and Jiang, L. Self-supervised pretraining for differentially private learning. *arXiv preprint arXiv:2206.07125*, 2022.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623. ACM, 2021.

Bennett, J. and Lanning, S. The Netflix prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, 2007.

Berrada, L., De, S., Shen, J. H., Hayes, J., Stanforth, R., Stutz, D., Kohli, P., Smith, S. L., and Balle, B. Unlocking accuracy and fairness in differentially private image classification. *arXiv preprint arXiv:2308.10888*, 2023.

Birhane, A. and Prabhu, V. U. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision*, WACV '21, pp. 1536–1546. IEEE, 2021.

Bittau, A., Erlingsson, Ú., Maniatis, P., Mironov, I., Raghunathan, A., Lie, D., Rudominer, M., Kode, U., Tinnes, J., and Seefeld, B. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th ACM Symposium on Operating Systems Principles*, SOSP '17, pp. 441–459, New York, NY, USA, 2017. ACM.

Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A. S., Shumailov, I., and Papernot, N. When the curious abandon honesty: Federated learning is not private. In *Proceedings of the 8th IEEE European Symposium on Security and Privacy*, EuroS&P '23, pp. 175–199. IEEE Computer Society, 2023.

Bommasani, R. et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brown, H., Lee, K., Mireshghallah, F., Shokri, R., and Tramèr, F. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 2280–2292. ACM, 2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20. Curran Associates, Inc., 2020.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, USENIX Security '19, pp. 267–284. USENIX Association, 2019.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium*, USENIX Security '21, pp. 2633–2650. USENIX Association, 2021.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Cummings, R., Kaptchuk, G., and Redmiles, E. M. "i need a better description": An investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3037–3052, 2021.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '09, pp. 248–255, Washington, DC, USA, 2009. IEEE Computer Society.

Douglas, D. M. Doxing: a conceptual analysis. *Ethics and information technology*, 18(3):199–210, 2016.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pp. 265–284, Berlin, Heidelberg, 2006. Springer.

Ellis, M. and Thomas, Z. Finding your personal data online is easy. taking it down is harder. https://www.wsj.com/podcasts/google-news-update/finding-your-personal-data-online-is-easy-taking-it-down-is-harder/e073e152-f5fd-43df-92be-6d5a6c6199ba, 2022.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM Symposium on the Theory of Computing*, STOC '20, pp. 954–959, New York, NY, USA, 2020. ACM.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pp. 2881–2891. Curran Associates, Inc., 2020.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gentry, C. *A fully homomorphic encryption scheme*. Stanford university, 2009.

Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., and Wernsing, J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International conference on machine learning*, pp. 201–210. PMLR, 2016.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gunasekar, S., Zhang, Y., Aneja, J., César Teodoro Mendes, C., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Singh Behl, H., Wang, X., Bubeck, S., Eldan, R., Tauman Kalai, A., Lee, Y. T., and Li, Y. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Hartmann, V., Suri, A., Bindschaedler, V., Evans, D., Tople, S., and West, R. SoK: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*, 2023.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22, pp. 30016–30030. Curran Associates, Inc., 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, ICLR '22, 2022.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI '19, pp. 590–597, 2019.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Klimt, B. and Yang, Y. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pp. 217–226. Springer, 2004.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, ICML '21, pp. 5637–5664. JMLR, Inc., 2021.

Krizhevsky, A. Learning multiple layers of features from tiny images, 2009.

Kurakin, A., Chien, S., Song, S., Geambasu, R., Terzis, A., and Thakurta, A. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '22, pp. 8424–8445. Association for Computational Linguistics, 2022.

Li, X., Tramèr, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *Proceedings of the 10th International Conference on Learning Representations*, ICLR '22, 2022.

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., and Lee, Y. T. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

McSherry, F. Statistical inference considered harmful, 2016a. URL https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md.

McSherry, F. Differential privacy and correlated data, 2016b. URL https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-29.md.

