DIFFERENTIALLY PRIVATE MIXING OF PUBLIC DATASETS IMPROVES PRIVATE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many machine learning applications involve sensitive data and therefore require training under differential privacy (DP). However, DP training often degrades model utility. In some cases, first pre-training the model on "public" data before finetuning with DP on the sensitive data can reduce the drop in utility. However, the success of this depends on how relevant the selected public dataset is to the sensitive data. We introduce the first pipeline that privately learns the mixture of several public datasets to pretrain on for a given sensitive downstream task. Our key insight is that we can privately find the best mixture of multiple public datasets by privately learning a low-dimensional linear model. We tested our method on the NIH dataset for X-ray classification and the ENRON email dataset for language modeling. Applying our method to find tailored mixtures of X-ray datasets to pretrain on for diseases in the NIH ChestX-ray14 dataset, we improved macro AUC by up to 0.037 across privacy budgets compared to the baselines, with gains as large as +22.8% relative AUC on Cardiomegaly at $\epsilon = 1$. For DP training on the ENRON dataset, pre-training on our mixture of The Common Pile (a collection of public-domain text datasets) decreased test perplexity by 16% relative to the baseline mixtures.

1 INTRODUCTION

For many machine learning tasks, the only data we have comes from sensitive sources, such as medical records (Johnson et al., 2019a) or conversations between individuals (Klimt & Yang, 2004b). Models trained on such data, even if the data is anonymized, are known to leak information about the individuals who contributed to the training dataset (Shokri et al., 2017; Carlini et al., 2021). Training with differential privacy (DP) is the de-facto framework for addressing this concern (Dwork et al., 2006; Dwork & Roth, 2014); DP guarantees that the models trained using an individual's data are indistinguishable to the models trained without their data. However, training with DP often comes with a large drop in model utility (Abadi et al., 2016; De et al., 2022). Past work has observed that first pre-training a model on public data (data that can be treated as non-sensitive, hence not contributing to privacy leakage) before DP finetuning the model on the sensitive data can sometimes reduce the drop in model utility (De et al., 2022; Yu et al., 2022). Importantly, the effectiveness of this pre-training depends on how related the selected public data is to the target (sensitive) data (Ben-David et al., 2010; Isik et al., 2024; Pouget et al., 2024; Dominguez-Olmedo et al., 2024; Ganesh et al., 2023; Tramèr et al., 2022). To the best of our knowledge, pre-training on all the public datasets uniformly is common practice and state-of-the art for privately mixing the public dataset: we improve this.

In this paper, we study how to privately choose the public pre-training mixture that gives the best downstream performance on a given sensitive dataset. Finding the best mixture privately can generically be approached with private hyperparameter selection techniques (Liu & Talwar, 2019; Papernot & Steinke, 2022), where one trains on several different mixtures and then releases the best found mixture. However, this comes with significant computational costs: the space of mixtures which we need to trial (i.e., train on) grows exponentially with the number of public datasets. Compounding this, every candidate mixture we evaluate touches the sensitive data, so the more mixtures we try, the higher the privacy cost (Papernot & Steinke, 2022). We show that it is possible to significantly improve downstream utility over standard approaches to public data mixing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

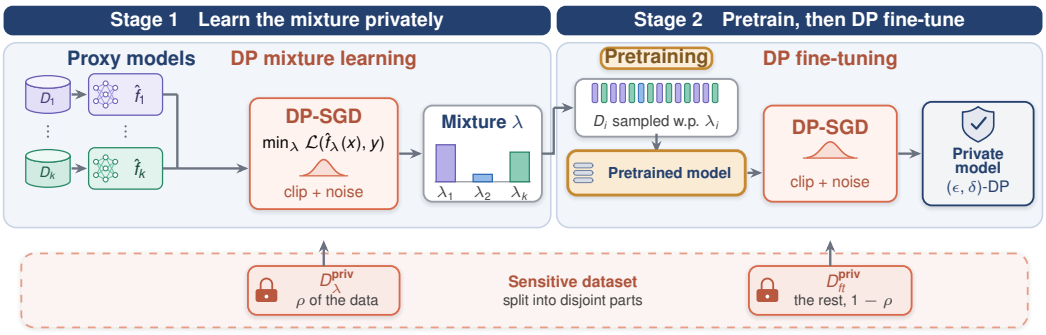


Figure 1: **Overview of DP-MixMin.** Proxy models trained on public data are combined with mixture weights λ learned with DP-SGD on a fraction ρ of the sensitive data (split A). We then pretrain on public data according to λ and DP fine-tune on the rest (split B).

Our key insight is that we can privately learn the weighting of public datasets by privately optimizing a convex objective. This builds off MixMin (Thudi et al., 2025), a non-private data mixing method which showed that as models approach Bayes optimal (the best a model can do with enough compute and data), finding the best weighting of pre-training datasets for a downstream task reduces to minimizing a convex objective. We build onto this insight by realizing this objective is amenable to private learning with little utility degradation. Formally, if k is the number of public datasets, by using existing DP algorithms (Bassily et al., 2019; Feldman et al., 2020; Asi et al., 2021; Bassily et al., 2021) we have the privacy cost for a fixed error grows at worst $O(\sqrt{k})$. In contrast, for certain data mixing problems, private hyperparameter search would need to try exponentially many mixtures, which translates to a privacy cost that grows linearly in k (Papernot & Steinke, 2022). However, a key limitation of this convex objective is that it requires having models near Bayes optimal on each of the individual pre-training sources; we hence explore empirical approaches.

Our pipeline, DP-MixMin, empirically instantiates the DP convex optimization for data mixing alongside further DP finetuning (Figure 1); our pipeline provides one end-to-end privacy guarantee. We first split the sensitive data into two disjoint parts: one for mixture learning with a variant of DP-SGD (Abadi et al., 2016), and the other for DP fine-tuning the model pre-trained on the mixture. Concretely, we empirically learn the mixture by replacing the Bayes optimal models for each source with a proxy model trained on each individual public dataset (following past work (Thudi et al., 2025)); these models do not see the private data, so leak no privacy. We then optimize the convex objective for data mixing with DP-SGD, providing a DP bound for the mixture. In the final stage we pre-train on this mixture, and then apply DP finetuning on the held out set. Because the two DP stages (mixing and fine-tuning) use disjoint datasets, parallel composition (Dwork & Roth, 2014) bounds the pipeline's overall privacy cost by the worse of the two steps, not their sum. So *learning the mixture costs data, not privacy*. DP-MixMin fine-tunes on less data than our baseline of finetuning a model trained uniformly over all public datasets, but we show that better mixtures more than make up for it.

We empirically tested DP-MixMin across medical imaging and language tasks, and found across varying DP guarantees it consistently improved baselines (Figure 2). For medical imaging, we used the 14 disease classification tasks in the NIH chest X-ray dataset (Wang et al., 2017) as sensitive datasets, and CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., 2019a) as public sources. These sources also contain classification tasks over chest X-rays, but differ in the diseases and how the images were labelled. DP-MixMin consistently outperformed the baselines, with an average AUC gain of ≈ 0.03, including a 22.8% relative gain on the Cardiomegaly class. For language modeling, we used Enron (Klimt & Yang, 2004b) as the sensitive dataset, targeting *user-level* DP; protecting all of a user's emails, a strictly stronger and costlier guarantee than protecting single emails (Kandpal et al., 2024; Levy et al., 2021; Chua et al., 2024). Pre-training on mixtures of The Common Pile (Kandpal et al., 2025) decreased test perplexity by up to 16%, with similar gains under example-level DP on OpenWebMath (Paster et al., 2023). We also found that DP-MixMin recovered mixtures close to the non-private solution in most settings, despite using a fraction of the sensitive data.

To conclude, our contributions are:

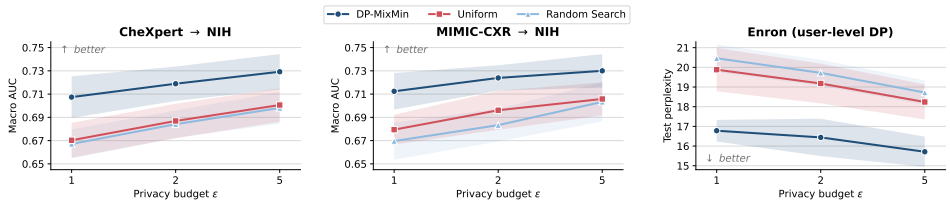


Figure 2: **DP-MixMin beats Uniform and Random Search at every ε.** Macro AUC on CheXpert/MIMIC-CXR→NIH (↑) and Enron test perplexity at $\rho = 0.3$ (↓).

1. DP-MixMin: The first end-to-end privatized pretraining mixture-learning pipeline.
2. Theoretically showing that when we train models near Bayes optimal, our procedure asymptotically outperforms the baseline of private hyperparameter search.
3. Empirically showing that across medical imaging and language modeling tasks, DP-MixMin improves the performance of public-pretraining over past mixture finding approaches.

2 Related Work

We now discuss past work on DP finetuning and non-private mixing we build on, and the limitations of previous public data selection for DP learning: they could not improve over a uniform mixture.

DP public pretraining and non-private data mixing. Public pretraining followed by private finetuning drives state-of-the-art DP results across language (Li et al., 2022; Yu et al., 2022) and vision (De et al., 2022; Mehta et al., 2022), and its effectiveness depends heavily on which public data is used (Ganesh et al., 2023; Tramèr et al., 2022). Non-privately, optimizing mixtures over many sources has shown to provide large gains in pre-training performance (Gao et al., 2020; Xie et al., 2023; Liu et al., 2024; Gururangan et al., 2020; Thudi et al., 2025); among these, MixMin (Thudi et al., 2025), which we build on, showed that the mixture-learning objective is (asymptotically) equivalent to fitting a linear model over per-source proxy model predictions.

Prior work on private selection of public datasets. Gu et al. (2025) propose GSD, which scores a candidate public dataset's relevance by the distance between its gradients' top-k subspace and the private task's, and picks the single source with the smallest distance to the private task. However, GSD breaks down by construction when the public sources share the same input space $\mathcal{X}$: their scoring uses a single model's gradient on randomly labeled examples from each source, so two sources with identical inputs become indistinguishable regardless of their actual label distributions.[1] Gu et al. (2025) also present no end-to-end private experiments—the privatized version of the selection step is proposed but never empirically validated. Our method, DP-MixMin, can operate when input distributions are shared and provides DP guarantees. Yu et al. (2024) train a DP domain classifier to rank examples within a single pre-chosen public corpus and pretrain on the top-ranked subset. While their method improves over the unfiltered baseline at smaller model scales, the benefit is no longer apparent at GPT2-Medium (355M parameters) [2]. In contrast, DP-MixMin continues to deliver substantial gains over unweighted pretraining at GPT2-Medium scale.

3 Preliminaries

In this section, we review differential privacy (DP), data mixing, and relevant past results.

Differential privacy. A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ϵ, δ)-differential privacy (Dwork et al., 2006) if, for every pair of adjacent datasets $D, D' \in \mathcal{D}$ (differing in a single

[1] We also implemented GSD (Gu et al., 2025), and confirmed we could not obtain meaningful results in our setting; implementation details and results are provided in Appendix A.4.

[2] Figure 2 in Yu et al. (2024) shows selective pretraining on 15% of OpenWebText and full pretraining on 100% of OpenWebText reaches the same perplexity.

record) and every measurable set $S \subseteq \mathcal{R}$,

$$\Pr[\mathcal{M}(D) \in S] \;\leq\; e^{\epsilon}\, \Pr[\mathcal{M}(D') \in S] \;+\; \delta,$$

where $\epsilon \geq 0$ is the privacy budget and $\delta \in [0, 1]$ bounds the probability of failure.

We will use DP-SGD, which privatizes the SGD algorithm by clipping individual gradients and adding noise (Abadi et al., 2016), and the post-processing property of DP which states any function (independent of the private dataset) of a DP algorithm is also a DP mechanism with the same guarantees (Dwork & Roth, 2014). We also leverage parallel composition, which states that if two DP mechanisms operate on disjoint parts of the private dataset, then the overall DP guarantee of releasing both mechanisms is at most the worst DP guarantee of the two mechanisms (McSherry, 2009).

User-level differential privacy. The definition above protects individual records (*record-level* or *example-level* DP). However, a single individual may contribute multiple records, and record-level guarantees can degrade substantially for such individuals. *User-level* DP (McMahan et al., 2018; Levy et al., 2021) aims to resolve this. Suppose each dataset D is partitioned by users $D = \bigcup_{u \in U} D_u$, where D_u is the (possibly multi-record) contribution of user u. Two datasets D, D' are *user-adjacent* if they only differ by the D_u of a single user. A mechanism $\mathcal{M}$ satisfies (ϵ, δ) user-level DP if the inequality above holds for every pair of user-adjacent datasets and every measurable $S \subseteq \mathcal{R}$. Note that while Record-level DP implies user-level DP through group privacy (Dwork & Roth, 2014), this bound is often loose. User-level guarantees are typically obtained directly without relying on record-level DP methods, e.g., by running DP-SGD (Abadi et al., 2016) with sampling, clipping, and noise applied per user rather than per example (McMahan et al., 2018). User-level DP satisfies both post-processing and parallel composition, provided the private dataset is split along user boundaries.

