A FRAMEWORK FOR COPYRIGHT AND RECONSTRUCTION PROTECTION VIA MAX INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a formal framework for understanding and limiting the impact of sensitive data on data analysis and model training procedures. Specifically, we introduce the notion of piecewise information limited (PIL) algorithms. We focus in particular on the power of this framework as a copyright protection standard, for which it has several practical advantages over existing copyright protection frameworks. First, our framework does not require that only one (or some prespecified number of) datapoints depend on a copyrighted work, as needed in previous frameworks. Second, our framework allows us to quantify at a granular level to what degree the prompt increases the risk of a generative model regurgitating a copyrighted work. Third, our definition is strictly more general than the commonly leveraged technique of differential privacy and allows for a class of noiseless mechanisms which are not differentially private. Finally, our framework is auditable via memorization testing, thus allowing regulators to verify whether training procedures adhere to the protection standard. In this vein, we propose a new algorithm for memorization testing and empirically demonstrate its capability. In addition to copyright, our framework also provides formal protection guarantees for the two closely related domains of data reconstruction and data memorization. As such, we believe our framework may serve more generally as a way to understand responsible data usage.

1 INTRODUCTION

It has long been known that extraction attacks can reconstruct training data from machine learning models (Fredrikson et al., 2015). Even more severely, recent work has shown that large generative models can regurgitate (i.e. output upon prompting) their training data on seemingly benign prompts (Yarkoni & Livni, 2025). This phenomenon raises both data privacy and copyright protection concerns. Consequently, several related lines of work, training data reconstruction, copyright protection, and training data memorization, have seen a surge of recent interest.

In this work, we propose a framework for protecting sensitive data used in data analysis mechanisms, such as training generative machine learning models. We introduce piecewise information limited (PIL) algorithms, which bound the amount of information a model retains from each datapoint in its training data. We further characterize their behavior on a general data sampling model that separates sensitive data records (i.e. the fundamental units we care about protecting) from dataset entries (which may result from modifying or mixing several data records). Our framework has several upshots over prior work: it admits upper bounds on the degree of memorization a model can exhibit, leads to formal protection guarantees against reconstruction attacks even when the dataset contains correlated data, and can quantify how much a prompt increases the risk of a model outputting a copyrighted work. We discuss mechanisms that satisfy our definition and construct an efficient auditing scheme to test memorization of PIL algorithms.

Copyright protection. The rapidly improving ability of generative models to produce content has resulted in concern over whether the output of these models infringes on the copyright of content used in their training data (Cooper et al., 2023; Lee et al., 2024). These copyright concerns motivate the need for a rigorous framework that limits the generation of infringing content. Unfortunately however, existing work on these formal frameworks has so far not provided protection guarantees under realistic data assumptions.

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.

The first formal framework for copyright protection was near access freeness (NAF), which required that the output distribution of a model be close to some "safe" model (Vyas et al., 2023). Subsequently, Cohen (2025) introduced the notion of *Blameless Copy Protection*, which roughly requires that on "blameless" users, the model does not output a copyrighted work with high probability. While the design of copyright safe generative models is a multifaceted problem, a key focus of both these and other works has been bounding the probability a model outputs an infringing work Golatkar et al. (2024); Chen et al. (2024b). Cohen (2025) observed that a more refined version of this goal is to avoid producing infringing content only on "reasonable" prompts. That is, we likely do not view the model as copyright unsafe if the prompt contains substantial information about the copyrighted work (i.e. a prompt which asks the model to simply regurgitate a book whose text was provided in the prompt itself does constitute a copyright unsafe model).

Unfortunately, existing work has provided formal protection guarantees only under the assumption that the dataset was "golden", a strong condition which requires an explicit a priori bound on the number of dataset entries correlated with any given copyrighted work. Such exact independence of the remaining data is impractical in general. Further, the definition of blameless copy protection involves restricting the model behavior only under "blameless" users, whose prompt distribution satisfies a certain constraint. As noted in Cohen (2025), it is unclear how a user can ensure/check that they are blameless. Consequently, we also lack ways to audit models which claim to satisfy blameless copy protection (because the standard only restricts model behavior on blameless users). Additionally, the framework separates users into a binary category of blameless or not blameless. In reality, the degree to which a user is responsible for a model outputting copyrighted work exists on a spectrum which depends on to what degree the user's prompts depend on the copyrighted work. Another important limitation is that these frameworks have so far only been instantiated with differential privacy, which limits the insight into new algorithm design.[1]

We address these limitations by providing a framework which provides protection even on non-golden datasets and gives users a concrete way to limit their involvement in a model regurgitating copyrighted work. Further, because our protection standard does not involve inaccessible quantities, such as the underlying prompt distribution or the corpus of all copyrighted works at train time, it is also auditable via memorization testing.

Data Reconstruction. A related line of work has sought to bound or measure to what extent adversaries can reconstruct training data. We believe the tools we develop for understanding and protecting against copyright infringement are also relevant to this line of work. Specifically, we highlight that most work on dataset reconstruction has focused on the ability of an adversary to reconstruct a dataset entry, often under the assumption that the entries are sampled independently (Balle et al., 2022; Cummings et al., 2024; Cohen et al., 2024; Swanberg et al., 2026). However, protecting dataset *entries* is only a proxy for the actual goal of protecting sensitive data *records*. For example, consider the case where a medical dataset contains a particular person's record but not their child's record. Due to correlations between a parent and child's medical information, an adversary may learn about the contents of the child's medical data even though the child was not in the dataset (the researcher may not even know if such a child exists). If both of the child's parents are included in the dataset, the data analysis may feasibly reveal more about the child than about either parent individually. Alternatively, a dataset entry may contain information from multiple sensitive records; consider for example a report discussing multiple individuals medical data. We provide more detailed discussion of these ideas and related work in Appendix B.2. In particular, in Appendix B.4 we provide a detailed comparison to the Narcissus resiliency framework proposed by Cohen et al. (2024) and highlight several advantages of PIL.

Memorization. Recent years have seen a surge of interest in measuring to what degree models memorize their training data, although the literature has not yet converged on a standard definition of memorization. Among such proposals are perplexity (Carlini et al., 2019), counterfactual memorization (Zhang et al., 2023), mutual information Brown et al. (2021), adversarial reconstruction (Schwarzschild et al., 2024), and Kolmogorov complexity (Morris et al., 2025; Cohen et al., 2024). A common pattern in most of these approaches is to measure how difficult it is to extract the training data from the model (as defined via some particular scheme), and normalize this quantity by some

[1] Vyas et al. (2023) provided a non-DP mechanism for NAF which was also studied by Golatkar et al. (2024), but the mechanism was later found to be capable of regurgitating copyrighted work Cohen (2025)

baseline which scales with the natural extraction difficulty. To our knowledge, our framework is the first to give explicit bounds on the degree of memorization under such extraction based definitions.

Differential Privacy. A common thread throughout much of the work on reconstruction and copyright protection has been the observation that differential privacy (DP) provides strong protection, even with large privacy parameters. This has been explored at length in both empirical and theoretical work (Bousquet et al., 2020; Hayes et al., 2023; Chen et al., 2024b; Cummings et al., 2024; Guo et al., 2022; Kaissis et al., 2023; Swanberg et al., 2026). However, differential privacy can be overly restrictive for the purpose of protecting against reconstruction. For example, consider two adjacent datasets $S, S' \in (\{0,1\}^d)^n$. Differential privacy precludes a mechanism from releasing any bit of any datapoint because it is possible for that bit to fully determine where the mechanism was run on S or S'. This goes against the intuition that any data which is non-trivially difficult to reconstruct would not be compromised by a single bit; this issue is resolved by our framework.

Additional Related Work: Additional discussion of related work can be found in Appendix B.

1.1 Our Contribution: A New Unifying Protection Standard

In this work, we propose an algorithm level protection standard which has implications for copyright protection, reconstruction attacks and memorization. Our method is also built around understanding the protection of underlying sensitive data records (which are the objects the framework protects but may not be entries in the training dataset), rather than realized dataset entries. We model this via the following.

A Two Stage Model. We consider the following setup for sensitive data. Let $U = \{R_1, R_2, ..., R_N\}$ be a universe of data records generated from some distribution, $\mathcal{D}$, some of which may be private or copyrighted. This could represent, for example, the set of all images that exist at a given point in time. A data analyst receives a dataset, $S = X_1, ..., X_n$, of n records chosen via the "sampling" process $\mathsf{Samp}(U)$. For clarity, we will refer to the random variables in U as data records and those in S as dataset entries. Note that Samp does not necessarily have to select records from U verbatim. It can be any possibly randomized function of U. For example, Samp could create a dataset entry by paraphrasing some string of text in U, recombine multiple records from U, or add an image augmented by some filter. The analyst then runs mechanism $\mathcal{M}$ on the received dataset and generates $Y \sim \mathcal{M}(S)$.