Mehta, H., Thakurta, A. G., Kurakin, A., and Cutkosky, A. Towards large scale transfer learning for differentially private image classification. *Transactions on Machine Learning Research*, 2023.

Micali, S., Goldreich, O., and Wigderson, A. How to play any mental game. In *Proceedings of the 19th Annual ACM Symposium on the Theory of Computing*, STOC '87, pp. 218–229. ACM, 1987.

Mireshghallah, N., Kim, H., Zhou, X., Tsvetkov, Y., Sap, M., Shokri, R., and Choi, Y. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. In *Proceedings of the 12th International Conference on Learning Representations*, ICLR '24, 2024.

Mironov, I. Rényi differential privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium*, CSF '17, pp. 263–275, Washington, DC, USA, 2017. IEEE Computer Society.

Mitchell, M., Luccioni, A. S., Lambert, N., Gerchick, M., McMillan-Major, A., Ozoani, E., Rajani, N., Thrush, T., Jernite, Y., and Kiela, D. Measuring data. *arXiv preprint arXiv:2212.05129*, 2022.

Mohassel, P. and Zhang, Y. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*, pp. 19–38. IEEE, 2017.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023a.

Nasr, M., Mahloujifar, S., Tang, X., Mittal, P., and Houmansadr, A. Effectively using public data in privacy preserving machine learning. In *International Conference on Machine Learning*, pp. 25718–25732. PMLR, 2023b.

Neel, S. and Chang, P. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*, 2023.

Nissenbaum, H. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

Ohrimenko, O., Schuster, F., Fournet, C., Mehta, A., Nowozin, S., Vaswani, K., and Costa, M. Oblivious multi-party machine learning on trusted processors. In *25th USENIX Security Symposium (USENIX Security 16)*, pp. 619–636, 2016.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Panda, A., Tang, X., Sehwag, V., Mahloujifar, S., and Mittal, P. Dp-raft: A differentially private recipe for accelerated fine-tuning. *arXiv preprint arXiv:2212.04486*, 2022.

Papernot, N., Chien, S., Song, S., Thakurta, A., and Erlingsson, U. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. https://openreview.net/forum?id=rJg851rYwH, 2019.

Paris, B. and Donovan, J. Deepfakes and cheap fakes: The manipulation of audio and visual evidence. *United States of America: Data & Society*, 1, 2019.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y., Tan, M., and Le, Q. V. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

Ponomareva, N., Bastings, J., and Vassilvitskii, S. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2182–2193, 2022.

Quach, K. Inside the 1TB ImageNet data set used to train the world's AI: Naked kids, drunken frat parties, porno stars, and more. https://www.theregister.com/2019/10/23/ai_dataset_imagenet_consent, 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pp. 3347–3357. Curran Associates, Inc., 2019.

Sander, T., Stock, P., and Sablayrolles, A. TAN without a burn: Scaling laws of DP-SGD. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pp. 29937–29949. JMLR, Inc., 2023.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 2*, NeurIPS Datasets and Benchmarks '22, 2022.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, SP '17, pp. 3–18, Washington, DC, USA, 2017. IEEE Computer Society.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., and Zhou, D. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pp. 2158–2170. Association for Computational Linguistics, 2020.

Tang, X., Panda, A., Sehwag, V., and Mittal, P. Differentially private image classification by learning priors from

random processes. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23, pp. 35855–35877. Curran Associates, Inc., 2023.

Tang, X., Panda, A., Nasr, M., Mahloujifar, S., and Mittal, P. Private fine-tuning of large language models with zeroth-order optimization. *arXiv preprint arXiv:2401.04343*, 2024.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Tramèr, F. and Boneh, D. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. In *Proceedings of the 7th International Conference on Learning Representations*, ICLR '19, 2019.

Tramèr, F. and Boneh, D. Differentially private learning needs better features (or much more data). In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, 2021.