Data mixing setup. We consider learning from a collection of k public datasets $\{D_i^{pub}\}_{i=1}^k$ in order to improve performance on a private dataset. Each public dataset D_i^{pub} is assumed to be drawn from an underlying source distribution d_i over a shared input–output domain $\mathcal{X} \times \mathcal{Y}$. The private downstream dataset D^{priv} is drawn from a target distribution d_t, which is also over $\mathcal{X} \times \mathcal{Y}$.

Let $\mathcal{H}$ denote the hypothesis space for our models. We represent models as functions $f : \mathcal{X} \to \mathcal{O}$, where $\mathcal{O}$ is some output space (e.g., label probabilities), and we evaluate performance using a loss function $L : \mathcal{O} \times \mathcal{Y} \to \mathbb{R}_+$. We require the loss function to be either cross-entropy (CE) or mean squared error (MSE), as required for the MixMin algorithm which we adapt (Thudi et al., 2025). Our goal is to learn sampling weights $\lambda \in \Delta_k$ over the public datasets. Operationally, we sample an index $i \sim \lambda$ and then a data point from D_i^{pub}, giving samples from $d_\lambda = \sum_{i=1}^k \lambda_i\, d_i$.

To simplify data mixing formulations, we assume no covariate shift among public sources: $d_i(x) = d_j(x)$ for all i, j. This assumption is used only to simplify how we frame the mixture problem, and it is also trivially satisfied in our experiments (i.e., generative tasks have no "covariates", and disease classification is done over the same distribution of images). In general, if there is covariate shift, one should use a more general data mixing objective that optimizes the joint distribution over the inputs and outputs to be the same as the target distribution (see Section 6.1 in Thudi et al. (2025)).

Data mixing as bi-level optimization. The data mixing problem asks which $\lambda \in \Delta_k$ yields a model trained on d_λ that performs best on d_t. This is a bi-level optimization,

$$\min_{\lambda \in \Delta_k} \; \mathbb{E}_{(x,y)\sim d_t}\big[L(f_\lambda(x),\, y)\big] \quad \text{where} \quad f_\lambda \in \arg\min_{f \in \mathcal{H}} \; \mathbb{E}_{(x,y)\sim d_\lambda}\big[L(f(x),\, y)\big], \tag{1}$$

that is generally hard to solve: the inner minimization over $\mathcal{H}$ typically has to be run for many candidates λ. Beyond computational complexity, we note that private hyperparameter search (which optimizes this bi-level problem with random search) also has worse privacy-utility trade-offs.

MixMin: a convex reduction. Thudi et al. (2025) showed that under the assumptions above, as our model f_λ approaches Bayes optimal for d_λ (i.e., large compute and data limit), eq. (1) reduces to a single convex problem in λ. Letting f_i be the Bayes optimal model for each d_i, the optimal λ solves:

$$\min_{\lambda \in \Delta_k} \; \mathbb{E}_{(x,y)\sim d_t} L\big(\sum_{i=1}^{k} \lambda_i\, f_i(x),\; y\big). \tag{2}$$

In practice, it was found that running this optimization with empirically trained proxy models $\hat{f}_i$ still outperformed other mixture-finding methods (Thudi et al., 2025).

4 Differentially Private Data Mixing

In this section we first highlight how privately optimizing the reduced data mixing objective (eq. (2)) leads to better utility-privacy trade-offs than private hyperparameter search. We then describe our empirical method for privately optimizing the reduced objective, which we call DP-MixMin.

4.1 Improved Utility over Private Hyperparameter Search Asymptotically

Privately optimizing the MixMin reduction leads to asymptotically better privacy-utility trade-offs than private hyperparameter search using Poisson sampling (Papernot & Steinke, 2022) as we increase the number of public datasets k. We prove that there are cases where private hyperparameter search requires the DP parameter ϵ to grow at least linearly with k, for a fixed error. In contrast, existing DP stochastic convex optimization (SCO) algorithms applied to the MixMin reduction only require the DP guarantee to grow with $\sqrt{k}$, meaning we can achieve better privacy at a given utility. We state this formally in Lemma 4.1, where we suppress constants that do not depend on k.

We quantify utility by the excess error of the recovered mixture $\hat{\boldsymbol{\lambda}}$ relative to the optimal mixture:

$$\alpha \;:=\; \mathbb{E}\big[F(\hat{\boldsymbol{\lambda}})\big] - F(\boldsymbol{\lambda}^\star), \qquad F(\boldsymbol{\lambda}) \;=\; \mathbb{E}_{(x,y)\sim d_t}\Big[L\Big(\textstyle\sum_{i=1}^k \lambda_i f_i(x),\, y\Big)\Big], \tag{3}$$

where F is the MixMin objective of eq. (2), $\boldsymbol{\lambda}^\star = \arg\min_{\boldsymbol{\lambda}\in\Delta_k} F(\boldsymbol{\lambda})$ is its minimizer.

Lemma 4.1. *Suppose the conditions for the MixMin convex data mixing reduction hold (eq. (2)). There exists a Lipschitz data mixing objective such that private hyperparameter search with Poisson sampling requires $\epsilon = \Omega(k)$ to achieve a fixed (sufficiently small) excess error α with constant probability. In contrast, there exist DP-SCO algorithms that, for any Lipschitz data mixing objective, have $\epsilon = O(\sqrt{k})$ to achieve excess error α, for mixture-learning data size $n = \Omega(1/\alpha^2)$.*

The key idea of the proof is to consider a specific Lipschitz data mixing problem where finding a good mixture is equivalent to finding a mixture within an ℓ_2 ball around the best mixture. We then show the probability of uniformly sampling a mixture $\lambda \in \Delta^k$ that is also in this ball vanishes exponentially with k, which implies private hyperparameter search needs to try many random mixtures and hence use larger privacy budget ϵ. In contrast, we are able to apply existing DP-SCO results to any reduced Lipschitz data mixing objectives to show we can achieve much less error at lower privacy budgets. The full proof is in Appendix A.1. In Appendix A.6 we empirically show our proposed DP data mixing algorithm (Algorithm 1) follows an $O(\sqrt{k})$ excess error rate.

While Lemma 4.1 shows optimizing eq. (2) should lead to better utility, this assumes we can train models close to Bayes optimal. We turn to how we empirically approximate eq. (2).

4.2 Implementing DP-MixMin

We now explain how we empirically optimize the MixMin objective (eq. (2)), train the proxy models $\hat{f}_i$ which we use instead of the Bayes optimal f_i, and how we allocate the privacy budget for private mixture finding and finetuning. The overall pipeline is summarized in Figure 1.

Making MixMin Differentially Private. To empirically optimize eq. (2) with DP, we replace the Bayes optimal models f_i with models $\hat{f}_i$ trained empirically on D_i^{pub}, and the expectation over d_t with the empirical expectation over a dataset $D_\lambda^{priv} \sim d_t$. To now optimize this we run a post-processing of the DP-SGD (Abadi et al., 2016) update: for each iteration, per-example gradients w.r.t λ are computed, clipped to a fixed norm, and perturbed with Gaussian noise before being aggregated. The resulting noisy gradient is then passed to the entropic-descent update. The overall procedure satisfies (ϵ, δ)-DP by the DP guarantees of the sampled Gaussian mechanism (Mironov et al., 2019), post-processing (Dwork & Roth, 2014), and composition theorems to accumulate DP guarantees

Algorithm 1 DP-MixMin

Require: Target privacy guarantees (ϵ,δ), step size η, number of steps n, loss function L (CE or MSE), private mixture-learning split $D_\lambda^{priv} \subseteq \mathcal{X} \times \mathcal{Y}$, proxy models $\{\hat{f}_i : i = 1, \ldots, k\}$, clipping norm C

Initialize $\lambda_i \leftarrow \frac{1}{k}$ for all $i = 1, \ldots, k$
Pre-compute $\{\hat{f}_i(x) : x \in D_\lambda^{priv},\ i = 1, \ldots, k\}$
for $j = 1$ to n **do**
  $\hat{f}_\lambda(x) \leftarrow \sum_{i=1}^k \lambda_i \hat{f}_i(x)$
  $\ell(x, y) \leftarrow L(\hat{f}_\lambda(x), y)$ for all $(x, y) \in D_\lambda^{priv}$
  $g(x, y) \leftarrow \nabla_\lambda \ell(x, y)$
  $\tilde{g}(x, y) \leftarrow \text{Clip}(g(x, y), C)$ {clip per-example gradients}
  $g_{\text{sum}} \leftarrow \sum_{(x,y) \in D_\lambda^{priv}} \tilde{g}(x, y)$
  $z \sim \mathcal{N}(0, \sigma^2 I)$ {noise calibrated to (ϵ, δ)}
  $\bar{g} \leftarrow \frac{1}{|D_\lambda^{priv}|}(g_{\text{sum}} + z)$
  $\lambda_i \leftarrow \lambda_i \, e^{-\eta \bar{g}_i}$ for all $i = 1, \ldots, k$
  $\lambda \leftarrow \lambda \,/\, \sum_{i=1}^k \lambda_i$ {project onto Δ_k}
end for
Return λ

across steps (Mironov, 2017) [3]. See Algorithm 1, with blue text describing changes to the non-private optimization described by Thudi et al. (2025). See Algorithm 2 in the Appendix for user-level DP.

Reducing Proxy Model Costs. Past work on empirically optimizing eq. (2) required training a separate proxy model on each public dataset (Thudi et al., 2025). We reduced costs for classification tasks by training a single multi-class proxy model to jointly predict the labels for each public dataset: $\hat{f} : \mathcal{X} \to \mathcal{O}^k$. To do so, we concatenated the label spaces of all k public datasets (representing labels as one-hot or multi-hot vectors depending on the datasets), and trained the proxy model to output a probability vector over this combined label space. The proxy model for the i-th dataset is then obtained as $\hat{f}_i(x) = \hat{f}(x)[i]$, i.e., the i-th coordinate of the output. If learned to Bayes optimality, this proxy would provide the same output probabilities as a model trained only on D_i^{pub}. We found this significantly reduced training and storage costs when scaling the number of public datasets to mix. For generative language modeling, we followed the standard procedure of training an individual model on each public dataset. Further details are provided in Section 5.2.

Privacy budget allocation. We split the private target dataset D^{priv} into two disjoint subsets: D_λ^{priv}, which is used to compute the mixture weights (Algorithm 1), and D_{ft}^{priv}, which is used to fine-tune the pre-trained model. By parallel composition (Dwork & Roth, 2014), if the mechanism used for mixture learning is (ϵ_1, δ_1)-DP and the mechanism used for finetuning is (ϵ_2, δ_2)-DP, then the overall guarantee is: $(\epsilon, \delta) = \big(\max(\epsilon_1, \epsilon_2), \max(\delta_1, \delta_2)\big)$. Note that the baseline of pre-training on the uniform mixture uses all of D^{priv} for finetuning; we found that DP-MixMin improves over uniform mixtures, showing that we can use less data for finetuning if we optimize the mixture.

5 Experiments

We compared the full DP-MixMin pipeline—private mixture learning, mixture-weighted public pretraining, and DP finetuning—against three baselines:

1. **Random Initialization**: privately fine-tune with no public pretraining, isolating the impact of public pretraining.
2. **Uniform Pretraining**: pretrain with equal weight on every public source, isolating the value of learning a target-specific public-data mixture.

[3] In particular, we use the DP accountant for DP-SGD provided by Opacus (Yousefpour et al., 2021)

3. **Random Search** (non-private): select the best mixture from a random sample of candidate mixtures (Bergstra & Bengio, 2012; Papernot & Steinke, 2022) by training a proxy model on each one and evaluating performance on the downstream task. The selection of the best mixture from the random sample was done non-privately, so this over-estimates the performance of private hyperparameter search. Full details are in Appendix A.10.

We evaluated these methods' downstream utility on the private dataset, across privacy budgets, sources of public data, and modalities. We also compare DP-MixMin to non-private MixMin.

5.1 Image Classification

We evaluated DP-MixMin for chest X-ray disease classification. Prior work applied DP-SGD to train models from scratch on these datasets (Tayebi Arasteh et al., 2024; Gu et al., 2025), which corresponds to our random initialization baseline. We found mixtures of other chest X-ray datasets (for different diseases) so that pretraining on them improves performance.

5.1.1 Experimental Setup

Datasets and Models. We considered two different collections of pretraining datasets: (1) CheXpert (Irvin et al., 2019) and (2) MIMIC-CXR (Johnson et al., 2019b), both standard public sources for chest X-ray transfer learning (Rajpurkar et al., 2017; Gu et al., 2025). Note each of these datasets also consist of 14 chest X-ray disease classification tasks, but differ in the diseases and how the images were labelled. We used a ResNet-50 as the proxy model and a randomly initialized ResNet-18 for downstream pretraining and finetuning. The smaller downstream model reduced the cost of pretraining, allowing us to run more experiments within our compute budget.