A notable aspect of this model is how it separates data correlations that arise from the underlying uncertainty in the data distribution versus the sampling procedure. Consider two different cases. In the first, $\mathcal{D}$ may be a product distribution such that elements of U are drawn i.i.d., but Samp may select only elements that are similar to each other. For example, the data collector could collect user data based on friend networks. Alternatively, consider when $\mathcal{D}$ is the (non-product) distribution such that any R_j provides high information about other elements of U. For example, U may contain many different images of a single character from popular media, leading to the scenario where S contains multiple depictions of this character even when Samp selects elements from U uniformly at random. In either case the dataset entries are correlated. This highlights two facts. First, assuming i.i.d. dataset entries is generally not a safe assumption for the purpose of reconstruction. Even if the sampling process is i.i.d., the dataset entries may not be i.i.d. due to an adversary's uncertainty in the underlying data distribution. Second, preventing adversarial reconstruction depends not only on the data analysis procedure, but also on the design of the sampling process. Specifically, practical techniques for mitigating memorization/copyright infringement/reconstruction often incorporate data de-duplication procedures (Chen et al., 2024a; Kandpal et al., 2022). Ensuring protection even against pathological sampling scenarios where most datapoints are highly correlated or repeated is likely to lead to poor utility guarantees. As such, our main protection results will focus on the scenario where Samp creates a dataset with "reasonably" independent/deduplicated entries.

Piecewise Information Limited Algorithms. The second part of our framework is a protection standard for data analysis mechanisms (such as the training process of a generative model). Specifically, we introduce the notion of Piecewise Information Limited algorithms, and show how this single definition simultaneously has implications for adversarial reconstruction, copyright protection,

and memorization. Our standard is built on the notion of approximate max information introduced in Dwork et al. (2015).

Definition 1. *Dwork et al. (2015) The β-approximate max information (AMI) between random variables X and Y with supports $\mathcal{X}$ and $\mathcal{Y}$, is* $I_\infty^\beta(X;Y) = \max_{\substack{\mathcal{O}\subseteq\mathcal{X}\times\mathcal{Y} \\ \mathbb{P}[(X,Y)\in\mathcal{O}]>\beta}} \log\left(\frac{\mathbb{P}[(X,Y)\in\mathcal{O}]-\beta}{\mathbb{P}[X\times Y\in\mathcal{O}]}\right)$. *A randomized function $\mathcal{M}:\mathcal{X}\mapsto\mathcal{Y}$ has β-AMI capacity defined as $I_\infty^\beta(\mathcal{M}) := \max_X\{I_\infty^\beta(X;\mathcal{M}(X))\}$, where the max is over all distributions over $\mathcal{X}$ for X.*

Above, $X \times Y$ denotes the product (rather than joint) distribution over X and Y. While originally proposed as a way to control generalization in adaptive data analysis, approximate max information has not to our knowledge been applied in the domains of copyright and data reconstruction. Crucially it provides strong control over the change in the posterior of X given an observation of Y. This motivates our definition of piecewise information limited algorithms.

Definition 2. *(Piecewise Information Limited) Let $\mathcal{M}:\mathcal{X}^n\mapsto\mathcal{Y}$ be a (possibly randomized) mechanism. For every $\sigma\subseteq[n]$ and $z\in\mathcal{X}^{n-|\sigma|}$, let $\mathcal{M}_{z,\sigma}:\mathcal{X}^{|\sigma|}\mapsto\mathcal{Y}$ denote the mechanism induced by fixing the inputs of $\mathcal{M}$ with indices not in σ to the corresponding elements of z. $\mathcal{M}$ is (α,β,k)-piecewise information limited (PIL) if $I_\infty^\beta(\mathcal{M}_{z,\sigma})\le\alpha$ for every $|\sigma|\subseteq[n], |\sigma|\le k$ and $z\in\mathcal{X}^{n-|\sigma|}$. When $k=1$, we say $\mathcal{M}$ satisfies (α,β)-PIL.*

PIL requires that for any set of indices σ of size at most k and any fixing of the inputs with indices not in σ, the AMI capacity of the channel formed by $\mathcal{M}$ with respect to all remaining inputs is bounded. We show our definition has a variety of desirable properties.

Protection Guarantees of PIL. In Section 2, we show how PIL can be interpreted as a standard for preventing copyright infringement or adversarial data reconstruction. In contrast to prior work, we show how PIL algorithms can provide meaningful guarantees even when many datapoints contain information about a copyrighted/private data record. This allows us to reason about sensitive data risk even without golden datasets or independence of the dataset entries, assumptions which are unlikely to hold in practice. At the same time, our standard is also designed with realistic data distributions in mind. Specifically, we leverage the intuition that few dataset entries contain high information with any one sensitive data record. Additionally, dataset deduplication is an important piece of ensuring sensitive data is handled responsibly in practice. Our guarantees naturally becomes more effective as data deduplication methods become more effective, but does not require perfect deduplication or an a priori guarantee on the success of the deduplication process. We show how to use PIL algorithms to reason about the reconstruction/regurgitation risk of any sensitive data record, regardless of whether it is "in" the dataset. The ambiguity of defining membership has been shown to be a real concern in practice; models have been shown to memorize data using only disjoint pieces or fuzzy duplicates (Liu et al., 2025; Shilov et al., 2026). However, this factor has been essentially ignored by existing reconstruction definitions. Further, our standard also leads to explicit lower bounds on the length of the shortest program that can extract a data record from the output of a PIL process. Consequently, PIL provides guarantees for memorization/extractability (under several recent definitions).

For copyright protection, our definition avoids placing model users/prompts into binary categories of blameless or not blameless. Instead, when a generative model is trained with a PIL training procedure, our framework naturally places model users on a continuum of responsibility based on how much AMI their prompts contain about a copyrighted work. This also leads to a concrete and intuitive way users can minimize their culpability for regurgitation, which is to use short prompts.

Implementing PIL Algorithms and Comparison to Differential Privacy. We show how PIL algorithms can be implemented in Section 3. An important aspect of the class of PIL algorithms is that it is broader than the class of DP algorithms. That is, (ϵ,δ)-DP algorithms are (ϵ,δ)-PIL, but additionally, PIL algorithms can be built from mechanisms with conditionally bounded description length (i.e. mechanisms which can be written in terms of few bit function of k inputs after fixing the remaining $n-k$ inputs). This is due to the fact that such mechanisms have bounded AMI capacity. Conditionally bounded description length mechanisms can be noiseless, and do not satisfy DP in general. Finally, the fact that AMI does not rely on adjacency relations leads to a more flexible approach when analyzing the impact of side information.

Auditing PIL Algorithms. A key practical benefit of PIL algorithms is that they are auditable, which we show in Section 4. That is, we provide a concrete procedure that can be performed on the trained model to accept or refute the claim that it was generated from a PIL training process provided synthetic data (i.e. canaries) has been inserted into the training data. More generally, we show that memorization testing doubles as auditing for PIL algorithms. Complementing this result, we develop a new practical memorization test and empirically compare its performance to the adversarial compression test of Schwarzschild et al. (2024).

2 Protection Guarantees of PIL Algorithms

In this section, we argue how Definition 2 serves as a standard for guarding against copyright violation, adversarial reconstruction, and memorization. We focus here primarily on copyright protection, and discuss reconstruction and memorization subsequently. Recall we consider a data model as in Section 1.1, where $U = \{R_1, R_2, ...R_N\}$ is a set of data records generated from some distribution, $\mathcal{D}$ and $X_1, ..., X_n$ are dataset entries generated from some sampling process over U, $X_1, ..., X_n \sim \mathsf{Samp}(U)$. As with existing reconstruction/copyright/memorization standards, it is important to quantify guarantees with respect to some baseline. Let $\mathsf{Rec}(x)$ denote the set of points considered a reconstruction of (or the in the case of copyright, "substantially similar" to) the point x. We define the reconstruction entropy as follows.

Definition 3. *For a random variable X, the β-smooth reconstruction entropy is defined as* $H^{\beta}_{\mathsf{Rec}}(X) = \sup_{\mathcal{E}\subseteq\mathcal{X}:\mathbb{P}[X\in\mathcal{E}]\leq\beta} \inf_{x\in\mathcal{X}} \left\{ -\log\left(\mathbb{P}\left[X \in \{x' : x \in \mathsf{Rec}(x')\} \setminus \mathcal{E}\right]\right)\right\}$.

We also define $H^{\beta}_{\mathsf{Rec}}(X|Y=y)$ in the obvious way (i.e. replacing the law of X with the conditional law). Note that $2^{-H^0_{\mathsf{Rec}}(X)} = \sup_{z\in\mathcal{X}} \mathbb{P}_X[z \in \mathsf{Rec}(X)]$. That is, H^0_{Rec} exactly characterizes the probability of reconstructing X in one guess. The $\beta = 0$ case is the baseline event used in Balle et al. (2022) and Cummings et al. (2024). Additionally, when $\mathsf{Rec}(x) = \{x\}$ and $\beta > 0$, this is a quantity known as the smooth min entropy (Renner & Wolf, 2005). Note $H^{\beta}_{\mathsf{Rec}} \geq H^0_{\mathsf{Rec}}$, so it is a more permissive baseline than assuming H^0_{Rec} is large. Intuitively, the β parameter allows for there to be some small probability of a (possibly exact) reconstruction even before receiving additional information from a data analysis procedure. In the following, we will present guarantees for data records with non-trivial reconstruction entropy.

2.1 Protecting Against Copyright Infringement

For copyright protection, we consider generative models. A model designer chooses a training process $\mathcal{T}$ [2], and a generative model is then sampled $h \sim \mathcal{T}(S)$. Similar to Cohen (2025), we take $\mathsf{Rec}(x)$ to characterize the set of points "substantially similar" to x (i.e. points which would be deemed infringing). Additionally, a model user chooses/samples a prompt P, and receives $h(P)$. Consider some copyrighted work $R \in U$. The high level interpretation of our main result is that when 1) R is non-trivial to imitate (i.e. high reconstruction entropy) 2) the prompt P contains little information about R and 3) Samp succeeds in generating a "reasonably" deduplicated dataset, then w.h.p. h does not output something substantially similar to R. We can even provide a stronger guarantee, which is that no short program transforms the output of h into R. "Reasonably deduplicated" in this instance means that, for some small k, all but the k most informative data records (indexed by σ in the following theorem) contain minimal aggregate information about R.