Tramèr, F., Kamath, G., and Carlini, N. Position: Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. Maxvit: Multi-axis vision transformer. In *Computer Vision - ECCV 2022 - 17th European Conference*, ECCV '22, pp. 459–479. Springer, 2022.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Xu, H., Xu, F., and Chen, B. Internet protocol cameras with no password protection: An empirical investigation. In *International Conference on Passive and Active Network Measurement*, pp. 47–59. Springer, 2018.

Yu, D., Zhang, H., Chen, W., and Liu, T.-Y. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, 2021a.

Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T.-Y. Large scale privacy learning via low-rank reparametrization. In *Proceedings of the 38th International Conference on Machine Learning*, ICML '21, pp. 12208–12218. JMLR, Inc., 2021b.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., Yekhanin, S., and Zhang, H. Differentially private fine-tuning of language models. In *Proceedings of the 10th International Conference on Learning Representations*, ICLR '22, 2022.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pp. 14774–14784. Curran Associates, Inc., 2019.

A. Large models require uploading private data

In order to achieve high accuracy, transfer learning with differential privacy currently requires using enormous pretrained models. For example, state-of-the-art differentially private ImageNet models have over 250 **million** parameters and require over 100 **billion** FLOPs to evaluate (De et al., 2022; Mehta et al., 2023). And while models of this size can still be run on high-end customer GPUs (250 million parameters require "only" 1 GB of memory), the size of state-of-the-art models is currently scaling at a much faster rate than (consumer) hardware.[11] (The largest language models, for example, are already larger than 500 **billion** parameters.)

As a result, using these enormous private models will likely require that data owners upload their private data to some remote cloud service. This causes a direct tradeoff between the privacy of the individuals who provide the private training data and the privacy of the end users of the trained model.

To illustrate, suppose that the final model *must* meet a minimum accuracy level to be viable (potentially at the cost of privacy). This accuracy could be reached in one of two ways: (1) use a very large pretrained model and finetune it with DP on sensitive data; (2) use a smaller model (possibly

[11] For example, from 2019 to 2022 the size of the largest language models grew by a factor $100 - 1000\times$, while the transistor count and memory size of consumer GPUs grew by less than $2\times$ (e.g., for Nvidia's GeForce series).

also pretrained) and finetune it without DP (or with very low privacy guarantees) on sensitive data.[12]

The latter approach—using a smaller (non-privately-finetuned) model—has the advantage that the model can be evaluated locally on each user's own device and thus poses no privacy risks to the model's ultimate end users. However, the sensitive training data is not guaranteed any protection from being memorized. In contrast, the former approach of privately tuning a larger pretrained model has the advantage that we can achieve more stringent DP guarantees for the finetuning data without sacrificing model utility, but requires the model's end users to upload their private data to a remote service.

The situation is actually slightly more complicated than this, because these models are also too large to be *finetuned* locally (e.g., using federated learning). While this burden could be reduced by employing parameter-efficient methods such as LoRA (Hu et al., 2022), enabling fine-tuning of the very largest models on-device seems far out of reach. Furthermore, methods such as *local* differential privacy (Kasiviswanathan et al., 2011) introduce too much noise to see practical deployment for such settings (Bittau et al., 2017). As a result, owners of the sensitive training data face a trade-off between either having their data be centrally collected for differentially private training, or keeping their data locally but not having any differential privacy guarantees.

In principle, the confidentiality of outsourced sensitive data (both for training and inference) could be guaranteed using cryptographic techniques such as fully homomorphic encryption (Gentry, 2009; Gilad-Bachrach et al., 2016) or secure multiparty computation (Micali et al., 1987; Mohassel & Zhang, 2017). Unfortunately, for the time being we are several orders of magnitude away from being able to efficiently apply these techniques in practice to large models. Outsourcing ML workloads to trusted execution environments (Ohrimenko et al., 2016; Tramèr & Boneh, 2019) is another possible alternative, but the security (and scalability) of existing platforms are also currently limited and are unlikely to handle billion-parameter models.

[12] Specifically, this hypothetical precludes finetuning a smaller model with strong DP guarantees, as the resulting accuracy would be too poor for use.