Data preprocessing and computational usage For all three chest X-ray datasets, CheXpert, MIMIC-CXR, and NIH, uncertain labels (-1) were mapped to 0 (negative); 0 and 1 were kept as-is. The total compute used was $\sim$ 20,000 GPU hours over a cluster of L40 GPUs.

Methodology. For each of the 14 downstream tasks, 80% of the NIH data was used for training and 20% was split between validation and test. DP-MixMin and Random Search accessed the private task only through parallel composition, using 10% of the training split for mixture learning and the rest for finetuning. Uniform pretraining used all of the training split for DP finetuning. We repeated each method 5 times with a different random partition of the NIH data into mixture-finding, finetuning, validation, and test sets, as well as a different random seed for model initialization and DP noise. See Appendix A.3.1 for details on training hyperparameters. We write $\rho = |D_\lambda^{\text{priv}}|/|D^{\text{priv}}|$ for the fraction used for private mixture-learning: see Appendix A.2 for guidance on choosing it.

5.1.2 Image Classification Results

Improvements were consistent across downstream tasks and privacy budgets. Across the 14 NIH tasks, we found DP MixMin always improved or was on-par (within one standard deviation) with the baselines of uniform pretraining and random search (Figure 3). In the case of Cardiomegaly, DP-MixMin's gain over Random Init was roughly $5\times$ Uniform's at $\epsilon = 1$.

Gains were consistent across privacy budgets DP-MixMin also outperformed the baselines at every privacy budget we tested: $+0.037, +0.032, +0.029$ absolute macro AUC gain over uniform at $\epsilon = 1, 2, 5$ for CheXpert pretraining(Figure 2), and $+0.033, +0.028, +0.024$ for MIMIC-CXR. Tables 7 and 8 in the Appendix report absolute AUC across settings.

5.2 Language Models

We evaluated DP-MixMin on two downstream tasks: Enron (Klimt & Yang, 2004a) email completion with user-level DP and OpenWebMath (Paster et al., 2023) with example-level DP. For Enron we compared to uniform pretraining and random search, and for OpenWebMath we only compared to uniform as it had often outperformed the other baseline (see Figure 3 and Table 1).

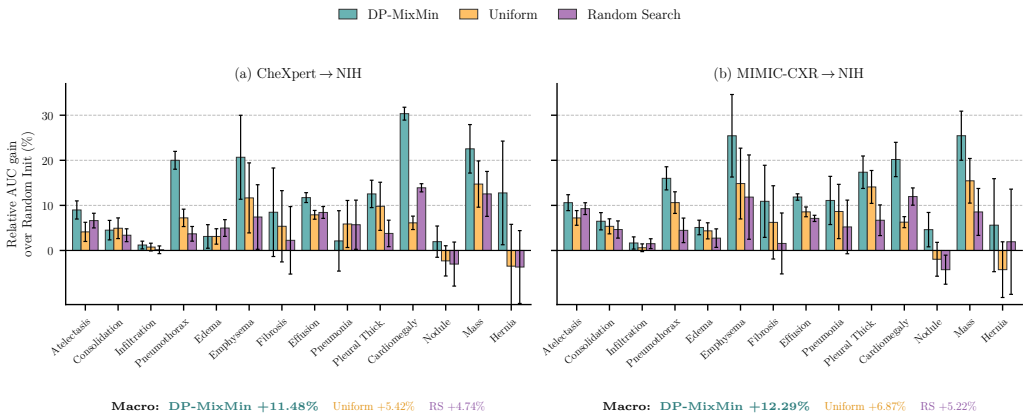


Figure 3: **Per-task relative AUC gain over Random Init on NIH Chest X-ray at $\epsilon = 1$.** Bars show $\big(\mathrm{AUC}(\mathrm{method}) - \mathrm{AUC}(\text{Random Init})\big)/\mathrm{AUC}(\text{Random Init})$ as a percentage, with propagated standard deviation over 3–5 seeds. The macro AUC gain (unweighted mean across 14 tasks) is reported next to each subplot title; DP-MixMin roughly doubles, and in most cases more than doubles, the gain of both Uniform and Random Search on each public pretraining source. Full raw AUCs are reported in Tables 7 and 8.

Table 1: **Enron test perplexity (mean ± std) after DP finetuning at δ=10^{-5}.** DP-MixMin beats both Uniform and Random Search at every privacy budget. At ϵ=1, the mixture quality (and stability) improves monotonically with the mixture weight finding fraction ρ.

Method	ρ	ϵ=1	ϵ=2	ϵ=5
DP-MixMin	0.3	$\mathbf{16.78 \pm 0.55}$	$\mathbf{16.44 \pm 0.95}$	$\mathbf{15.71 \pm 0.77}$
DP-MixMin	0.2	17.52 ± 1.55	—	—
DP-MixMin	0.1	17.97 ± 2.72	—	—
Uniform	—	19.88 ± 1.10	19.18 ± 1.01	18.24 ± 0.89
Random Search	0.3	20.46 ± 0.70	19.72 ± 0.66	18.72 ± 0.62

5.2.1 Experimental Setup

Datasets and Models. We used 16 domains from The Common Pile (Kandpal et al., 2025) as candidate public pretraining sources, selected from the 18 domains containing at least 10GB of data [4]. For each domain we trained a GPT-2 Small (124M parameters) proxy on documents with at least 1024 GPT-2 tokens until roughly 45M tokens had been seen. The downstream model was GPT-2 Medium (355M parameters). DP finetuning was only done on the last two layers and output head of the model, due to memory allocation constraints on our GPUs.

Data preprocessing and computational usage. For Enron, we restricted to the `_sent_mails` and `sent_mails` folders, yielding 148 users and approximately 80,000 emails. This avoided multi-author records such as received emails, forwarded messages, and duplicated threads, which require more complex multi-attribution privacy treatment (Ponomareva et al., 2025). For OpenWebMath, we streamed 50,000 samples to form the dataset. The total compute was $\approx 5{,}000$ L40 GPU-hours.

Methodology For Enron, we ablated $\rho = |D_\lambda^{\mathrm{priv}}|/|D^{\mathrm{priv}}|$ between $\{10\%, 20\%, 30\%\}$. For OpenWebMath, DP-MixMin used 10%. See Appendix A.3.2 for further details.

5.2.2 Language Results

DP-MixMin's utility advantage grew under tighter ϵ: As shown in Table 1, the relative perplexity reduction of DP-MixMin over the baselines widened as ϵ shrank on Enron: 13.9% at ϵ=5, 14.3% at

[4] Two were excluded due to streaming limitations.

Table 2: **OpenWebMath test perplexity (mean ± std) after DP finetuning at δ=10^{-5}.** DP-MixMin lowers test perplexity by roughly 13% over uniform Common Pile pretraining at ϵ=1. The uniform baseline is essentially flat in ϵ on this domain, showing DP-MixMin at $\epsilon = 1$ is still better.

Method	ρ	ϵ=1	ϵ=2	ϵ=5
DP-MIXMIN	0.1	$\mathbf{12.22 \pm 0.17}$	—	—
UNIFORM	—	14.05 ± 0.17	14.04 ± 0.17	14.02 ± 0.17

ϵ=2, and 15.6% at ϵ=1. On OpenWebMath, we found the performance of uniform mixture weights was stable across $\epsilon = 1, 2, 5$, and DP-MixMin improved over it at $\epsilon = 1$ (Table 2).

Public pretraining mixture mattered more than amount of finetuning data. While each of the mixtures at ratios $\rho \in 0.1, 0.2, 0.3$ outperformed uniform pretraining, allocating private data toward mixture estimation mattered more than reserving it for finetuning for Enron. The proxy-fraction ablation in Table 1 exhibited monotonic improvement with ρ in both mean perplexity ($17.97 \to 17.52 \to 16.78$) and cross-seed standard deviation ($2.72 \to 1.55 \to 0.45$).

5.3 COMPARING DP-MIXMIN TO NON-PRIVATE MIXMIN

Preventing privacy leakage. We audited whether the data mixture λ leaked private information in practice, following the canary-based auditing framework of Jagielski et al. (2020) using both natural and worst-case (max-clip) canaries. Across all 14 NIH tasks and both pretraining sources, every measured ϵ_{LB} stayed far below the claimed $\epsilon = 1$, with the highest being 0.131. Removing the DP noise from the same audit recovered ϵ_{LB} as high as 5.6, confirming that this step is genuinely privacy-sensitive and that DP is necessary (Tables 16, 17; further details in Appendix A.8).

DP-MixMin algorithm often recovered non-private MixMin weights. To understand the impact of DP on mixture finding, we compared the DP-MixMin weights $\boldsymbol{\lambda}_{\mathrm{DP}}$ to the non-private MixMin $\boldsymbol{\lambda}^\star$ (which used the entire D^{priv} training split for mixture finding). Across modalities, DP-MixMin recovered mixtures close to $\boldsymbol{\lambda}^\star$ even with small privacy budgets and limited private data (Tables 11, 14, 15 in the Appendix). By $\epsilon = 1$, DP-MixMin achieved cosine similarity of 0.97 and 0.93 for CheXpert→NIH and MIMIC-CXR→NIH respectively. On Enron, DP-MixMin reached cosine similarity 0.92 having seen only 12 users at $\epsilon = 2$ user-level DP. On OpenWebMath, using only 2.8M tokens ($\rho = 0.1$), DP-MixMin achieved cosine similarity above 0.9 at $\epsilon = 0.2$.

DP-MixMin's downstream performance was on-par with non-private MixMin. Given the similarity in the weights, we further checked whether it translated to matched downstream utility. We compared the full DP-MixMin pipeline at $\varepsilon = 1$ to a non-private MixMin baseline that paid no privacy for mixture finding, and also reused the *entire* private dataset (both the mixing and finetuning splits) for weight-finding and DP finetuning steps. Across both pretraining sources, DP-MixMin was comparable to non-private MixMin: the macro AUC difference between DP-MixMin and non-private MixMin was only 0.717 vs. 0.723 for MIMIC-CXR and 0.711 vs. 0.712 for CheXpert. Most per-class differences lay within one standard deviation (Tables 19, 18).

CONCLUSION

In this paper, we initiated the study of privately optimizing public-data pretraining mixtures for a target sensitive dataset. Across medical imaging and language modeling, DP-MixMin improved private downstream utility over uniform pretraining and random mixture search, with large gains under low-prevalence labels, tight privacy budgets, and user-level privacy. Extending DP-MixMin to tabular and other domain settings remains important future work. In particular, public tabular sources often differ in their input features, violating the no-covariate-shift assumption our reduction relies on (Section 3). Future work could also look into how to improve our current DP-MixMin pipeline. Initial results showed our approach was robust to less proxy models training (Appendix A.5) and whether parallel or sequential composition was used (Appendix A.7). However, the theoretical results of Bassily et al. (2021) suggest using DP Franke-Wolfe instead of DP mirror descent.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks requiring disclosure. We only used generative AI tools for formatting the calculations presented in Appendix A.2, debugging code, formatting the figures and tables in this paper, and checking if there was literature on optimizing public pretraining data we missed. All code and references were verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H 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*, pp. 308–318, 2016.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Daogao Liu, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. Mind the privacy unit! user-level differential privacy for language model fine-tuning. *arXiv preprint arXiv:2406.14322*, 2024.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Ricardo Dominguez-Olmedo, Florian E Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. *arXiv preprint arXiv:2407.07890*, 2024.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? *arXiv preprint arXiv:2302.09483*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Xin Gu, Gautam Kamath, and Zhiwei Steven Wu. Choosing public datasets for private machine learning via gradient subspace distance, 2025. URL https://arxiv.org/abs/2303.01256.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. URL https://arxiv.org/abs/1901.07031.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd?, 2020. URL https://arxiv.org/abs/2006.07709.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019a.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019b. URL https://arxiv.org/abs/1901.07042.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. User inference attacks on large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 18238–18265, 2024.

Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, John Kirchenbauer, Shayne Longpre, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben Allal, Elie Bakouch, John David Pressman, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, Tom Goldstein, Brian R. Bartoldson, Bhavya Kailkhura, and Tyler Murray. The common pile v0.1: An 8tb dataset of public domain and openly licensed text, 2025. URL https://arxiv.org/abs/2506.05209.

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi (eds.), *Machine Learning: ECML 2004*, pp. 217–226, Berlin, Heidelberg, 2004a. Springer Berlin Heidelberg. ISBN 978-3-540-30115-8.

Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *CEAS*, volume 45, pp. 92–96, 2004b.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=BJ0hF1Z0b.

Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 19–30, 2009.

Harsh Mehta, Abhradeep Thakurta, Alexey Kurakin, and Ashok Cutkosky. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. In *International Conference on Learning Representations*, 2022.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text, 2023. URL https://arxiv.org/abs/2310.06786.