Theorem 1. *Let $\beta' \in [0,1]$, $K \in [n]$, and $R \in U$. Assume that $\mathcal{T}$ is $(\alpha(k), \beta, k)$-PIL for each $k \in [K]$. For any realization of $S = \{x_1, ..., x_n\}$ and $P = p$, define the random variable $H^* := \max_{\sigma\subseteq[n],|\sigma|\leq K} \{H^{\beta'}_{\mathsf{Rec}}(R|X_{\notin\sigma} = x_{\notin\sigma}, P = p) - \alpha(|\sigma|)\} - 1$, where $X_{\notin\sigma} = \{X_i : i \notin \sigma\}$. If $\mathcal{T}$ is run on a uniformly random permutation of S then,*

1. ***Regurgitation Prevention:*** $\mathbb{P}_{R,S,\mathcal{T},h}[h(P) \in \mathsf{Rec}(R) \mid H^* \geq \log(1/\beta')] \leq 2\sqrt{\beta' + 2\beta}$.

[2] As is common in the literature, we implicitly assume the model generation process is defined in a manner independent of the underlying data. Alternatively, one could additionally condition on the realization of $\mathcal{T}$ when defining H^* in Theorem 1 to account for a process which does not satisfy this independence.

2. ***Extraction Difficulty:*** *Let $\mathcal{L} = \mathcal{L}(\mathcal{D})$ be a programming language. With probability at least $1 - 2\sqrt{\beta' + 4\beta}$, no program in $\mathcal{L}$ of description length at most $H^* - 2\log(1/\beta')$ given as input h and P outputs an element in $\mathsf{Rec}(R)$.*

The proof, as well as a more precise definition of "programming language" is in Appendix C.1. The proof proceeds by bounding the conditional reconstruction entropy of R using the accumulation of max information. Note in this regard that max information does not share the same chain rule properties of mutual information. For the rest of this section, we discuss the theorem statement itself. For the sake of intuition, we note that similar (unconditional) probability bounds can be obtained under the scenario where the quantity $H_{\mathsf{Rec}}(R) - I_\infty^{\beta'}(R; X_{\neq\sigma}, P) - \alpha(k)$ (instead of H^*) is sufficiently large (we provide an explicit statement in Appendix C.5). That is, one prevents reconstruction when the random variable storing the $n-k$ least informative (about R) dataset entries do not contain too much information about R. However, such a statement ignores how the guarantees of PIL are adaptive to the sampled dataset instance. Theorem 1 shows that, if $\mathcal{M}$ satisfies the PIL condition for a variety of group sizes, the protection guarantees scale adaptively based on how many datapoints are reasonably informative to R. One does not need to know the group size in advance, nor does one need to ensure the group parameter k is large enough to cover every datapoint that may depend on R (which leads to poor scaling in α). See Appendix C.4 for an illustrative example.

Our framing is built on the intuition that for most realistic settings, the bulk of the information about a copyrighted work will be contained in a few highly correlated dataset entries; information which is "held back" due to the PIL constraint. This also better aligns with realistic deduplication efforts, which are more likely to remove dataset entries with similar information, rather than be able to scrub the dataset of any entry containing any information about the copyrighted work.

Similar to prior work on copyright protection (Vyas et al., 2023; Cohen, 2025), an important aspect of Theorem 1 and Definition 2 is that the designer of $\mathcal{T}$ does not need access to the substantial similarity function, Rec, which is likely impractical to determine algorithmically. Further, while Theorem 1 uses a random permutation of the dataset, whose application is tailored to applications where the position of an entry in the dataset does not matter, similar guarantees can be obtained for unpermuted data at the expense of higher entropy requirements on R (see Appendix C.5). We also provide discussion in Appendix C.6 on how max information compares to other potential information choices.

How model users minimize regurgitation risk. We see that another key benefit of the PIL definition is that it places model users on a continuum of responsibility rather than the binary category of blameless or non-blameless used by Cohen (2025). Specifically, users which add less information about a copyrighted work into the generation process via their prompts make it less likely to cause the model to generate copyrighted works. Further, instead of viewing P as side information provided to the model, the extraction length guarantee of Theorem 1 means that it *suffices for users to make short prompts to minimize the chance of generating infringing content* (see Appendix C.3 for formal details). The upshot is that PIL training leads to concrete and intuitive advice for how users themselves can minimize the chance of generating infringing work, which previous frameworks have been unable to provide (Vyas et al., 2023; Cohen, 2025).

Infringement protection beyond regurgitation. In contrast to existing copyright protection standards, our definition explicitly leads to a lower bound on the program length required to transform $h(P)$ into a reconstruction of R. This program length notion was proposed by Scheffler et al. (2022) as a formal way of testing whether a work is similar enough to be infringing.

PIL Training vs. PIL Generation Theorem 1 considers the case where the model creation process is piecewise information limited. An alternative condition would be to only ensure that, for any prompt p', the aggregate process which generates $h \sim \mathcal{T}(S)$ and $Y \sim h(p')$ is PIL. This weaker condition is satisfied if either the training procedure or inference procedure is PIL, and is sufficient to ensure regurgitation prevention and difficulty of extraction. That said, PIL training may have additional copyright related benefits, such as ensuring that the training process was "substantially transformative". We provide further details in Appendix C.2.

2.2 Protection Against Adversarial Reconstruction

PIL algorithms also prevent data reconstruction. An analog of Theorem 1 is obtainable for the setting of adversarial reconstruction; see Theorem 5 in Appendix C.1 for a full statement. In brief, rather than a generative model reconstructing the record R given additional information from a prompt, we can consider an adversary reconstructing the record using some side information (which is possibly unknown to the mechanism designer). In contrast to previous protection guarantees for reconstruction, such as those leveragin differential privacy, our result does not assume independently sampled dataset entries. Indeed, the degree to which entries are independent or dependent depends on the side information held by the adversary. As such, we focus on providing guarantees which scale smoothly with the level of side information held by the adversary, rather than assuming strict properties on the data distribution. We provide more discussion in Appendix B.3, and in Appendix B.4 discuss in detail how our standard compares to the recently proposed Narcissus resiliency framework of Cohen et al. (2024), while highlighting several advantages of PIL algorithms.

2.3 Limiting Memorization

From the second guarantee of Theorem 1, we can see how PIL training bounds memorization/extractability. A fairly general recent definition of extractability comes from Cohen et al. (2024), which states that a value r is $(\mathsf{Rec}, \mathcal{L})$-extractable from h if there exists a program, $\mathcal{A}$, in language $\mathcal{L}$ which generates r when given h as input. The quality of the extraction is given by $1 - \frac{\mathsf{length}(\mathcal{A})}{K_{\mathcal{L}}(r)}$, where $\mathsf{length}(\mathcal{A})$ is the description length of program $\mathcal{A}$ in $\mathcal{L}$ and $K_{\mathcal{L}}(r)$ is the shortest program which generates r (without h as input). Note that $\mathcal{K}_{\mathcal{L}}(r)$ is independent of the data analysis mechanism $\mathcal{T}$. Simultaneously, Theorem 1 (taking the prompt to be empty), implies that any program which extracts R from h has lower bounded length with high probability, thus bounding the extraction quality obtainable by an adversary. Relatedly, Schwarzschild et al. (2024) proposed measuring memorization via the adversarial compression ratio, $\mathsf{ACR}(h, r) = \frac{|r|}{|p^*|}$, where $p^* = \arg\min_{p:h(p)=r}\{|p|\}$. That is, the length of the string r over the length of the shortest prompt that generates the string. Since the prompt can be viewed as the description of a program which generates r given h as input, memorization is similarly bounded by Theorem 1.

3 Implementing PIL Algorithms

An important property of PIL algorithms is that they inherent a rich class of DP mechanisms, as well as a new class of conditionally bounded description length mechanisms. The following lemmas show that approximate DP directly leads to PIL guarantees with $k = 1$, and a slightly stronger notion known as zero concentrated DP leads to more robust protections that can handle large group sizes. Proofs for the results in this section are deferred to Appendix D.

Lemma 1. *Let $\epsilon > 0$, $\delta \in [0, 1]$, and $\mathcal{M}$ be an (ϵ, δ)-DP algorithm. Then $\mathcal{M}$ is (ϵ, δ)-PIL.*

Lemma 2. *Let $\rho \geq 0$. If $\mathcal{M}$ is a ρ-zCDP algorithm, then for any $k \in [n]$ and $\delta > 0$ $\mathcal{M}$ satisfies $(2\rho k^2 + \sqrt{\rho k^2 \cdot \log(\sqrt{\pi \rho k^2}/\delta)}, \delta, k)$-PIL.*

Separately, mechanisms with conditionally bounded description length also have bounded max information capacity. Notably, such mechanisms may be noiseless.

Lemma 3. *Let $k \in [n]$. Let $\mathcal{M} : \mathcal{X}^n \mapsto \mathcal{Y}$ be a (possibly randomized) mechanism such that for any fixing of $n - k$ of its inputs and any fixing of the random seed of $\mathcal{M}$, the induced range (with respect to the remaining inputs) of $\mathcal{M}$ is a discrete set with cardinality at most M. Then for any $\beta \in (0, 1]$, $\mathcal{M}$ satisfies $(\log(M/\beta), \beta, k)$-PIL.*

This generalizes the bounded description length mechanism of Dwork et al. (2015), which is a special case when $k = n$. In particular, by lowering k, we can obtain new mechanisms such as *releasing exact histograms* with efficient α budget (see Appendix D). In the context of model training, a clear practical application of bounded range mechanisms is for hyperparameter tuning. Differential privacy alone requires adding noise to the hyperparameters, but by simply constraining hyperparameters to a grid and doing DP training, one can achieve a PIL model creation process without ever adding noise to the hyperparameters.

Composition Similar to DP, PIL mechanisms enjoy certain composition and subsampling guarantees. Roughly, composing T PIL mechanisms each with parameters $\{(\alpha_t, \beta_t)\}_{t\in[T]}$ leads to an aggregate mechanism which is $(\sum_{t=1}^{T} \alpha_t, \sum_{t=1}^{T} \beta_t)$-PIL. We provide details in Appendix D.

Parameter range. In contrast to a standard like differential privacy, the necessary range of (α, β) for PIL algorithms is problem dependent. High entropy data and high accuracy reconstruction relations will tolerate much higher settings of α than low dimensional data. Evidence suggests obtaining meaningful reconstruction prevention with $\alpha \gg 1$ is plausible for many scenarios involving high dimensional data. Experiments using differential privacy have already shown that (ϵ, δ)-DP with ϵ in the hundreds have empirically had success preventing reconstruction attacks on realistic data (Balle et al., 2022; Hayes et al., 2023).

4 Auditing PIL Generative Models

In this section, we give a procedure for auditing PIL generative models by way of a new data extraction procedure. Our extractor may be of independent interest for the purpose of memorization testing, which several works have measured by normalizing the extraction program length by a baseline difficulty of generating the datapoint (Schwarzschild et al., 2024; Cohen et al., 2024). Motivated by this, we provide experimental results on the general efficacy of our data extraction procedure on real world data.

4.1 Auditing Procedure

The goal of auditing is for a third-party to test the claim that a model is generated from a PIL algorithm (Definition 2). The auditing procedure we propose leverages the extraction difficulty guarantee provided by PIL algorithms (Theorem 1). In particular, if a model is generated via a PIL training process, we know that with high probability there is a lower bound on the length of the program (e.g., specifying the prompt) that outputs a particular data record. The auditing procedure finds an upper bound on this minimal program length required to reconstruct the data record of interest. If the extraction program is short enough, we show the auditor can use this as evidence to reject the PIL claim with high confidence.

Specifying edit operations. Our auditing procedure uses a pre-specified series of edit operations. Let $\mathcal{F}$ be a finite set of functions. We define the edit "distance" (it is not necessarily symmetric) with respect to $\mathcal{F}$ of points x_1, x_2 as $d_{\mathcal{F}}(x_1, x_2) := \min_{l\in\mathcal{Z}} \{\exists\, f_1, \ldots, f_l \in \mathcal{F} : (f_l \circ \cdots \circ f_1)(x_2) \in \mathsf{Rec}(x_1)\}$. Note that the instantiation of Rec used for the auditing procedure does not necessarily have to be the reconstruction function of interest, which is important for applications where we do not have algorithmic access to Rec. In particular, we can always use $\mathsf{Rec}(x) = \{x\}$. We also define a prompt alphabet $\mathcal{P}$ (i.e. the set of all ASCII characters, or the set of all input tokens).

Performing the audit. To perform the audit, we assume an auditor is given a training process $\mathcal{T}$ which is claimed to be (α, β)-PIL. The auditor constructs a dataset $x_1, ..., x_{n-1}$ and a canary distribution $\mathcal{C}$. The auditor then samples $X' \sim \mathcal{C}$ and runs $\mathcal{T}$ on a random shuffling of $\{x_1, ..., x_{n-1}, X'\}$, S, to generate model h. The auditor then attempts to find a prompt in $\mathcal{P}^*$ (where $\mathcal{P}$ is the vocabulary of the programming language the auditor is using, e.g. possible tokens if directly searching over prompts to a generative model) and sequence of edit operations in $\mathcal{F}$ which are (in aggregate) short. Formally, the auditor determines an upper bound for the quantity,

$$m^*(y, h) = \min_{p\in\mathcal{P}^*} \{|p| \cdot \log(|\mathcal{P}|) + d_{\mathcal{F}}(y, h(p)) \cdot \log(|\mathcal{F}|)|\}. \tag{1}$$

Solving this minimization exactly will generally not be practical; we must settle for an upper bound on the minimum program length m^*. We provide an approximate algorithm in Appendix E.3 that finds an upper bound. We have the following guarantee, which says that models which are actually PIL require Algorithm 1 to use a prompt and sequence of edit operations which are long (in aggregate) to reconstruct the canary. Thus obtaining a small upper bound on m^* suggests a large α parameter of the PIL process. Note that an explicit probability that $\mathcal{T}$ is not PIL requires additional

modeling assumptions on the prior probability that $\mathcal{T}$ is PIL; the following result is more akin to a confidence interval.

Theorem 2. *Let $\beta, \beta', \alpha > 0$. Assume $\mathcal{T}$ is (α, β)-PIL, $h \sim \mathcal{T}(S)$, and m^* as defined in equation 1. Then with probability at least $1 - 2\sqrt{\beta' + 4\beta}$ over the randomness in $\mathcal{T}$, h, and X' it holds that, $m^*(X', h) \geq H^{\beta'}_{\mathsf{Rec}}(X') - \alpha - 2\log(1/\beta') - 1$.*

The proof is in Appendix E.1. We have here assumed for simplicity that the model h is a deterministic function, but one can alternatively leverage randomness in h by approximating m^* for several random seeds. Theorem 2 then implies a high probability lower bound on each m^* after a union bound. Note also that $\mathcal{P}^*$ and $\mathcal{F}$ may depend on the canary distribution $\mathcal{C}$, but not the canary itself.

4.2 Efficacy of the Data Extraction Procedure

A result similar to Theorem 2 could be obtained via other extraction/memorization testing algorithms, such as the MiniPrompt algorithm of Schwarzschild et al. (2024), which searches only over prompts rather than prompts and edit operations (see Algorithm 3 in Appendix E.5). However, our proposed auditing procedure achieves tighter results when the extraction procedure is able to reconstruct data with shorter programs. We show in this section how the added flexibility afforded to our procedure leads to shorter program lengths on real world data, which may be of independent interest for memorization testing. We provide further explanation on how our process serves as a memorization test in Appendix E.2.

Experimental setup. We run experiments with our proposed memorization test ExtData (Algorithm 1) and Schwarzschild et al. (2024)'s MiniPrompt (Algorithm 3). The experiments use open-source Pythia models (Biderman et al., 2023) trained on the Pile (Gao et al., 2020) with parameter counts of 410M, 1.4B, 6.9B, and 12B. We evaluate the auditing procedures using Theorem 2. The algorithm then returns an estimate of the minimum length of the extraction program, giving a bound on memorization. We include more details about the experimental setup, including the edit operations we consider in Appendix E.4.

Datasets. We evaluate Algorithm ExtData on two datasets. First, Famous Quotes, a set of 100 famous quotes curated by Schwarzschild et al. (2024). These famous quotes likely recur in many examples of the training data, and therefore are likely to be memorized (Schwarzschild et al., 2024). Second, a set of 50 Enron emails included in the Pile (Gao et al., 2020). The Enron emails dataset consists of much longer token sequences than the Famous Quotes dataset (52.12 vs 14.31 tokens per dataset entry on average, respectively). The Enron emails experiments were only run on the models with parameter counts of 410M and 1.4B.

Results. On the Famous Quotes and Enron emails datasets, we find that ExtData returned shorter program lengths across all model sizes than the previous baseline, indicating that ExtData provides tighter results for memorization testing. See Figure 2 in Appendix E.4 for details. Of particular interest, on the Famous Quotes dataset, ExtData achieved shorter program lengths while using significantly shorter prompts; highlighting that ExtData leverages the fact that the bulk of the information stored in the models is easily extractable. See Table 1 in Appendix E.4 for details.

5 Future Work

We see several important avenues for future work. First, while our framework shows how noiseless bounded description length mechanisms can be used to protect against reconstruction, it is still unclear how to leverage such mechanisms in the training of large models beyond basic applications such as hyperparameter tuning. Additionally, while our framework is sensitive to how both the sampling process and data analysis mechanism affect reconstruction risk, our main aim is to provide guidance for the design of the data analysis mechanism. Studying how new sampling and data deduplication procedures interface with our guarantees is an interesting avenue for future work.

AI USE STATEMENT

This paper was developed using the assistance of AI tools to check proofs, debug experiments, and catch typos. Additionally, the proof of Lemma 12 was found by AI tools and then verified and rewritten by the authors. The authors are responsible for all content in the paper. AI tools were not used for developing conceptual frameworks, writing exposition, generating datasets, designing experiments, or interpreting results.

REFERENCES

Mário Alvim, Kostas Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. The science of quantitative information flow. 01 2020. doi: 10.1007/978-3-319-96131-6.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1138–1156, 2022. URL https://api.semanticscholar.org/CorpusID:245906243.

Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. Coupled-Worlds Privacy: Exploiting Adversarial Uncertainty in Statistical Data Privacy . In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 439–448, Los Alamitos, CA, USA, October 2013. IEEE Computer Society. doi: 10.1109/FOCS.2013.54. URL https://doi.ieeecomputersociety.org/10.1109/FOCS.2013.54.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/pdf/2304.01373.