Natalia Ponomareva, Zheng Xu, H. Brendan McMahan, Peter Kairouz, Lucas Rosenblatt, Vincent Cohen-Addad, Cristóbal Guzmán, Ryan McKenna, Galen Andrew, Alex Bie, Da Yu, Alex Kurakin, Morteza Zadimoghaddam, Sergei Vassilvitskii, and Andreas Terzis. How to dp-fy your data: A practical guide to generating synthetic data with differential privacy, 2025. URL https://arxiv.org/abs/2512.03238.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas P Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496, 2024.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.org/abs/1711.05225.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Soroosh Tayebi Arasteh, Mahshad Lotfinia, Teresa Nolte, Marwin-Jonathan Sähn, Peter Isfort, Christiane Kuhl, Sven Nebelung, Georgios Kaissis, and Daniel Truhn. Securing collaborative medical AI by using differential privacy: Domain transfer for classification of chest radiographs. *Radiology: Artificial Intelligence*, 6(1):e230212, 2024. doi: 10.1148/ryai.230212.

Anvith Thudi, Evianne Rovers, Yangjun Ruan, Tristan Thrush, and Chris J. Maddison. Mixmin: Finding data mixtures via convex minimization, 2025. URL https://arxiv.org/abs/2502.10510.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471. IEEE, 2017. doi: 10.1109/cvpr.2017.369. URL http://dx.doi.org/10.1109/CVPR.2017.369.

Sang Michael Xie, Hieu Pham, Xinyun Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*, 2023.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

Da Yu, Sivakanth Gopi, Janardhan Kulkarni, Zinan Lin, Saurabh Naik, Tomasz Lukasz Religa, Jian Yin, and Huishuai Zhang. Selective pre-training for private fine-tuning, 2024. URL https://arxiv.org/abs/2305.13865.

A Appendix

A.1 Proof of Lemma 4.1

Proof. We first provide a specific data mixing problem (under the lemma assumptions) that requires private hyperparameter search to have $\epsilon = \Omega(k)$ for fixed excess error with constant probability. We then show, under the lemma assumptions, the reduced MixMin objective always has $\epsilon = O(\sqrt{k})$.

To show worst-case scaling for private hyperparameter search, we need only find one "hard" example. Consider the case where the data mixture objective is $\|\lambda - \lambda^*\|_2^2$ for λ^* in the relative interior of the simplex (i.e., $\lambda_i^* \neq 0$ for any i). This corresponds to taking L to be MSE, $f_i(x) = e_i$ for all x, and $y = \lambda^*$ always (which hence defines the target distribution). Then for random search to achieve α error, we need $\|\lambda - \lambda^*\|_2 \leq \sqrt{\alpha}$. Note the probability density function for uniform sampling on the simplex is $(k-1)!$ (where k is the number of public datasets we are mixing). Thus the probability of a random mixture being in $S_\alpha = \{\lambda : \|\lambda - \lambda^*\|_2 \leq \sqrt{\alpha}, \lambda \in \Delta^k\}$ is $(k-1)!\, Vol(S_\alpha) = \frac{(k-1)!\,\pi^{(k-1)/2}}{\Gamma((k-1)/2+1)}\alpha^{(k-1)/2}$ for α sufficiently small. This follows from noting that in a sufficiently small ball around any point in the interior of the simplex, the simplex is a hyperplane defined by $\langle \mathbf{1}, \lambda \rangle = 1$, and the intersection of a hyperplane around a k-dim sphere centered at a point on it is yet another sphere of dimension $k-1$ and the same radius (by rotational symmetry).

Now recall Poisson hyperparameter search for mixture finding first samples a number r following the Poisson distribution, and then trains on r random mixtures and picks the best. As $\alpha \to 0$, we have the probability that the closest of the r random mixtures is in S_α, where r is also a random variable, is $\Theta\Big(\mathbb{E}r\, \frac{(k-1)!\,\pi^{(k-1)/2}}{\Gamma((k-1)/2+1)}\, \alpha^{(k-1)/2}\Big)$ (see Appendix A.2.1 in Papernot & Steinke (2022) for the first-order approximation which becomes tight as $\alpha \to 0$). Hence for the probability to not vanish with k, we need $\mathbb{E}r = \Omega\Big(\big(\frac{1}{\alpha}\big)^{(k-1)/2}\Big)$.

Plugging this into Theorem 6 in Papernot & Steinke (2022) shows the ϵ guarantee for (ϵ, δ) grows $\propto \log(\mathbb{E}r) = \Omega(k)$.

In contrast, the MixMin reduction presents a Lipschitz DP-SCO objective. It is known that an achievable expected excess error vs. privacy trade-off is $\alpha = O(\frac{1}{\sqrt{n}} + \frac{\sqrt{k \ln(1/\delta)}}{\epsilon n})$ (Bassily et al., 2019;

Feldman et al., 2020; Asi et al., 2021) [5]. We hence need $\epsilon = O(\sqrt{k}/(n\alpha)) = O(\sqrt{k} \cdot \alpha) = O(\sqrt{k})$, noting $n = \Omega(1/\alpha^2)$ by assumption and $\alpha \leq 1$. Note that by Lipschitzness and the fact the simplex has ℓ_2 diameter bounded by 1, the loss of the mixture is always bounded by a constant independent of k, and so always has at least constant probability around the expectation (independent of k), and we can conclude $O(\sqrt{k})$ is sufficient for excess error less than α with constant probability. □

A.2 Guidance on choosing the mixture-selection budget ρ

DP-MixMin sets aside a fraction ρ of the private data to estimate the mixture weights. In the main text we validate a choice of ρ *a posteriori*, by checking that the privately recovered mixture $\boldsymbol{\lambda}_{\mathrm{DP}}$ has high cosine similarity to the non-private solution $\boldsymbol{\lambda}^\star$ (Tables 11–13). Here we give a complementary *a-priori* guideline, derived directly from our error bound (Lemma 4.1), that tells a practitioner how large ρ must be before running anything.

Setup. We measure quality by the excess error α of eq. (3), here instantiated with the empirically trained proxies $\hat{f}_i$ in place of the Bayes-optimal f_i. Write $T_\lambda = \rho N$ for the number of private examples used for mixing, with N the private set size and k the number of public sources.

An a-priori bound on ρ. The DP-SCO trade-off for this objective (Lemma 4.1) is $\alpha = O\big(1/\sqrt{T_\lambda} + \sqrt{k\ln(1/\delta)}/(\epsilon\, T_\lambda)\big)$. Requiring each term to contribute at most $\alpha/2$ yields a sufficient budget:

$$\alpha = O\left(\frac{1}{\sqrt{T_\lambda}} + \frac{\sqrt{k\ln(1/\delta)}}{\epsilon\, T_\lambda}\right), \tag{4}$$

$$\text{Requiring each term} \leq \alpha/2: \quad \frac{1}{\sqrt{T_\lambda}} \leq \frac{\alpha}{2} \;\Rightarrow\; T_\lambda \geq \frac{4}{\alpha^2}, \tag{5}$$

$$\frac{\sqrt{k\ln(1/\delta)}}{\epsilon T_\lambda} \leq \frac{\alpha}{2} \;\Rightarrow\; T_\lambda \geq \frac{2\sqrt{k\ln(1/\delta)}}{\epsilon\alpha}, \tag{6}$$

$$\therefore \quad T_\lambda \;\geq\; \max\left(\frac{4}{\alpha^2},\; \frac{2\sqrt{k\ln(1/\delta)}}{\epsilon\alpha}\right), \qquad \rho = \frac{T_\lambda}{N}. \tag{7}$$

This is a guideline rather than a single universal constant. The required budget is governed by two regimes: a *statistics-limited* regime where $T_\lambda \gtrsim 4/\alpha^2$ (the cost of estimating the objective from finite data), and a *privacy-limited* regime where $T_\lambda \gtrsim 2\sqrt{k\ln(1/\delta)}/(\epsilon\alpha)$ (the cost of the DP noise). The two cross over at $\epsilon^\star = \frac{\alpha}{2}\sqrt{k\ln(1/\delta)}$: for $\epsilon > \epsilon^\star$ the statistical term dominates and ρ is independent of the privacy budget; for $\epsilon < \epsilon^\star$ the required ρ grows as $1/\epsilon$.

Reading the guideline at the budgets we use. For $\epsilon \geq 1$ we operate in the statistics-limited regime (e.g. for our vision setting $k = 14$, $\delta = 10^{-5}$ give $\epsilon^\star \approx 0.63 < 1$), so the achievable error is of order $\alpha \approx 1/\sqrt{T_\lambda}$. Evaluated at $\rho = 0.1$, this single relation reproduces both of our empirical operating points:

- **Vision (per-example DP).** $\rho = 0.1$ gives $T_\lambda \approx 8{,}970$ examples, so $\alpha \approx 1/\sqrt{T_\lambda} \approx 0.01$, consistent with the strong recovery we observe (Table 11: cosine similarity 0.97 at $\epsilon = 1$ across the 14 NIH tasks).
- **Language (example-level DP).** $\rho = 0.1$ gives $\approx$ 2.8M OpenWebMath tokens, so $\alpha \approx 1/\sqrt{T_\lambda} \approx 0.001$, consistent with the near-exact recovery we observe (Table 15: cosine similarity 1.00 at $\epsilon = 1$).

The achievable α therefore tracks data availability: with millions of tokens, language reaches a far tighter α than vision, whose mixture-learning splits contain only thousands of examples. In both cases $\rho = 0.1$ clears the threshold of eq. (7), which is why it is a robust default; a practitioner facing a new dataset can invert the same relation to size ρ for a target α.

[5] In fact, in the specific case of generalized linear losses, this can be far better in dimension dependence (Bassily et al., 2021)

A.3 Extended Experimental Details

A.3.1 Image Classification

DP-MixMin Implementation. Proxy models were trained for 4 and 10 epochs respectively for CheXpert MIMIC-CXR at learning rate 10^{-4} and batch size 256 (the largest batch size that fit on a single L40 GPU). DP-MixMin used clipping norm 13.0, selected from $\{1.0, 2.0, 5.0, 10.0, 13.0, 15.0, 20.0\}$ with a batch size of 1000 selected from $\{500, 1000, 1500\}$, and learning rate chosen from $\{0.5, 1.0, 2.0\}$.

Training details. Pretraining and private fine-tuning on the ResNet18s both used batch size 200 (the largest batch that fit on an L40). Given mixture weights $\lambda \in \Delta^{14}$, we pretrained the ResNet-18 by minimizing the binary cross-entropy loss between its scalar output and the mixture-weighted target $\sum_{c=1}^{14} \lambda_c \, y_c$, where y_c was the per-disease binary label. [6] We evaluated the pretrained ResNet-18 at 3, 6, and 9 epochs and selected the checkpoint with the highest downstream validation AUC before fine-tuning. Fine-tuning ran for 36,000 steps, with learning rate 10^{-4} and DP-SGD clipping norm 2.0. All hyperparameters were tuned once on a single pilot configuration—Atelectasis classification as the NIH downstream task with CheXpert pretraining at $\varepsilon = 1$—and then held fixed across all 14 tasks, both pretraining sources, all seeds, and all three privacy budgets $\varepsilon \in \{1, 2, 5\}$ at $\delta = 10^{-5}$.

A.3.2 Language Modeling

DP-MixMin Implementation. For Enron, DP-MixMin was run for 200 steps with per-sample clipping norm 0.01 and step size 5.0. For OpenWebMath, DP-MixMin was run for 200 steps with clipping norm 0.2 and step size 5.0.

Training details. Pretraining uses GPT-2 Medium from random initialization for 7.1×10^9 tokens at context length 1024, effective batch size 64, and learning rate $5 \cdot 10^{-3}$. Enron fine-tuning uses 5 users per gradient step with 20 accumulation steps (effective batch of 100 users), G_ULS of 15, context length 512, and learning rate $5 \cdot 10^{-4}$ (chosen as the largest stable rate from $\{5 \cdot 10^{-5}, 5 \cdot 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}\}$). Per-device batch sizes were set to the largest power of 2 that fit in available GPU memory, with gradient accumulation used to reach the target effective batch. Private Enron fine-tuning uses clipping norm 0.5 and a token budget of 25M. Private OpenWebMath fine-tuning used batch size 500 at learning rate $5 \cdot 10^{-3}$ for pretraining and a 50M-token fine-tuning budget. All results are averaged over 3 seeds.

A.4 On the infeasibility of GSD in our setting

Gu et al. (2025) propose Gradient Subspace Distance (GSD), which scores a candidate public source by training a model on randomly labeled examples from both the public source and the private task, then measuring the distance between their gradient subspaces. The lowest-distance source is selected.

GSD is inapplicable to our setting in Section 5.1. We binarize each of CheXpert and MIMIC-CXR along their 14 disease axes (Section 5.1), so all 14 candidate sources within a corpus share the same image set $\mathcal{X}$ and differ only in per-image labels. With random labels, gradient subspaces depend only on $\mathcal{X}$, so any two sources sharing inputs receive a GSD of zero, and the candidates DP-MixMin must rank are precisely such sources. GSD is, by construction, only informative when candidates have distinct input distributions, for example disjoint vision benchmarks of the kind on which Gu et al. (2025) evaluate it.