Sayan Biswas, Mark Dras, Pedro Faustini, Natasha Fernandes, Annabelle McIver, Catuscia Palamidessi, and Parastoo Sadeghi. Bayes' capacity as a measure for reconstruction attacks in federated learning, 2024. URL https://arxiv.org/abs/2406.13569.

Olivier Bousquet, Roi Livni, and Shay Moran. Synthetic data generators – sequential and private. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pp. 123–132, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Martin Hirt and Adam Smith (eds.), *Theory of Cryptography*, pp. 635–658, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-53641-4.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8425–8434, 2024a. doi: 10.1109/CVPR52733.2024.00805.

Wei-Ning Chen, Peter Kairouz, Sewoong Oh, and Zheng Xu. Randomization techniques to mitigate the risk of copyright infringement, 2024b. URL https://arxiv.org/abs/2408.13278.

Aloni Cohen. Blameless users in a clean room: Defining copyright protection for generative models, 2025. URL https://arxiv.org/abs/2506.19881.

Edith Cohen, Haim Kaplan, Yishay Mansour, Shay Moran, Kobbi Nissim, Uri Stemmer, and Eliad Tsfadia. Data reconstruction: When you see it and when you don't, 2024. URL https://arxiv.org/abs/2405.15753.

A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A. Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, Jack M. Balkin, Nicholas Carlini, Christopher De Sa, Jonathan Frankle, Deep Ganguli, Bryant Gipson, Andres Guadamuz, Swee Leng Harris, Abigail Z. Jacobs, Elizabeth Joh, Gautam Kamath, Mark Lemley, Cass Matthews, Christine McLeavey, Corynne McSherry, Milad Nasr, Paul Ohm, Adam Roberts, Tom Rubin, Pamela Samuelson, Ludwig Schubert, Kristen Vaccaro, Luis Villa, Felix Wu, and Elana Zeide. Report of the 1st workshop on generative ai and law, 2023. URL https://arxiv.org/abs/2311.06477.

Rachel Cummings, Shlomi Hod, Jayshree Sarathy, and Marika Swanberg. Attaxonomy: Unpacking differential privacy guarantees against practical adversaries, 2024. URL https://arxiv.org/abs/2405.01716.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 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.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506.02629, 2015. URL http://arxiv.org/abs/1506.02629.

Vitaly Feldman, Guy Kornowski, and Xin Lyu. Trade-offs in data memorization via strong data processing inequalities. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1935–1973. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/feldman25a.html.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813677. URL https://doi.org/10.1145/2810103.2813677.

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.

Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. Cpr: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12374–12384, June 2024.

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning, 2022. URL https://arxiv.org/abs/2201.12383.

Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Jamie Hayes, Borja Balle, and Saeed Mahloujifar. Bounding training data reconstruction in DP-SGD. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=7LZ4tZrYlx.

Georgios Kaissis, Jamie Hayes, Alexander Ziller, and Daniel Rueckert. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy, 2023. URL https://arxiv.org/abs/2307.03928.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10697–10707. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kandpal22a.html.

Shiva P. Kasiviswanathan and Adam Smith. On the 'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), Jun. 2014. doi: 10.29012/jpc.v6i1.634. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/634.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'bout ai generation: Copyright and the generative-ai supply chain (the short version). In *Proceedings of the 2024 Symposium on Computer Science and Law*, CSLAW '24, pp. 48–63, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703331. doi: 10.1145/3614407.3643696. URL https://doi.org/10.1145/3614407.3643696.

Ken Liu, Christopher A. Choquette-Choo, Matthew Jagielski, Peter Kairouz, Sanmi Koyejo, Percy Liang, and Nicolas Papernot. Language models may verbatim complete text they were not explicitly trained on. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=bLcXkIasck.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 81–90, 2010. doi: 10.1109/FOCS.2010.14.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize?, 2025. URL https://arxiv.org/abs/2505.24832.

Renato Renner and Stefan Wolf. Simple and tight bounds for information reconciliation and privacy amplification. In Bimal Roy (ed.), *Advances in Cryptology - ASIACRYPT 2005*, pp. 199–216, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32267-2.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-Information, Differential Privacy, and Post-selection Hypothesis Testing . In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–494, Los Alamitos, CA, USA, October 2016. IEEE Computer Society. doi: 10.1109/FOCS.2016.59. URL https://doi.ieeecomputersociety.org/10.1109/FOCS.2016.59.

Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing human ingenuity: A quantitative framework for copyright law's substantial similarity. CSLAW '22, pp. 37–49, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392341. doi: 10.1145/3511265.3550444. URL https://doi.org/10.1145/3511265.3550444.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Chase Lipton, and J Zico Kolter. Rethinking LLM memorization through the lens of adversarial compression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=KFmRMvzAZy.

Igor Shilov, Matthieu Meeus, and Yves-Alexandre Montjoye. The mosaic memory of large language models. *Nature Communications*, 17, 01 2026. doi: 10.1038/s41467-026-68603-0.

Marika Swanberg, Meenatchi Sundaram Muthu Selva Annamalai, Jamie Hayes, Borja Balle, and Adam Smith. A unified framework for adversary-aware differential privacy bounds, 2026. URL https://arxiv.org/abs/2507.08158.

Nikhil Vyas, Sham Kakade, and Boaz Barak. On provable copyright protection for generative models, 2023. URL https://arxiv.org/abs/2302.10870.

Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. CCS '19, pp. 225–240, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367479. doi: 10.1145/3319535.3354261. URL https://doi.org/10.1145/3319535.3354261.

Sol Yarkoni and Roi Livni. Low resource reconstruction attacks through benign prompts. In *The Impact of Memorization on Trustworthy Foundation Models: ICML 2025 Workshop*, 2025. URL https://openreview.net/forum?id=zZ2F3Gfish.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A Additional Preliminaries

Differential privacy. We reference two types of differential privacy in this work.

Definition 4. *(Approximate Differential Privacy, (Dwork et al., 2006)) An algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all datasets S and S' differing in one data point and all events $\mathcal{E}$ in the range of the $\mathcal{A}$, we have, $\mathbb{P}[\mathcal{A}(S) \in \mathcal{E}] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(S') \in \mathcal{E}] + \delta$*

Definition 5. *(Zero Concentrated Differential Privacy, (Bun & Steinke, 2016)) An algorithm $\mathcal{A}$ is (ξ, ρ)-zero concentrated differentially private (zCDP) if for all datasets S and S' differing in one data point and all $\alpha \in (1, \infty)$, it holds that $D_\alpha(\mathcal{A}(S)||\mathcal{A}(S')) \leq \xi + \rho\alpha$, where $D_\alpha(X||Y) = \frac{1}{\alpha-1}\int \rho_X(x)^\alpha \rho_Y(x)^{1-\alpha} dx$ denotes the α-Rényi divergence. We define ρ-zCDP as $(0, \rho)$-zCDP.*

B Additional Discussion of Related Work

B.1 Relaxations of Differential Privacy and its Connection to Information

The notion of max information and its connection to pure DP was originally introduced in Dwork et al. (2015). Subsequently, Rogers et al. (2016) established a connection between differential privacy and approximate max information under i.i.d. dataset entries. It was shown in McGregor et al. (2010) that approximate DP does not bound even the mutual information with the dataset when the elements are not independently distributed.

Several works have also studied relaxations of differential privacy, often centered on how an adversary's posterior belief about the data changes given the output of some data analysis procedure (Bassily et al., 2013; Kasiviswanathan & Smith, 2014). However, their formulations are not primarily motivated by reconstruction and differ substantially from ours.

B.2 Related Work in Data Reconstruction

While there is a long line of work on methods to empirically reduce reconstruction potential (Fredrikson et al., 2015; Yang et al., 2019; Carlini et al., 2023), work on formal information theoretic protection standards are more recent. A common theme in recent attempts has been to compare the probability that an adversary can reconstruct an element of the dataset given access to $\mathcal{M}(S)$, the output of mechanism $\mathcal{M}$ on dataset S drawn from distribution $\mathcal{D}$, to the probability of some "baseline" event. For example, one could require for some $\epsilon > 0$ and any adversary Adv,

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D},\mathcal{M}}\left[\mathsf{Adv}(\mathcal{M}(S)) \in \cup_{x\in S}\mathsf{Rec}(x)\right] \leq e^{\epsilon}\max_{z\in\mathcal{X}}\mathop{\mathbb{P}}_{S\sim\mathcal{D}}\left[z \in \cup_{x\in S}\mathsf{Rec}(x)\right]. \tag{2}$$

Here, $\mathsf{Rec}(x)$ denotes the set of points considered a reconstruction of x, and $\mathcal{X}$ is a discrete data domain. In words, the above states that no adversary can reconstruct a datapoint in S with probability much larger than they could had they not been given the output of $\mathcal{M}$. This is roughly the approach taken by Balle et al. (2022) and Cummings et al. (2024), although the adversary information and distributional assumptions vary between the two. More recently, Cohen et al. (2024) proposed that a better baseline event is the probability that the adversary's particular output reconstructs a datapoint in a fresh dataset sample from the distribution, among other modifications. See Appendix B.4 for a detailed discussion of how Narcissus resiliency relates to our framework.

However, seemingly all prior work focuses on reconstruction of dataset entries specifically, but we argue this is only a proxy for the actual goal of protecting sensitive data. For example, consider the case where a medical dataset contains a person's data but not their child's data. Due to correlations between a parent and child's medical information, an adversary may learn about the contents of the child's medical data even though the child was not in the dataset (the researcher may not even know if such a child exists). If both of the child's parents are included in the dataset, the data analysis may feasibly reveal more about the child than about either parent individually. Similarly, it has been noted in the context of copyright protection that a copyrighted work may influence a mechanism's output through derivative work, regardless of whether the copyrighted work itself was in the dataset Cohen (2025). Such examples show that defining what it means for a sensitive data record to be "in" the dataset is not straightforward and the important impact data correlations can have on reconstruction risk. These concerns are not theoretical either; work has already show that models themselves can generate training data they were never explicitly trained on (Liu et al., 2025). Relatedly, we note that most existing work on reconstruction has shown protection gaurantees only under the assumption that the dataset entries are independent. In this work, we provide a framework for understanding and protecting against reconstruction of sensitive records, regardless of whether or not they are in the dataset.

Finally, we highlight several other works that have studied reconstruction using information theoretic tools. Hannun et al. (2021) and Guo et al. (2022) studied reconstruction risk in terms of leaked Fisher information leakage, although their guarantees do not generalize to arbitrary reconstruction functions. Biswas et al. (2024) used Bayes capacity as measure of reconstruction risk in federated learning, but do not provide guarantees for inexact reconstruction. See also Appendix C.6 for further discussion of these works.

B.3 PIL Algorithms Prevent Data Reconstruction

Theorem 5 shows how a PIL data analysis procedure prevents an adversary, $\mathcal{A}$, holding side information B, from reconstructing some data record $R \in U$. Many of the points stated for copyright protection in Section 2.1 remain true for reconstruction. A few additional points are worth highlighting.

First, the protection guarantees degrade smoothly as one increases the amount of information held by the adversary, regardless of how that information affects the level of correlation between dataset entries. Specifically, every extra bit of information given to the adversary can roughly decrease the reconstruction entropy of the record R by 1-bit, which at most results in a increase in the successful reconstruction probability by roughly a factor of 2. This avoids imposing a strict requirement that the distribution over the dataset entries remain independent under differing side information.

Second, implicit in Theorem 5 is that PIL provides a continuum of guarantees for different levels of reconstruction entropy based on how the reconstruction entropy of the corresponding reconstruction relations changes. Consider for example the family of reconstruction relations, $\{\mathsf{Rec}_\tau\}_{\tau \geq 0}$ such that for any $\tau \leq \tau'$, $\mathsf{Rec}_\tau(x) \subseteq \mathsf{Rec}_{\tau'}(x)$ for all $x \in \mathcal{X}$. That is, $\mathsf{Rec}_\tau(x)$ only contains reconstructions of x which are at least as accurate as those in $\mathsf{Rec}_{\tau'}$. It is then straightforward to see that for any distribution over X and any $\beta' \in [0,1]$ that $H^{\beta'}_{\mathsf{Rec}_\tau}(X) \geq H^{\beta'}_{\mathsf{Rec}_{\tau'}}(X)$. This is to say that even if the PIL parameter α is insufficient to prevent reconstruction at accuracy level τ', it may still provide some protection by preventing more accurate reconstructions.

B.4 Comparison to Narcissus Resiliency

In this section, we elaborate on the comparison to Narcissus resiliency and how it compares to piecewise information limited algorithms. We recall the definition here,

Definition 6 (Cohen et al. (2024))**.** *Let $\mathcal{F}$ be a family of distributions of domain $\mathcal{X}$. An Algorithm $\mathcal{M}$ is $(\epsilon, \delta, \mathcal{F})$-*Rec*-Narcissus-resilient if for all $\mathcal{D} \in \mathcal{F}$ and all attackers $\mathcal{A}$, it holds that*

$$\mathbb{P}_{S\sim\mathcal{D},\mathcal{M},\mathcal{A}}\left[\mathcal{A}(\mathcal{M}(S)) \in \mathsf{Rec}(S)\right] \leq e^{\epsilon} \mathbb{P}_{S,S'\sim\mathcal{D},\mathcal{M},\mathcal{A}}\left[\mathcal{A}(\mathcal{M}(S)) \in \mathsf{Rec}(S')\right] + \delta.$$

where $\mathsf{Rec}(S)$ *denotes* $\cup_{x\in S}\mathsf{Rec}(x)$.

Towards a comparison to PIL, let us first consider for this section that we are only interested in protecting dataset entries from reconstructions, rather than any underlying data records (as discussed in Section 1.1). Second, we note that some form of distributional assumptions are likely necessary for a reconstruction protection standard to be practical. PIL implicitly does this via a definition that quantifies over all possible distributions and provides guarantees only for sufficiently nice distributions, whereas Narcissus resiliency requires a property holds only for a subclass of distributions. To a large extent, these differences are a matter of presentation.

Instead, our main focus of this section is to highlight how Narcissus resiliency and PIL lead to differing or similar updates in the posterior given the output of the mechanism. For simplicity, we focus on the case where $n = 1$ here, which allows for particularly clean comparisons. Note that when $n = 1$, a mechanism $\mathcal{M}$ is (ϵ, δ)-PIL if and only if it's δ-approximate max information capacity is bounded by ϵ, $I_\infty^\delta(\mathcal{M}) \leq \epsilon$.

Narcissus resiliency can be equivalent to pure max information capacity. Our first result is an equivalence. When $n = 1$, $(\epsilon, 0)$-PIL is essentially equivalent to $(\epsilon, 0)$-Narcissus resiliency. That is, $(\epsilon, 0)$-Narcissus resiliency is just a bound on the non-approximate max information capacity.

Theorem 3. *Let $\epsilon \geq 0$, $n = 1$, $\mathsf{Rec}(x) = \{x\}$ and $\mathcal{F}$ be all possible distributions over $\mathcal{X}$. Then $\mathcal{M}$ is $(\epsilon, 0, \mathcal{F})$-*Rec*-Narcissus resilient if and only if $I_\infty(\mathcal{M}) = \epsilon$.*

Proof. By Yao's minimax principle, without loss of generality we can assume $\mathcal{A}$ is a deterministic function. Consider some $\mathcal{D} \in \mathcal{F}$ and let X denote the sample given to $\mathcal{M}$ and denote the fresh sample as X'. We can first upper bound the ratio of probabilities.

$$\begin{aligned}
\frac{\mathbb{P}\left[\mathcal{A}(\mathcal{M}(X)) = X\right]}{\mathbb{P}\left[\mathcal{A}(\mathcal{M}(X)) = X'\right]} &= \frac{\sum_y \mathbb{P}\left[\mathcal{A}(y) = X | Y = y\right]\mathbb{P}\left[Y = y\right]}{\sum_y \mathbb{P}\left[\mathcal{A}(y) = X'\right]\mathbb{P}\left[Y = y\right]} \\
&\leq \max_y \left\{ \frac{\mathbb{P}\left[\mathcal{A}(y) = X | Y = y\right]\mathbb{P}\left[Y = y\right]}{\mathbb{P}\left[\mathcal{A}(y) = X'\right]\mathbb{P}\left[Y = y\right]} \right\} \\
&\leq \max_{y,x} \left\{ \frac{\mathbb{P}\left[X = x | Y = y\right]\mathbb{P}\left[Y = y\right]}{\mathbb{P}\left[X' = x\right]\mathbb{P}\left[Y = y\right]} \right\} \\
&= 2^{I_\infty(X;Y)}. \qquad (3)
\end{aligned}$$

The first inequality uses that for non-negative sequences the ratio $\frac{a_1+\ldots+a_k}{b_1+\ldots+b_k} \leq \max_{j\in[k]} \left\{\frac{a_j}{b_j}\right\}$.

To complete the proof, we construct an instantiation of $\mathcal{A}$ for which the inequality is tight. Let $(x^*, y^*) = \arg\max_{x,y} \left\{ \frac{\mathbb{P}[X=x|Y=y]}{\mathbb{P}[X=x]} \right\}$. Consider the adversary instantiation,

$$\mathcal{A}(y) = \begin{cases} x^* & \text{if } y = y^* \\ \perp & \text{else} \end{cases}$$

That is, the attacker does not attempt to reconstruct X unless $Y = y^*$. Then we see that $\mathbb{P}\left[\mathcal{A}(\mathcal{M}(X) = X\right] = \mathbb{P}\left[X = x^* | Y = y^*\right]\mathbb{P}\left[Y = y^*\right]$ and $\mathbb{P}\left[\mathcal{A}(\mathcal{M}(X)) = X'\right] = \mathbb{P}\left[X' = x^*\right]\mathbb{P}\left[Y = y^*\right]$. The ratio of these probability exactly gives the RHS of equation 3. This information equality holds for any $\mathcal{D}$, which establishes the capacity equivalence. □

While PIL and Narcissus resiliency are functionally equivalent in this case, they differ on their motivation for why this is useful. Cohen et al. (2024) motivates Narcissus resiliency as using the reconstruction attack itself to define the baseline success rate (and thus delineate which data is non-trivial to reconstruct). We find this non-intuitive given that, in the above proof, we show the worst case adversary never tries to reconstruct X except on one possible output of Y. In this regard, we find the explanation that max information bounds the increase in any adversaries best possible guess for X more direct.

Narcissus resiliency can differ from AMI and does not bound degree of extractability. Given the previous equivalence, a natural next question is whether (ϵ, δ)-Narcissus resiliency leads to an ϵ bound on I_∞^δ. We show an example for which this is not the case and where the approximate max information leakage is much higher than what is suggested by the Narcissus resiliency parameters. Further, it is due to this fact that piecewise information limited algorithms bound the *degree* of extractability and Narcissus resilient algorithms do not. The following counterexample shows that (ϵ, δ)-Narcissus resiliency cannot in general lead to lower bounds on the extraction length greater than $O(\log(1/\delta))$; notably, this extraction length does not improve even as ϵ goes to 0. Finally, the result also shows Narcissus resiliency can be brittle to side information in the following sense. If $\mathcal{M}$ is $(\epsilon, \delta, \mathcal{F})$-Narcissus resilient, there exists a case where an adversary with $\log(1/\delta + 1)$ bits of side information can obtain a reconstruction with probability 1 when given the output of $\mathcal{M}$, even when no adversary with access to the side information alone is able to obtain a reconstruction except with very small probability.