The coarsest GSD variant with signal in our setting selects between *whole corpora*: CheXpert versus MIMIC-CXR. Either choice is dominated by DP-MixMin, which improves over uniform pretraining on each corpus independently (Table 10). Even granting GSD an oracle picker over the two corpora, the resulting pipeline still falls short of DP-MixMin.

[6] Note that this is the importance weighting defined by the mixture, and it reduces variance relative to sampling y_c according to λ_c.

A.5 Robustness to Proxy Model Quality

Lemma 4.1 assumes proxy models near Bayes optimal. To check how much this matters in practice, we varied the number of epochs used to train the CheXpert proxy (1 to 6, versus 4 in our main experiments) and ran DP-MixMin for NIH Atelectasis at $\epsilon = 1$. Results were similar across all proxy budgets (Table 3), and all remained well above uniform pretraining (0.707 AUC), suggesting small changes in proxy quality have little effect on DP-MixMin's downstream performance.

Table 3: **DP-MixMin is insensitive to proxy training budget.** NIH Atelectasis AUC (mean $\pm$ std over 3 seeds) with CheXpert pretraining at $\epsilon = 1$.

Proxy epochs	Atelectasis AUC
1	0.744 ± 0.005
2	0.745 ± 0.005
4 (main experiments)	0.742 ± 0.006
6	0.749 ± 0.003

A.6 Empirically Validating Lemma 4.1

To show that Lemma 4.1 holds empirically under its assumptions, we analyze the excess error as a function of the number of public sources being mixed, k, and trace its trajectory as k increases.

Setup. We train ResNet-50 proxies on k ImageNet-1k classes (Russakovsky et al., 2015) ($k \in \{2, 4, 8, 16, 64, 128, 256, 512, 1000\}$, the k classes being the k sources) and run DP-MixMin on the first 10 binarized CIFAR-100 classes (Krizhevsky, 2009) at ε=1, δ=10^{-5} over all N=50,000 images.

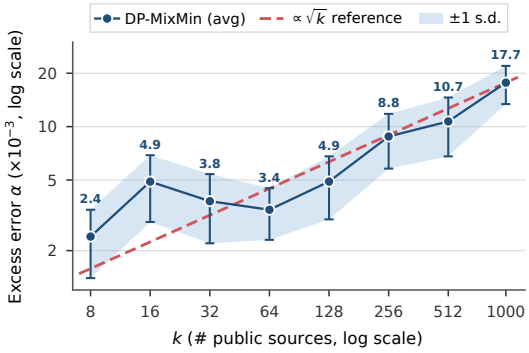


Figure 4: **Excess error scales roughly as $\sqrt{k}$ for large k.** Average excess error of DP-MixMin over the 10 CIFAR-100 tasks as the number of public sources k increases (both axes log scale). Error bars and the shaded band show ± 1 standard deviation across tasks. The dashed red line is a $\sqrt{k}$ reference; the fitted slope for $k \geq 32$ is 0.49, consistent with the predicted $\sqrt{k}$ rate.

Metric. We report the excess error $\alpha = L(\boldsymbol{\lambda}_{\mathrm{DP}}) - L(\boldsymbol{\lambda}^\star)$ (eq. (3)) against the non-private optimum $\boldsymbol{\lambda}^\star$, averaged over 5 seeds and the first 10 binarized CIFAR100 tasks

Result. Excess error grows sub-linearly in k (Table 4). A log-log fit over $k \geq 32$ gives slope 0.49, matching the $\sqrt{k}$ prediction (0.5) and well below the linear cost of private hyperparameter search.

A.7 Sequential vs. Parallel Composition

Instead of splitting D^{priv} into disjoint D^{priv}_{λ} and D^{priv}_{ft} (parallel composition), one can use all of D^{priv} for both steps and split the privacy budget, giving $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP by sequential composition (Dwork & Roth, 2014). We compared the two on NIH with CheXpert pretraining at total $\epsilon = 1$: sequential used $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.9$, and parallel used our main setting ($\rho = 0.1$,

Table 4: **DP-MixMin excess error $\alpha = L(\boldsymbol{\lambda}_{\mathrm{DP}}) - L(\boldsymbol{\lambda}^\star)$ across the number of public sources k** (ImageNet-1k $\to$ CIFAR-100, ε=1, δ=10^{-5}), for all 10 CIFAR-100 one-vs-rest tasks. Mean $\pm$ std over 5 seeds, in units of 10^{-3}. The last column (**Avg**) averages over all 5 seeds and 10 tasks.

k	class 0	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9	**Avg**
2	0.1±0.1	0.1±0.1	0.1±0.2	0.1±0.2	0.1±0.1	0.1±0.2	0.1±0.1	0.1±0.2	0.1±0.1	0.1±0.2	**0.1±0.1**
4	0.1±0.0	0.0±0.0	0.0±0.0	0.1±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.1±0.0	**0.0±0.0**
8	1.1±0.1	1.4±0.2	1.3±0.2	3.7±0.2	3.3±0.2	2.1±0.1	2.3±0.2	2.8±0.1	4.1±0.2	2.1±0.1	**2.4±1.0**
16	5.7±0.4	7.0±0.2	3.3±0.4	2.9±0.3	2.4±0.4	7.0±0.6	4.2±0.3	2.8±0.4	5.8±0.5	7.8±0.5	**4.9±2.0**
32	2.8±0.2	6.7±0.5	3.4±0.3	2.0±0.1	1.9±0.1	3.4±0.2	3.7±0.2	2.7±0.3	5.8±0.3	5.3±0.3	**3.8±1.6**
64	3.8±0.2	2.1±0.3	3.7±0.2	3.4±0.3	2.6±0.3	4.4±0.3	4.4±0.2	5.0±0.5	1.8±0.1	2.3±0.3	**3.4±1.1**
128	5.9±0.5	1.6±0.3	5.7±0.3	5.2±0.4	6.4±0.4	2.7±0.2	7.4±0.2	6.8±1.0	3.8±0.4	3.6±0.7	**4.9±1.9**
256	12.2±0.7	2.7±0.3	9.2±0.3	10.6±0.3	11.5±0.3	4.6±0.4	10.9±0.2	10.5±0.7	7.9±0.6	8.3±0.9	**8.8±3.0**
512	8.6±0.4	6.7±0.7	4.5±0.5	16.2±0.3	10.1±0.6	10.3±0.4	13.7±0.4	15.7±0.9	13.9±0.5	7.0±1.0	**10.7±3.9**
1000	18.6±1.0	14.9±0.9	16.3±0.7	23.4±0.3	14.9±0.7	9.6±0.4	21.9±0.3	22.3±0.5	20.6±0.5	14.8±1.3	**17.7±4.3**

$\epsilon_1 = \epsilon_2 = 1$). Neither was consistently better, and both were close to non-private MixMin (Table 6). This is expected: DP noise in each step scales roughly with $1/(\epsilon \cdot n)$ (Appendix A.2), and with $\epsilon_1 = \rho\,\epsilon$ and $\epsilon_2 = (1-\rho)\,\epsilon$, sequential composition matches parallel composition's $\epsilon \cdot n$ in both steps, differing only in using somewhat more data per step.

Table 5: **Learned mixture is stable across proxy training budget.** Cosine similarity of the DP-MixMin mixture (early-stop, ε=1) at proxy epochs $\{1, 2, 6\}$ to the main E=4 mixture, per NIH class (mean $\pm$ std, 3 seeds). E=4 is the self-reference (1.000).

Class	**E1**	**E2**	**E6**
Atelectasis	0.868 ± 0.015	0.986 ± 0.003	0.972 ± 0.014
Consolidation	0.886 ± 0.004	0.941 ± 0.035	0.944 ± 0.012
Infiltration	0.895 ± 0.018	0.873 ± 0.046	0.848 ± 0.023
Pneumothorax	0.952 ± 0.019	0.976 ± 0.003	0.975 ± 0.003
Edema	0.978 ± 0.003	0.965 ± 0.003	0.887 ± 0.029
Emphysema	0.986 ± 0.006	0.988 ± 0.002	0.970 ± 0.009
Fibrosis	0.993 ± 0.005	0.989 ± 0.002	0.916 ± 0.024
Effusion	0.870 ± 0.020	0.884 ± 0.006	0.938 ± 0.006
Pneumonia	0.975 ± 0.003	0.984 ± 0.004	0.889 ± 0.008
Pleural Thickening	0.982 ± 0.011	0.988 ± 0.001	0.931 ± 0.013
Cardiomegaly	0.904 ± 0.004	0.975 ± 0.007	0.726 ± 0.067
Nodule	0.967 ± 0.010	0.972 ± 0.012	0.932 ± 0.031
Mass	0.934 ± 0.021	0.982 ± 0.004	0.974 ± 0.009
Hernia	0.983 ± 0.001	0.998 ± 0.000	0.914 ± 0.007

Table 6: **Sequential and parallel composition perform similarly.** NIH AUC (mean $\pm$ std over 3 seeds), CheXpert pretraining, total $\epsilon = 1$.

Class	**Sequential**	**Parallel**	**Non-private**
Atelectasis	0.740 ± 0.010	0.747 ± 0.005	0.746 ± 0.006

A.8 Auditing DP-MixMin's Privacy Guarantee

Threat model. We consider example-level (ϵ, δ)-differential privacy, where neighboring private datasets differ by the addition or removal of one example. We partition the private dataset as $D^{\text{priv}} = D_{\text{mix}} \sqcup D_{\text{ft}}$, where D_{mix} is used to privately learn the mixture weights λ, and D_{ft} is used for DP-SGD fine-tuning. Because the two mechanisms operate on disjoint subsets, parallel composition gives an end-to-end guarantee of (ϵ, δ)-DP. In our experiments, this is $(\epsilon, \delta) = (1, 10^{-5})$.

We assume a deliberately strong adversary who observes everything released by the pipeline, including all public data, the algorithm and its hyperparameters, the learned mixture weights λ, and the final model. The adversary additionally knows every private example except the target example whose membership it seeks to infer.

Audit scope and methodology. We audit the component novel to DP-MixMin, $M_{\text{mix}} : D_{\text{mix}} \to \lambda$. For an example in D_{mix}, all subsequent model outputs depend on that example only through λ and are therefore post-processing of λ. We consequently audit the released mixture weights directly.

Following the empirical privacy-auditing framework and algorithms of Jagielski et al. (2020), we repeatedly run the mechanism on neighboring datasets and estimate an empirical lower bound ϵ_{LB}. Each audit configuration is evaluated over 2000 independent runs. We use a mixture-finding split ratio of $\rho = 0.1$. Public sources are the 14 CheXpert classes (proxy predictions). We run DP-MixMin once per NIH target class, auditing 14 mixture-finding instances in total.

We evaluate two types of canaries. The **natural canary** is a real, unmodified private training example whose membership the adversary tries to infer. This captures realistic leakage of a genuine data point. The **max-clip canary** is a strong construction. Its per-example gradient attains the clipping norm, the maximum influence any single example can have under clipping, in a fixed direction aligned with the mixture's most-weighted sources, since the released mixture is most sensitive to perturbations of those sources.

Results. At the claimed privacy level $\epsilon = 1$, the natural-canary audit gives $\epsilon_{LB} = 0$ for all 14 NIH tasks. The stronger max-clipped-gradient canary gives $\max_{\text{task}} \epsilon_{LB} = 0.131$. Thus, all measured private lower bounds satisfy $\epsilon_{LB} \leq 0.131 < 1$.

When privacy noise is removed, the same audit obtains values as high as $\epsilon_{LB} = 5.60$, demonstrating that it can detect substantial leakage when present. Overall, we find no empirical evidence of a privacy violation. Full per-class results for both CheXpert and MIMIC-CXR pretraining are reported in Tables 16 and 17.

A.9 User-Level DP-MixMin

User-level DP protects the contribution of an entire user rather than a single example: two neighboring datasets may differ in all examples contributed by one user (Levy et al., 2021). For Enron, we therefore partition the private mixture-learning split by user, $D_\lambda^{priv} = \bigsqcup_{u=1}^{N} D_u$, where D_u contains the examples attributed to user u. At each DP-MixMin step, we sample a batch $\mathcal{U}_j$ of B users, compute one gradient per user by averaging the loss over that user's sampled examples, clip this per-user gradient, and add Gaussian noise before updating the mixture weights. This differs from example-level DP-MixMin, where clipping and noise are applied to individual examples.

The parameter G_{ULS} is the user-level subsample size. For each sampled user u, we use $k_u = \min(G_{\text{ULS}}, |D_u|)$ examples, so users with many emails cannot dominate a single update, while users with fewer than G_{ULS} emails contribute all available examples. Because the gradient is averaged within each user before clipping, the sensitivity is controlled at the user level: the clipped object is the user's aggregate contribution, not an individual email.