Theorem 4. *Let $\epsilon > 0$, $\delta \in (0, 1/2]$, $n = 1$, $N > 0$, $\mathsf{Rec}(x) = \{x\}$, and $\mathcal{X} = \{0, ..., N-1\}$. There exists a distribution $\mathcal{D}$ and mechanism $\mathcal{M}$ such that $\mathcal{M}$ is $(0, \delta, \mathcal{D})$-Narcissus resilient but for $X \sim \mathcal{D}$ and $Y \sim \mathcal{M}(X)$ the following hold:*

1. ***No AMI bound:*** *$I_\infty^\delta(Y; X) = \Omega(\log(N) - \log(1/\delta))$*
2. ***No extraction length lower bound:*** *There exists a programming language (defined independently of X) such that for any x and $y \in \mathsf{Supp}(\mathcal{M}(x))$, there exists a program, $\mathcal{P}$, of description length at most $\log(1/\delta + 1)$ such that $\mathcal{P}(y) = x$*
3. ***Brittle to side information:*** *There exists an attacker $\mathcal{A}$ with access to side information B such that $I_\infty(B; X) = O(\log(1/\delta + 1))$ and $\mathbb{P}[\mathcal{A}(Y, B) \in \mathsf{Rec}(X)] = 1$, but for any adversary $\mathcal{A}'$ with access to B alone, $\mathbb{P}[\mathcal{A}'(B) \in \mathsf{Rec}(X)] \leq \frac{1}{N\delta}$.*

Proof. Let $\mathcal{X} = \{0, ..., N-1\}$ for some N sufficiently large and let $M = \lceil 1/\delta \rceil$. Let $\mathcal{D}$ be the uniform distribution over $[N]$. Let $\mathcal{M}$ be the mechanism for each $x \in [N-1]$, outputs a random sample from $\mathcal{X}'(x) = \{x \mod N, x+1 \mod N ..., x+M-1 \mod N\}$. Let $Y = \mathcal{M}(X)$. We have,

$$\mathbb{P}[X = x | Y = y] = \frac{\mathbb{P}[Y = y | X = x]\,\mathbb{P}[X = x]}{\mathbb{P}[Y = y]} = \frac{1}{M} \cdot \mathbf{1}_{[y \in \mathcal{X}'(x)]}. \tag{4}$$

Thus conditional on any value of Y, X is uniform over some set of size M. Consequently, the Bayes optimal adversary does not reconstruct X with probability greater than $1/M$, which means $\mathcal{M}$ is $(0, \frac{1}{M}, \mathcal{D})$-Narcissus resilient.

On the other hand, let $\mathcal{O} \subset [N]^2$ denote the set of points (x, y) such that $y \in \mathcal{X}'(x)$. We have,

$$I_\infty^\delta(\mathcal{M}(X); X) \geq \log\left(\frac{\mathbb{P}[(X, Y) \in \mathcal{O}] - \delta}{\mathbb{P}[X \times Y \in \mathcal{O}]}\right) = \log\left(\frac{N(1-\delta)}{M}\right) = \Omega(\log(N) - \log(1/\delta)).$$

This proves the first part of the lemma.

To show how X can be recovered with a short program, consider the family of programs where on some input x, the family contains the programs which output each unique element of $\mathcal{X}'(x)$. These programs can be specified with $\log(M)$ bits. Since $Y \in \mathcal{X}'(X)$ w.p. 1, there always exists a program (of description length $\log(M)$) which satisfies $\mathcal{P}(Y) = x$.

For the side information claim, let $B = X \mod M$. Conditional on any realization of $B = b$, it is straightforward to show that X is uniform over the set $\mathcal{X}_b = \{x \in [N] : b = x \mod M\}$, which has size $|\mathcal{X}_b| \leq N/M$. Thus, $I(B; X) \leq \log(\frac{M/N}{1/N}) = \log(M) \leq \log(1/\delta + 1)$. Simultaneously, we know from equation 4 that when $Y = y$ it holds that $X \in \{x : y \in \mathcal{X}'(x)\}$, w.p. 1. This set contains only one element such that $B = X \mod M$, as such the attacker can always reconstruct X by selecting this element. The fact that no adversary can reconstruct with probability greater than M/N comes from the fact that the conditional distribution of X given B is always uniform over a set of size N/M (shown above). □

Note that in the above, we think of $N \gg 1/\delta$ (i.e. for d-dimensional data, N is exponentially large in d, whereas δ is typically only polynomially small in d or n). We see that approximate Narcissus resiliency allows $\mathcal{M}$ to add substantial information about X so long as the Bayes optimal adversary is prevented from reconstructing X with probability more than δ. By contrast, the way PIL differs in how it handles the slack parameter δ allows us to provide stronger guarantees that allow us to bound the degree of extractability and make it robust to small changes in the adversary's side information.

We provide some additional explanation on why part 3 of the theorem is important. One could rightly point out that the conditional distribution of X after an adversary receives side information B is no longer in the class of distributions for which $\mathcal{M}$ is assumed to be Narcissus resilient, which in this case contains only $\mathcal{D}$. However, rarely can we guarantee $\mathcal{D}$ is in some distribution class exactly. In such case, we would hope that the Narcissus resilient algorithms still provide reasonable protection under distributions close to this class. More specifically, we hope data which is non-trivial to reconstruct under these "adjacent" distributions is still difficult to reconstruct after receiving the output of a Narcissus resilient mechanism. Theorem 4 shows we cannot hope for such behavior in general, even when $\epsilon = 0$.

Remark 1. *The fact that the chosen program in Theorem 4 depends on both the realization of y and x is necessary and consistent with the definition of an extraction in Cohen et al. (2024). Otherwise, no program (of any length) reconstructs X with reasonable probability, as guaranteed by Narcissus resiliency whenever the data is non-trivially reconstructible.*

C Supplement to Section 2

C.1 Generic Form of PIL Guarantees

We here give a generic form of the protection guarantees of PIL algorithms that can be used to easily obtain the copyright and reconstruction protection guarantees, given in subsequent sections. For the purpose of this work, a **programming language** need only be a set of symbols such that some strings made with these symbols delineate a particular function. As a simple example, a set of functions $\mathcal{F}$ could form a programming language where specifying any number of these functions specifies the program which sequentially applies them to some input.

Theorem 5. *Let $\beta' \in [0, 1]$ and $K \in [n]$. Assume that $\mathcal{M}$ is $(\alpha(k), \beta, k)$-PIL for each $k \in [K]$. Let $\mathcal{A}$ be a possibly randomized algorithm and B a random variable. For any realization of $S = \{x_1, ..., x_n\}$ and $B = b$ and define the random variable H^* as $H^* = \max_{\substack{\sigma \subseteq [n] \\ |\sigma| \leq K}} \{H^{\beta'}_{\mathsf{Rec}}(R | X_{\notin \sigma} = x_{\notin \sigma}, B = b) - \alpha(|\sigma|)\} - 1$ and let σ be argmax of the same expression. If $\mathcal{M}$ is run on a uniformly random permutation of S then,*

1. ***Regurgitation Prevention:*** $\mathbb{P}\left[\mathcal{A}(Y, B, X_{\notin \sigma}) \in \mathsf{Rec}(R) \mid H^* \geq \log(1/\sqrt{\beta'})\right] \leq 2\sqrt{\beta' + 2\beta}$.

2. ***Extraction Difficulty:*** *Let $\mathcal{L} = \mathcal{L}(\mathcal{D})$ be a programming language. With probability at least $1 - 2\sqrt{\beta' + 4\beta}$, no program in $\mathcal{L}$ of description length at most $H^* - 2\log(1/\beta')$ given as input Y, B, and $X_{\notin \sigma}$ outputs an element in $\mathsf{Rec}(R)$.*

The application to reconstruction follows from considering the case where $\mathcal{A}$ is some adversary trying to reconstruct R and B is some side information held by this adversary. Proving the copyright protection guarantees are similarly simple using Theorem 5.

Proof of Theorem 1 (Copyright Protection Guarantees). The proof follows from the generic statement given by Theorem 5 where the random variable B is equal to the prompt, $B = P$, the training process $\mathcal{T}$, is the (α, β)-PIL mechanism, and $Y = h = \mathcal{T}(S)$. □

C.1.1 Proof of Theorem 5

Before proving the theorem, it will be useful to establish the following lemma, which controls how the posterior updates given the output of an AMI bounded process.