A.10 Random Search

Random Search is the standard hyperparameter optimization baseline across fields (Bergstra & Bengio, 2012; Papernot & Steinke, 2022; Thudi et al., 2025): rather than searching the space of configurations in a structured way, one simply draws a fixed number of candidates independently at random, evaluates each on a held-out validation set, and returns the best performer. Despite its simplicity, Random Search is known to be competitive with, and often outperform, grid search and many model-based optimizers, particularly in moderate-to-high-dimensional spaces where only a few hyperparameters meaningfully affect performance (Bergstra & Bengio, 2012). Note that private hyperparameter search algorithms also do random search, but randomize the number of candidates to test (Papernot & Steinke, 2022).

We adapted this baseline to mixture selection as follows. We sampled m candidate mixture weights $\lambda_1, \ldots, \lambda_m$ independently from the probability simplex Δ_k, ran mixture-weighted public pretraining followed by downstream DP fine-tuning for each λ_i, and selected the best-performing candidate on the same held-out set D_λ^{priv} that DP-MixMin used for its mixture-selection step. We controlled compute by choosing m so that the total cost of training and evaluating all m mixtures matched or exceeded the end-to-end compute of DP-MixMin, ensuring any gains over Random Search could not

Algorithm 2 User-Level DP-MixMin

Require: Target privacy guarantees (ϵ, δ), step size η, number of steps n, loss function L (CE or MSE), private mixture-learning split $D_\lambda^{priv} \subseteq \mathcal{X} \times \mathcal{Y}$ partitioned by user as $D_\lambda^{priv} = \bigsqcup_{u=1}^{N} D_u$, proxy models $\{\hat{f}_i : i = 1, \ldots, k\}$, user batch size B, user-level subsample size G_{ULS}, clipping norm C

Initialize $\lambda_i \leftarrow \frac{1}{k}$ for all $i = 1, \ldots, k$
Pre-compute $\{\hat{f}_i(x) : x \in D_\lambda^{priv},\ i = 1, \ldots, k\}$
for $j = 1$ to n **do**
  Sample a batch $\mathcal{U}_j$ of B users from $\{1, \ldots, N\}$
  $\hat{f}_\lambda(x) \leftarrow \sum_{i=1}^{k} \lambda_i \hat{f}_i(x)$
  for $u \in \mathcal{U}_j$ **do**
    $k_u \leftarrow \min(G_{\text{ULS}}, |D_u|)$
    Sample $D_u^{(j)} \subseteq D_u$ with $|D_u^{(j)}| = k_u$
    $\ell_u \leftarrow \frac{1}{k_u} \sum_{(x,y) \in D_u^{(j)}} L(\hat{f}_\lambda(x), y)$
    $g_u \leftarrow \nabla_\lambda \ell_u$
    $\tilde{g}_u \leftarrow \text{Clip}(g_u, C)$ {clip per-user gradients}
  end for
  $g_{\text{sum}} \leftarrow \sum_{u \in \mathcal{U}_j} \tilde{g}_u$
  $z \sim \mathcal{N}(0, \sigma^2 I)$ {noise calibrated to (ϵ, δ)}
  $\bar{g} \leftarrow \frac{1}{B}(g_{\text{sum}} + z)$
  $\lambda_i \leftarrow \lambda_i\, e^{-\eta \bar{g}_i}$ for all $i = 1, \ldots, k$
  $\lambda \leftarrow \lambda \,/\, \sum_{i=1}^{k} \lambda_i$
end for
Return λ

Table 7: **Per-class final AUC on NIH ChestX-ray14 after transfer from CheXpert.** Each entry is the mean ± standard deviation of AUC over 3–5 seeds at $\delta = 10^{-5}$. Bold indicates the best method per (class, ϵ) cell. DP-MixMin achieves the best macro AUC across all three privacy budgets, with absolute gains of $+0.029$ to $+0.037$ over uniform pretraining.

	DP-MixMin			Uniform			Random Search			No Pretraining		
Class	ϵ=1	ϵ=2	ϵ=5	ϵ=1	ϵ=2	ϵ=5	ϵ=1	ϵ=2	ϵ=5	ϵ=1	ϵ=2	ϵ=5
Atelectasis	**0.740±0.012**	**0.747±0.005**	**0.747±0.007**	0.707±0.013	0.724±0.009	0.736±0.008	0.724±0.009	0.733±0.011	0.739±0.010	0.679±0.006	0.696±0.008	0.715±0.006
Consolidation	0.742±0.014	0.745±0.017	0.748±0.014	**0.745±0.015**	**0.749±0.014**	**0.752±0.013**	0.734±0.008	0.740±0.011	0.742±0.010	0.710±0.006	0.716±0.010	0.725±0.010
Infiltration	**0.671±0.004**	**0.674±0.005**	**0.678±0.002**	0.668±0.004	0.670±0.003	0.672±0.003	0.664±0.004	0.670±0.003	0.675±0.003	0.663±0.004	0.672±0.003	0.680±0.002
Pneumothorax	**0.780±0.007**	**0.788±0.002**	**0.798±0.004**	0.697±0.008	0.728±0.008	0.753±0.009	0.674±0.005	0.701±0.009	0.726±0.007	0.650±0.009	0.684±0.005	0.723±0.005
Edema	0.829±0.017	0.836±0.014	0.845±0.013	0.829±0.006	0.835±0.007	0.838±0.008	**0.844±0.008**	**0.850±0.008**	**0.857±0.009**	0.804±0.012	0.816±0.010	0.829±0.010
Emphysema	**0.683±0.028**	**0.715±0.022**	**0.725±0.016**	0.632±0.015	0.652±0.010	0.688±0.016	0.608±0.008	0.637±0.006	0.664±0.009	0.566±0.037	0.592±0.027	0.631±0.013
Fibrosis	**0.627±0.039**	**0.660±0.016**	**0.668±0.021**	0.609±0.022	0.641±0.029	0.658±0.024	0.591±0.019	0.618±0.013	0.632±0.012	0.578±0.038	0.612±0.023	0.653±0.013
Effusion	**0.848±0.007**	**0.850±0.002**	**0.852±0.006**	0.819±0.006	0.830±0.009	0.833±0.007	0.823±0.009	0.835±0.003	0.839±0.003	0.759±0.004	0.784±0.004	0.807±0.002
Pneumonia	0.626±0.035	0.626±0.025	0.650±0.022	0.649±0.023	**0.660±0.032**	**0.661±0.030**	0.648±0.025	0.652±0.039	**0.663±0.045**	0.613±0.021	0.634±0.019	0.642±0.019
Pleural Thick.	**0.655±0.008**	**0.668±0.009**	**0.676±0.027**	0.639±0.027	0.652±0.022	0.670±0.017	0.604±0.009	0.629±0.009	0.643±0.009	0.582±0.014	0.593±0.010	0.610±0.013
Cardiomegaly	**0.872±0.008**	**0.878±0.010**	**0.883±0.010**	0.710±0.009	0.733±0.006	0.748±0.010	0.762±0.004	0.792±0.002	0.820±0.004	0.669±0.004	0.679±0.010	0.691±0.010
Nodule	**0.574±0.009**	0.566±0.017	0.584±0.015	0.550±0.009	0.564±0.007	0.582±0.004	0.546±0.022	0.564±0.018	0.583±0.017	0.563±0.017	**0.578±0.013**	**0.584±0.015**
Mass	**0.674±0.009**	**0.688±0.009**	**0.700±0.012**	0.631±0.010	0.648±0.009	0.666±0.005	0.619±0.009	0.634±0.006	0.652±0.007	0.550±0.023	0.571±0.018	0.593±0.007
Hernia	**0.583±0.054**	**0.624±0.054**	**0.655±0.044**	0.499±0.043	0.531±0.042	0.551±0.048	0.498±0.036	0.523±0.027	0.538±0.045	0.517±0.022	0.551±0.050	0.561±0.047
Macro AUC	**0.707**	**0.719**	**0.729**	0.670	0.687	0.701	0.667	0.684	0.698	0.636	0.656	0.675
Δ vs. Uniform	+0.037	+0.032	+0.029	—	—	—	−0.003	−0.003	−0.003	−0.034	−0.031	−0.026

be attributed to a larger compute budget. The selection step itself was also implemented non-privately: our implementation of Random Search read D_λ^{priv} directly when picking the winning candidate, paying no privacy cost for the selection step. Random Search was therefore a strong baseline along two axes: equal or greater compute, and no utility degradation from private selection.

A.11 Tables and Figures

B Dataset Access and Asset Licenses

All restricted or credentialed datasets were accessed under their applicable data-use terms. For PhysioNet credentialed-access datasets, the authors completed the required credentialing process and signed the applicable Data Use Agreement before accessing the data. We do not redistribute restricted datasets.

Table 8: **Per-class final AUC on NIH ChestX-ray14 after transfer from MIMIC-CXR.** Each entry is the mean ± standard deviation of AUC over 3–5 seeds at $\delta = 10^{-5}$. Bold indicates the best method per (class, ϵ) cell. DP-MixMin achieves the best macro AUC across all three privacy budgets, with absolute gains of $+0.024$ to $+0.033$ over uniform pretraining.

	DP-MixMin			Uniform			Random Search			No Pretraining		
Class	ϵ=1	ϵ=2	ϵ=5	ϵ=1	ϵ=2	ϵ=5	ϵ=1	ϵ=2	ϵ=5	ϵ=1	ϵ=2	ϵ=5
Atelectasis	**0.751±0.010**	0.749±0.010	0.737±0.000	0.728±0.009	0.734±0.008	0.740±0.008	0.742±0.006	**0.749±0.006**	**0.755±0.005**	0.679±0.006	0.696±0.008	0.715±0.006
Consolidation	**0.756±0.012**	**0.758±0.011**	**0.759±0.015**	0.748±0.010	0.751±0.011	0.753±0.011	0.743±0.012	0.746±0.012	0.749±0.012	0.710±0.006	0.716±0.010	0.725±0.010
Infiltration	**0.674±0.008**	**0.679±0.004**	**0.682±0.005**	0.667±0.004	0.673±0.005	0.677±0.006	0.673±0.006	0.676±0.006	0.680±0.003	0.663±0.004	0.672±0.003	0.680±0.002
Pneumothorax	**0.754±0.013**	**0.770±0.004**	**0.782±0.005**	0.719±0.012	0.741±0.015	0.754±0.015	0.679±0.015	0.707±0.016	0.732±0.016	0.650±0.009	0.684±0.005	0.723±0.005
Edema	**0.845±0.003**	**0.849±0.004**	**0.853±0.005**	0.839±0.007	0.841±0.008	0.845±0.007	0.826±0.011	0.834±0.008	0.841±0.011	0.804±0.012	0.816±0.010	0.829±0.010
Emphysema	**0.710±0.023**	**0.728±0.012**	**0.742±0.012**	0.650±0.013	0.674±0.010	0.691±0.011	0.633±0.033	0.645±0.022	0.693±0.034	0.566±0.037	0.592±0.027	0.631±0.013
Fibrosis	**0.641±0.019**	**0.653±0.019**	**0.680±0.028**	0.614±0.024	0.638±0.026	0.637±0.018	0.587±0.006	0.620±0.004	0.644±0.010	0.578±0.038	0.612±0.023	0.653±0.013
Effusion	**0.849±0.003**	**0.849±0.005**	**0.851±0.005**	0.824±0.007	0.830±0.005	0.833±0.006	0.813±0.003	0.822±0.004	0.829±0.006	0.759±0.004	0.784±0.004	0.807±0.002
Pneumonia	**0.681±0.023**	**0.677±0.027**	0.674±0.020	0.666±0.029	0.664±0.041	**0.682±0.031**	0.645±0.029	0.655±0.029	0.667±0.029	0.613±0.021	0.634±0.019	0.642±0.019
Pleural Thick.	**0.683±0.013**	**0.688±0.017**	**0.704±0.022**	0.664±0.014	0.676±0.011	0.684±0.010	0.621±0.013	0.641±0.005	0.663±0.012	0.582±0.014	0.593±0.010	0.610±0.013
Cardiomegaly	**0.804±0.025**	**0.816±0.016**	**0.835±0.009**	0.711±0.007	0.765±0.026	0.782±0.027	0.749±0.012	0.770±0.019	0.801±0.017	0.669±0.004	0.679±0.010	0.691±0.010
Nodule	**0.589±0.012**	**0.591±0.016**	**0.598±0.008**	0.552±0.013	0.569±0.012	0.585±0.009	0.539±0.008	0.548±0.009	0.559±0.010	0.563±0.017	0.578±0.013	0.584±0.015
Mass	**0.690±0.008**	**0.707±0.005**	**0.719±0.009**	0.635±0.006	0.654±0.004	0.667±0.007	0.597±0.014	0.628±0.017	0.653±0.017	0.550±0.023	0.571±0.018	0.593±0.007
Hernia	**0.546±0.048**	**0.620±0.003**	**0.604±0.058**	0.495±0.024	0.536±0.053	0.551±0.033	0.527±0.056	0.525±0.040	0.580±0.053	0.517±0.022	0.551±0.050	0.561±0.047
Macro AUC	**0.712**	**0.724**	**0.730**	0.679	0.696	0.706	0.670	0.683	0.703	0.636	0.656	0.675
Δ vs. Uniform	+0.033	+0.028	+0.024	—	—	—	−0.010	−0.013	−0.003	−0.044	−0.041	−0.031

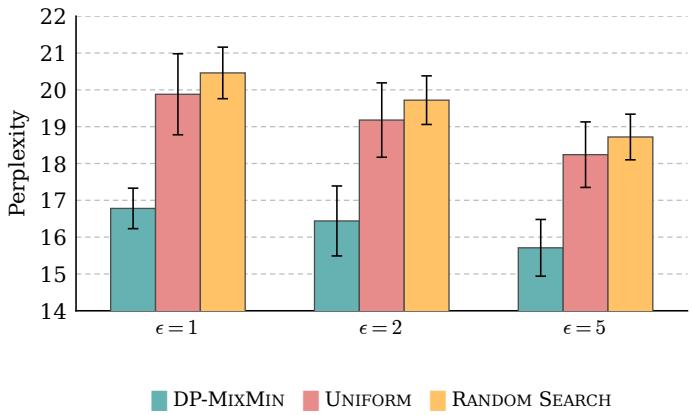


Figure 5: **DP-MixMin beats both uniform-mixture and random-search Common Pile pretraining from scratch on Enron email completion across all three privacy budgets, with the largest gain at the tightest budget.** Test-set perplexity after DP fine-tuning on Enron (δ=10^{-5}) at $\epsilon \in \{1, 2, 5\}$. All three methods pretrain GPT-2 medium from random initialization on Common Pile and are then DP-finetuned on Enron under identical settings. DP-MixMin reduces test perplexity by roughly $16\%/14\%/14\%$ over uniform and $18\%/17\%/16\%$ over random-mixture pretraining at $\epsilon \in \{1, 2, 5\}$.

NIH Chest X-rays. Accessed under the CCO: public domain license which gives free-use. The dataset is available at `https://www.kaggle.com/datasets/nih-chest-xrays/data`

CheXpert. Accessed under the Stanford CheXpert Dataset Research Use Agreement, which permits non-commercial research use. The dataset is available at `https://stanfordmlgroup.github.io/competitions/chexpert`.

MIMIC-CXR. Accessed under the PhysioNet Credentialed Health Data License 1.5.0. Access requires completion of a credentialing process and signing a Data Use Agreement through PhysioNet (Johnson et al., 2019a).

Enron Email Dataset. The Enron Corpus was released into the public domain following federal litigation and is freely available for research use (Klimt & Yang, 2004b).

The Common Pile. Released under the Open Data Commons Attribution License (ODC-By-1.0). Individual constituent sources carry their own licenses; we refer readers to the dataset documentation for a full provenance breakdown (Kandpal et al., 2025).

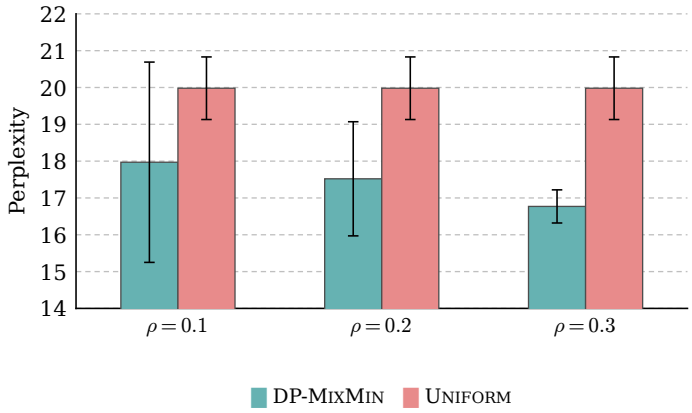


Figure 6: **DP-MixMin consistently outperforms uniform Common Pile pretraining on Enron at ϵ=1 across all proxy-split fractions, with larger fractions yielding both lower perplexity and tighter cross-seed variance.** Test-set perplexity after DP fine-tuning on Enron (δ=10^{-5}) at ϵ=1, varying the proxy fraction $\rho \in \{0.1, 0.2, 0.3\}$ used by MixMin to estimate domain weights. The uniform baseline is independent of ρ and is shown for reference at all three values. DP-MixMin reduces test perplexity by $10\%/12\%/16\%$ over uniform at $\rho \in \{0.1, 0.2, 0.3\}$, with ρ=0.3 giving the most stable mixture (cross-seed σ drops from $2.72 \rightarrow 0.45$).

Table 9: **NIH ChestX-ray14 class distribution and correspondence to CheXpert/MIMIC-CXR labels.** The **Match** column indicates whether each NIH class has a direct (*Yes*), *Partial*, or no (*None*) equivalent in CheXpert/MIMIC-CXR.

NIH Idx	Class Name	# Pos.	% Pos.	Match	CheXpert/MIMIC-CXR Label
0	Atelectasis	11,559	10.31%	Yes	Atelectasis
1	Consolidation	4,667	4.16%	Yes	Consolidation
2	Infiltration	19,894	17.75%	Partial	Lung Opacity
3	Pneumothorax	5,302	4.73%	Yes	Pneumothorax
4	Edema	2,303	2.05%	Yes	Edema
5	Emphysema	2,516	2.24%	None	—
6	Fibrosis	1,686	1.50%	None	—
7	Effusion	13,317	11.88%	Yes	Pleural Effusion
8	Pneumonia	1,431	1.28%	Yes	Pneumonia
9	Pleural Thickening	3,385	3.02%	Partial	Pleural Other
10	Cardiomegaly	2,776	2.48%	Yes	Cardiomegaly
11	Nodule	6,331	5.65%	Partial	Lung Lesion
12	Mass	5,782	5.16%	Partial	Lung Lesion
13	Hernia	227	0.20%	None	—

Table 10: **Average AUC gains of DP-MixMin over Uniform pretraining, grouped by class type.** Group A: well-represented classes with a direct CheXpert match (Atelectasis, Cardiomegaly, Effusion, Pneumonia, Pneumothorax, Consolidation, Edema). Group B: rare or unmatched classes (Infiltration, Mass, Nodule, Emphysema, Fibrosis, Pleural Thickening, Hernia); see Table 9. Each cell is the within-group mean of per-class absolute (abs) and relative (rel) AUC gain of DP-MixMin over Uniform pretraining at $\delta = 10^{-5}$.

	$\epsilon = 1$		$\epsilon = 2$		$\epsilon = 5$	
	abs	rel	abs	rel	abs	rel
CheXpert → NIH ChestX-ray14						
Group A ($n = 7$)	+0.040	+5.6%	+0.030	+4.0%	+0.029	+3.8%
Group B ($n = 7$)	+0.034	+6.0%	+0.034	+5.7%	+0.028	+4.7%
MIMIC-CXR → NIH ChestX-ray14						
Group A ($n = 7$)	+0.029	+4.0%	+0.020	+2.7%	+0.015	+1.8%
Group B ($n = 7$)	+0.037	+6.2%	+0.035	+5.8%	+0.034	+5.3%

Table 11: **Weight reconstruction of DP-MixMin across 14 NIH ChestX-ray14 tasks.** Mean ± std over n=70 paired comparisons (14 tasks × 5 seeds). We compare the private weights $\boldsymbol{\lambda}_{\mathrm{DP}}$ to the non-private MixMin solution $\boldsymbol{\lambda}^\star$ using cosine similarity $\cos_\star$ and L_1 distance. Uniform references are CheXpert: $\cos_{\mathbf{u}}$=0.75, $L_{1,\mathbf{u}}$=0.68; MIMIC-CXR: $\cos_{\mathbf{u}}$=0.80, $L_{1,\mathbf{u}}$=0.60. The DP-MixMin selection budget $\rho = 0.1$. Bold values improve over the corresponding uniform reference.

		$\epsilon = 0.2$		$\epsilon = 0.5$		$\epsilon = 1$		$\epsilon = 2$		$\epsilon = 5$	
src	n	$\cos_\star$	L_1	$\cos_\star$	L_1	$\cos_\star$	L_1	$\cos_\star$	L_1	$\cos_\star$	L_1
CheXpert → NIH ChestX-ray14											
CheXpert	70	0.70±.08	0.81±.14	**0.93**±.03	**0.37**±.09	**0.97**±.01	**0.24**±.06	**0.98**±.01	**0.19**±.05	**0.99**±.01	**0.16**±.04
MIMIC-CXR → NIH ChestX-ray14											
MIMIC-CXR	70	0.58±.17	1.04±.30	**0.82**±.09	**0.57**±.20	**0.93**±.04	**0.33**±.10	**0.97**±.01	**0.22**±.05	**0.98**±.01	**0.16**±.03

Table 12: **Weight reconstruction of DP-MixMin for NIH Hernia** (0.20% prevalence in D^{priv}, 227 total positives). Mean ± std over 5 paired comparisons (5 seeds). We compare the private weights $\boldsymbol{\lambda}_{\mathrm{DP}}$ to the non-private MixMin solution $\boldsymbol{\lambda}^\star$ using cosine similarity $\cos_\star$ and L_1 distance. Uniform references: CheXpert $\cos_{\mathbf{u}}$=0.64, $L_{1,\mathbf{u}}$=0.96; MIMIC-CXR $\cos_{\mathbf{u}}$=0.73, $L_{1,\mathbf{u}}$=0.82. Bold values improve over the corresponding uniform reference.

		ρ=0.05 n≈4,485, #pos≈9		ρ=0.1 n≈8,970, #pos≈18		ρ=0.2 n≈17,939, #pos≈36	
Pretrain	ε	$\cos_\star$	L_1	$\cos_\star$	L_1	$\cos_\star$	L_1
CheXpert	0.2	**0.79**±0.01	**0.64**±0.02	**0.94**±0.01	**0.32**±0.01	**0.98**±0.00	**0.17**±0.01
	0.5	**0.96**±0.01	**0.27**±0.03	**0.99**±0.00	**0.15**±0.01	**0.99**±0.00	**0.10**±0.01
	1.0	**0.99**±0.00	**0.16**±0.03	**0.99**±0.00	**0.10**±0.01	**1.00**±0.00	**0.08**±0.01
	2.0	**0.99**±0.00	**0.11**±0.02	**1.00**±0.00	**0.08**±0.01	**1.00**±0.00	**0.08**±0.01
	5.0	**1.00**±0.00	**0.09**±0.02	**1.00**±0.00	**0.08**±0.02	**1.00**±0.00	**0.08**±0.01
MIMIC-CXR	0.2	0.62±0.00	1.08±0.00	**0.79**±0.00	**0.67**±0.00	**0.92**±0.00	**0.35**±0.00
	0.5	**0.83**±0.00	**0.57**±0.00	**0.94**±0.00	**0.30**±0.00	**0.98**±0.00	**0.17**±0.00
	1.0	**0.94**±0.00	**0.31**±0.00	**0.98**±0.00	**0.17**±0.00	**0.99**±0.00	**0.10**±0.00
	2.0	**0.98**±0.00	**0.18**±0.00	**0.99**±0.00	**0.11**±0.00	**1.00**±0.00	**0.08**±0.00
	5.0	**0.99**±0.00	**0.10**±0.00	**1.00**±0.00	**0.07**±0.00	**1.00**±0.00	**0.06**±0.00

Table 13: **Weight reconstruction of DP-MixMin for NIH Fibrosis** (1.50% prevalence in D^{priv}, 1,686 total positives). Mean ± std over 5 paired comparisons (5 seeds). We compare the private weights $\boldsymbol{\lambda}_{\mathrm{DP}}$ to the non-private MixMin solution $\boldsymbol{\lambda}^\star$ using cosine similarity $\cos_\star$ and L_1 distance. Uniform references: CheXpert $\cos_{\mathbf{u}}$=0.73, $L_{1,\mathbf{u}}$=0.77; MIMIC-CXR $\cos_{\mathbf{u}}$=0.83, $L_{1,\mathbf{u}}$=0.60. Bold values improve over the corresponding uniform reference.

		ρ=0.05 n≈4,485, #pos≈57		ρ=0.1 n≈8,970, #pos≈114		ρ=0.2 n≈17,939, #pos≈229	
Pretrain	ε	$\cos_\star$	L_1	$\cos_\star$	L_1	$\cos_\star$	L_1
CheXpert	0.2	**0.78**±0.02	**0.70**±0.03	**0.94**±0.01	**0.38**±0.02	**0.98**±0.00	**0.23**±0.01
	0.5	**0.95**±0.01	**0.33**±0.04	**0.98**±0.00	**0.20**±0.02	**0.99**±0.00	**0.16**±0.01
	1.0	**0.98**±0.01	**0.22**±0.03	**0.99**±0.00	**0.16**±0.01	**0.99**±0.00	**0.16**±0.01
	2.0	**0.99**±0.00	**0.18**±0.03	**0.99**±0.00	**0.16**±0.01	**0.99**±0.00	**0.15**±0.01
	5.0	**0.99**±0.00	**0.17**±0.02	**0.99**±0.00	**0.15**±0.01	**0.99**±0.00	**0.15**±0.01
MIMIC-CXR	0.2	0.59±0.00	1.10±0.01	0.78±0.00	0.69±0.01	**0.92**±0.00	**0.37**±0.01
	0.5	0.82±0.00	0.59±0.01	**0.94**±0.00	**0.32**±0.01	**0.98**±0.00	**0.19**±0.00
	1.0	**0.93**±0.00	**0.34**±0.01	**0.98**±0.00	**0.21**±0.00	**0.99**±0.00	**0.15**±0.01
	2.0	**0.97**±0.00	**0.21**±0.01	**0.99**±0.00	**0.16**±0.00	**0.99**±0.00	**0.13**±0.01
	5.0	**0.99**±0.00	**0.15**±0.01	**0.99**±0.00	**0.14**±0.01	**1.00**±0.00	**0.10**±0.01

Table 14: **Enron weight recovery under user-level DP with K=16 Common Pile proxies.** Mean $\pm$ std over 3 paired comparisons (3 seeds). We compare the private mixture weights $\boldsymbol{\lambda}_{\mathrm{DP}}$ to the non-private MixMin solution $\boldsymbol{\lambda}^{\star}$ using cosine similarity $\cos_{\star}$ and L_1 distance. Columns vary the DP-MixMin selection budget ρ, the fraction of the private mixture-learning split used to estimate the mixture weights; the second header row reports the corresponding number of Enron users. The uniform reference is $\cos_{\mathbf{u}}$=0.40 and $L_{1,\mathbf{u}}$=1.47. Bold values improve over the uniform reference on that metric.

	ρ=0.05 (6 users)		ρ=0.1 (12 users)		ρ=0.15 (18 users)		ρ=0.2 (24 users)		ρ=0.3 (35 users)		ρ=0.5 (59 users)	
ε	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1
0.2	—	—	0.08±0.11	1.89±0.13	0.01±0.00	1.98±0.01	0.34±0.44	**1.41**±0.72	0.19±0.16	1.65±0.16	**0.44**±0.24	**1.25**±0.28
0.5	0.12±0.15	1.85±0.19	0.15±0.20	1.75±0.30	0.07±0.04	1.75±0.13	**0.70**±0.30	**0.80**±0.46	**0.86**±0.13	**0.70**±0.13	**0.95**±0.01	**0.54**±0.08
1	0.12±0.15	1.83±0.19	**0.61**±0.26	**1.08**±0.38	**0.61**±0.26	**1.01**±0.27	**0.97**±0.03	**0.36**±0.15	**0.97**±0.03	**0.46**±0.07	**0.99**±0.00	**0.40**±0.03
2	0.16±0.13	1.68±0.15	**0.92**±0.05	**0.65**±0.22	**0.93**±0.05	**0.45**±0.14	**0.99**±0.00	**0.28**±0.06	**0.99**±0.01	**0.37**±0.03	**0.99**±0.00	**0.36**±0.01
5	**0.93**±0.04	**0.48**±0.14	**0.98**±0.00	**0.41**±0.10	**0.99**±0.00	**0.17**±0.05	**1.00**±0.00	**0.28**±0.02	**0.99**±0.00	**0.33**±0.01	**0.99**±0.00	**0.34**±0.01
10	**0.98**±0.01	**0.28**±0.05	**0.99**±0.00	**0.17**±0.04	**1.00**±0.00	**0.09**±0.02	**1.00**±0.00	**0.29**±0.01	**0.99**±0.00	**0.32**±0.01	**0.99**±0.00	**0.33**±0.01

Table 15: **OpenWebMath weight recovery under example-level DP with K=16 Common Pile proxies.** Mean $\pm$ std over 3 paired comparisons (3 seeds). We compare the private mixture weights $\boldsymbol{\lambda}_{\mathrm{DP}}$ to the non-private MixMin solution $\boldsymbol{\lambda}^{\star}$ using cosine similarity $\cos_{\star}$ and L_1 distance. Columns vary the DP-MixMin selection budget ρ, the fraction of the private mixture-learning split used to estimate the mixture weights; the second header row reports the corresponding number of OpenWebMath tokens. The uniform reference is $\cos_{\mathbf{u}}$=0.40 and $L_{1,\mathbf{u}}$=1.28. Bold values improve over the uniform reference on that metric.

	ρ=0.01 (275K tokens)		ρ=0.02 (552K tokens)		ρ=0.05 (1.4M tokens)		ρ=0.1 (2.8M tokens)	
ε	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1	$\cos_{\star}$	L_1
0.2	0.17±0.08	1.66±0.14	0.34±0.17	1.38±0.23	**0.78**±0.10	**0.88**±0.14	**0.92**±0.02	**0.72**±0.04
0.4	0.32±0.14	1.39±0.20	**0.68**±0.13	**0.99**±0.16	**0.92**±0.02	**0.72**±0.04	**0.94**±0.00	**0.67**±0.01
0.6	**0.49**±0.15	**1.19**±0.20	**0.82**±0.07	**0.84**±0.10	**0.94**±0.01	**0.69**±0.02	**0.95**±0.00	**0.67**±0.01
0.8	**0.64**±0.12	**1.02**±0.15	**0.88**±0.04	**0.77**±0.06	**0.94**±0.01	**0.68**±0.01	**0.95**±0.00	**0.67**±0.01
1	**0.73**±0.09	**0.92**±0.12	**0.91**±0.02	**0.74**±0.04	**0.94**±0.00	**0.67**±0.00	**1.00**±0.00	**0.05**±0.03
2	**0.90**±0.01	**0.74**±0.04	**0.94**±0.00	**0.68**±0.02	**0.95**±0.00	**0.67**±0.01	**1.00**±0.00	**0.04**±0.01
5	**0.94**±0.01	**0.67**±0.01	**0.94**±0.01	**0.67**±0.02	**0.95**±0.00	**0.67**±0.01	**1.00**±0.00	**0.03**±0.01

Table 16: **Empirical privacy audit of DP-MixMin's mixture-learning step, CheXpert → NIH ChestX-ray14, at $\varepsilon = 1$.** We audit the released mixture weights $\boldsymbol{\lambda} = M_{\mathrm{mix}}(D_{\mathrm{mix}})$, reporting the Clopper–Pearson empirical lower bound $\varepsilon_{\mathrm{LB}}$ at $\varepsilon = 1$, $\delta = 10^{-5}$, $\rho = 0.1$, and $k = 14$ proxy sources, over **2000** runs on add/remove-neighboring datasets (attack fit on a SELECT half, bound computed on a disjoint EVAL half; 95% confidence). The *natural* canary is a real held-out example; the *max-clip* canary injects a worst-case gradient at the clipping norm. *Private* audits at $\varepsilon = 1$; *No-noise control* removes the DP noise. Every private $\varepsilon_{\mathrm{LB}} \ll 1$ while controls recover large leakage, confirming audit fidelity.

		Natural canary		Max-clipped-gradient canary	
Class	**% pos**	Private (ε=1)	No-noise ctrl.	Private (ε=1)	No-noise ctrl.
Atelectasis	10.31	0.000	0.261	0.080	3.71
Consolidation	4.16	0.000	0.501	0.052	5.19
Infiltration	17.74	0.000	0.088	0.000	2.27
Pneumothorax	4.73	0.000	0.068	0.080	5.18
Edema	2.05	0.000	0.404	0.000	4.93
Emphysema	2.24	0.000	0.619	0.054	4.93
Fibrosis	1.50	0.000	1.188	0.091	5.18
Effusion	11.88	0.000	0.000	0.028	3.73
Pneumonia	1.28	0.000	0.344	0.098	5.60
Pleural Thickening	3.02	0.000	1.226	**0.131**	5.60
Cardiomegaly	2.48	0.000	0.148	0.000	4.93
Nodule	5.65	0.000	0.539	0.000	5.60
Mass	5.16	0.000	0.000	0.000	5.60
Hernia	0.20	0.000	2.360	0.086	5.60
max over tasks	—	0.000	2.360	**0.131**	5.60

Table 17: **Empirical privacy audit of DP-MixMin's mixture-learning step, MIMIC-CXR → NIH ChestX-ray14, at $\varepsilon = 1$.** Identical protocol to Table 16: Clopper–Pearson $\varepsilon_{\mathrm{LB}}$ at $\varepsilon = 1$, $\delta = 10^{-5}$, $\rho = 0.1$, and $k = 14$ MIMIC-CXR proxy sources, over 2000 neighboring-dataset runs (SELECT/EVAL split, 95% confidence). As for CheXpert, every private $\varepsilon_{\mathrm{LB}} \ll 1$ while controls recover substantial leakage.

		Natural canary		**Max-clipped-gradient canary**	
Class	**% pos**	Private (ε=1)	No-noise ctrl.	Private (ε=1)	No-noise ctrl.
Atelectasis	10.31	0.000	0.000	0.000	4.02
Consolidation	4.16	0.000	0.501	0.000	4.72
Infiltration	17.74	0.000	0.000	0.001	3.21
Pneumothorax	4.73	0.000	0.022	0.000	4.92
Edema	2.05	0.000	0.672	0.000	5.59
Emphysema	2.24	0.000	0.000	0.000	5.40
Fibrosis	1.50	0.000	2.627	0.056	4.44
Effusion	11.88	0.000	0.000	0.000	2.87
Pneumonia	1.28	0.000	0.480	0.000	5.17
Pleural Thickening	3.02	0.000	1.147	0.000	5.07
Cardiomegaly	2.48	0.000	1.895	**0.101**	5.51
Nodule	5.65	0.000	0.582	0.000	4.73
Mass	5.16	0.000	0.551	0.000	4.76
Hernia	0.20	0.000	4.934	0.061	4.92
max over tasks	—	0.000	4.934	**0.101**	5.51

Table 18: **DP-MixMin vs. its non-private version, CheXpert → NIH ChestX-ray14 at $\varepsilon = 1$.** Per-class final AUC (mean ± std over 3–5 seeds, $\delta = 10^{-5}$). Classes are ordered by NIH index (Table 9). DP-MixMin matches the non-private upper bound to within 0.001 macro AUC.

Class	**Non-private MixMin**	**DP-MixMin** (ε=1)
Atelectasis	0.745 ± 0.007	0.747 ± 0.006
Consolidation	0.753 ± 0.014	0.756 ± 0.012
Infiltration	0.669 ± 0.005	0.669 ± 0.006
Pneumothorax	0.796 ± 0.008	0.789 ± 0.018
Edema	0.819 ± 0.012	0.809 ± 0.019
Emphysema	0.715 ± 0.026	0.746 ± 0.025
Fibrosis	0.645 ± 0.024	0.655 ± 0.019
Effusion	0.850 ± 0.006	0.851 ± 0.005
Pneumonia	0.595 ± 0.015	0.592 ± 0.019
Pleural Thickening	0.688 ± 0.017	0.671 ± 0.022
Cardiomegaly	0.878 ± 0.007	0.868 ± 0.013
Nodule	0.585 ± 0.012	0.570 ± 0.012
Mass	0.691 ± 0.004	0.670 ± 0.011
Hernia	0.545 ± 0.030	0.567 ± 0.029
Macro AUC	**0.712**	**0.711**

Table 19: **DP-MixMin vs. its non-private version, MIMIC-CXR → NIH ChestX-ray14 at $\varepsilon = 1$.** Per-class final AUC (mean ± std over seeds, $\delta = 10^{-5}$). *Non-private MixMin* finds the mixture with no privacy and fine-tunes on the *full* private set (both splits), so it is a strict upper bound; DP-MixMin reserves the mixing split (parallel composition) and privatizes both steps. DP-MixMin matches the non-private upper bound to within 0.006 macro AUC.

Class	**Non-private MixMin**	**DP-MixMin** (ε=1)
Atelectasis	0.758 ± 0.006	0.749 ± 0.010
Consolidation	0.764 ± 0.015	0.761 ± 0.013
Infiltration	0.678 ± 0.008	0.676 ± 0.003
Pneumothorax	0.779 ± 0.004	0.754 ± 0.014
Edema	0.838 ± 0.007	0.835 ± 0.010
Emphysema	0.745 ± 0.018	0.720 ± 0.028
Fibrosis	0.654 ± 0.024	0.672 ± 0.029
Effusion	0.855 ± 0.005	0.851 ± 0.004
Pneumonia	0.656 ± 0.022	0.658 ± 0.016
Pleural Thickening	0.687 ± 0.023	0.692 ± 0.014
Cardiomegaly	0.850 ± 0.011	0.832 ± 0.030
Nodule	0.613 ± 0.007	0.580 ± 0.007
Mass	0.700 ± 0.017	0.675 ± 0.013
Hernia	0.543 ± 0.064	0.584 ± 0.069
Macro AUC	**0.723**	**0.717**