Lemma 4. *Let $\beta, \beta' \in [0, 1]$ and let X and Y be random variables. Then,*

$$\mathop{\mathbb{P}}_{y \leftarrow Y}\left[H^{\sqrt{\beta'+2\beta}}_{\mathsf{Rec}}(X|Y=y) \geq H^{\beta'}_{\mathsf{Rec}}(X) - I^{\beta}_{\infty}(X;Y) - 1\right] \geq 1 - \sqrt{\beta' + 2\beta}.$$

Proof. For any $y \in \mathsf{Supp}(Y)$, let $\mathsf{Bad}(y) = \{x : \log\left(\frac{\mathbb{P}[x|y]}{\mathbb{P}[x]}\right) \geq I^{\beta}_{\infty}(X;Y)+1\}$. We will first show that $\mathop{\mathbb{P}}_{X,Y}[X \in \mathsf{Bad}(Y)] \leq 2\beta$. To see this, if it were instead the case that $\mathop{\mathbb{P}}_{X,Y}[X \in \mathsf{Bad}(Y)] > 2\beta$, then the definition of max information and the bad set implies

$$\begin{aligned}
I^{\beta}_{\infty}(X;Y) &\geq \log\left(\frac{\left(\sum_y \mathbb{P}\left[X \in \mathsf{Bad}(y)|Y=y\right]\mathbb{P}\left[Y=y\right]\right) - \beta}{\sum_y \mathbb{P}\left[X \in \mathsf{Bad}(y)\right]\mathbb{P}\left[Y=y\right]}\right) \\
&\geq \log\left(\frac{0.5\sum_y \mathbb{P}\left[X \in \mathsf{Bad}(y)|Y=y\right]\mathbb{P}\left[Y=y\right]}{\sum_y \mathbb{P}\left[X \in \mathsf{Bad}(y)\right]\mathbb{P}\left[Y=y\right]}\right) \\
&> \log\left(\frac{2^{I^{\beta}_{\infty}(X;Y)}\sum_y \mathbb{P}\left[X \in \mathsf{Bad}(y)\right]\mathbb{P}\left[Y=y\right]}{\sum_y \mathbb{P}\left[X \in \mathsf{Bad}(y)\right]\mathbb{P}\left[Y=y\right]}\right) \\
&= I^{\beta}_{\infty}(X;Y),
\end{aligned}$$

which is a contradiction. This establishes that $\mathop{\mathbb{P}}_{X,Y}[X \in \mathsf{Bad}(Y)] \leq 2\beta$.

Now let $\mathsf{Bad}' \subset \mathsf{Supp}(X)$ be the largest measure set such that $\forall x \notin \mathsf{Bad}'$, $\mathbb{P}[X = x] \leq 2^{-H^{\beta'}_{\infty}(X)}$. By the definition of smooth min-entropy, we have $\mathop{\mathbb{P}}_{X}\left[X \in \mathsf{Bad}'\right] \leq \beta'$. We thus have, $\mathop{\mathbb{P}}_{X,Y}\left[X \in \mathsf{Bad}(Y) \vee X \in \mathsf{Bad}'\right] \leq 2\beta + \beta'$. Because,

$$\mathop{\mathbb{P}}_{X,Y}\left[X \in \mathsf{Bad}(Y) \vee X \in \mathsf{Bad}'\right] = \mathop{\mathbb{E}}_{y \sim Y}\left[\mathop{\mathbb{P}}_{X}\left[X \in \mathsf{Bad}(Y) \vee X \in \mathsf{Bad}'|Y=y\right]\right],$$

Markov's inequality now implies,

$$\mathop{\mathbb{P}}_{y \leftarrow Y}\left[\mathbb{P}\left[X \in \mathsf{Bad}(y)|Y=y\right] \geq \sqrt{2\beta+\beta'}\right] \leq \sqrt{2\beta+\beta'}.$$

This means that with probability at least $1 - \sqrt{2\beta+\beta'}$ under the draw of Y, there exists a set $\mathcal{O} \subset \mathsf{Supp}(X)$ of measure at least $1 - \sqrt{2\beta+\beta'}$ such that $\forall x \in \mathcal{O}$, $\mathbb{P}[X = x|Y = y] \leq 2^{I^{\beta}_{\infty}(X;Y)+1}\mathbb{P}[X = x]$. This in turn bounds the approximate reconstruction entropy. □

We now prove Theorem 5, noting that we will also use the composition properties of max information given in Appendix F.

Proof of Theorem 5. Condition on any realization $B = b$ and $X_{\neq\sigma} = x_{\neq\sigma}$ throughout. Since $X_{\neq\sigma}$ determines $|\sigma|$, to simplify notation, let $\alpha := \alpha(k)$ throughout the following. Let Y denote the random variable $Y = \mathcal{M}(S)$. By the PIL property of $\mathcal{M}$ and the composition properties of AMI (Lemma 12), we have (recalling we condition on $X_{\notin\sigma} = x_{\notin\sigma}$ and $B = b$),

$$\begin{aligned}
I^{\beta}_{\infty}(R;Y) &\leq I^{\beta}_{\infty}(R;Y,\sigma) \\
&\leq \max_{\sigma'}\{I^{\beta}_{\infty}(R;Y|\sigma=\sigma')\} + I^{0}_{\infty}(R;\sigma) \\
&\leq \alpha. \qquad (5)
\end{aligned}$$

The last inequality uses two facts. First, if the dataset is permuted uniformly at random before applying $\mathcal{M}$, then both the marginal and conditional distribution (conditioning on any realization of $X_{\neq\sigma}$) of σ is a uniformly random subset of $[n]$. Thus, σ is independent of R, implying $I^0_\infty(R;\sigma|X_{\neq\sigma}=x_{\neq\sigma})=0$. Further, since $\mathcal{M}$ is (α,β)-PIL, for any fixed input indices, σ', the β-AMI capacity of the channel formed by $\mathcal{M}$ w.r.t. inputs σ' is at most α.

Now by Lemma 4,

$$\mathbb{P}_{y\leftarrow Y}\left[H^{\sqrt{\beta'+2\beta}}_{\mathsf{Rec}}(R|Y=y)\geq H^{\beta'}_{\mathsf{Rec}}(R)-\alpha-1\right]\geq 1-\sqrt{\beta'+2\beta}.$$

Note the LHS is equal to $\mathbb{P}_{y\leftarrow Y}\left[H^{\sqrt{\beta'+2\beta}}_{\mathsf{Rec}}(R|Y=y)\geq H^*\right]$. Since we have conditioned on any realization of $X_{\notin\sigma}$ and B, we have $\mathbb{P}_{y\sim Y,S,B}\left[H^{\sqrt{\beta'+2\beta}}_{\mathsf{Rec}}(R|Y=y)\geq\log(1/\sqrt{\beta'})\mid H^*\geq\log(1/\sqrt{\beta'})\right]\geq 1-\sqrt{\beta'+2\beta}$. Further, whenever $H^{\sqrt{\beta'+2\beta}}_{\mathsf{Rec}}(R|Y=y,X_{\neq\sigma}=x_{\neq\sigma},B=b)\geq\log(1/\sqrt{\beta'})$, no function of $b,x_{\notin\sigma}$ and y reconstructs R with probability at least $1-\max\{\sqrt{\beta'+2\beta},\sqrt{\beta'}\}\geq 1-\sqrt{\beta'+2\beta}$. A union bound then establishes claim (1).

For the second claim, observe that any program of description length at most L has β-AMI capacity at most $L+\log(1/\beta')$ (see Lemma 3). Letting Z denote the output of the program, we have in a manner similar to equation 5,

$$\begin{aligned}I^{2\beta}_\infty(R;Y)&\leq I^{\beta}_\infty(R;Y,\sigma,Z)\\&\leq\max_{\sigma',y}\{I^\beta_\infty(R;Z|Y=y,\sigma=\sigma')+\max_{\sigma'}\{I^\beta_\infty(R;Y|\sigma=\sigma')\}+I^0_\infty(R;\sigma)\\&\leq L+\log(1/\beta')+\alpha.\end{aligned}\tag{6}$$

Now if $L\leq H^*-2\log(1/\beta')$, we have,

$$\mathbb{P}_{y\leftarrow Y}\left[H^{\sqrt{\beta'+4\beta}}_{\mathsf{Rec}}(R|Y=y,X_{\notin\sigma}=x_{\notin\sigma},B=b,Z=z)\geq\log(1/\beta')\right]\geq 1-\sqrt{\beta'+4\beta}.$$

Finally, we observe that if the reconstruction entropy of R is still at least $\log(1/\beta')$ after conditioning on the output of the program, Z, then Z itself is not in $\mathsf{Rec}(R)$ with probability at least $1-2\sqrt{\beta'+4\beta}$. □

C.2 PIL Generation

We here give a result analogous to Theorem 1 under the weaker condition that only the aggregate process of generating the model and evaluating it is PIL. The regurgitation prevention and extraction difficulty guarantees are largely similar, with the only difference being that we lower bound the program length only for programs which take $h(P)$ and P (instead of h and P) as input.

Theorem 6. *Let $\beta'\in[0,1]$ and $K\in[n]$. Assume that for any fixed prompt $p'\in\mathcal{P}^*$, the aggregate process of drawing $h\sim\mathcal{T}$ and $Y\sim h(p')$ is $(\alpha(k),\beta,k)$-PIL for any $k\in[K]$. For any realization of $S=\{x_1,...,x_n\}$ and $P=p$ define the random variable $H^*=\max_{\substack{\sigma\subseteq[n]\\|\sigma|\leq K}}\{H^{\beta'}_{\mathsf{Rec}}(R|X_{\notin\sigma}=x_{\notin\sigma},P=p)-\alpha(|\sigma|)\}-1$. If $\mathcal{T}$ is run on a uniformly random permutation of S then,*

1. ***Regurgitation Prevention:*** $\mathbb{P}\left[h(p)\in\mathsf{Rec}(R)\mid H^*\geq\log(1/\beta')\right]\leq 2\sqrt{\beta'+2\beta}$.
2. ***Extraction Difficulty:*** *Let $\mathcal{L}=\mathcal{L}(\mathcal{D})$ be a programming language. With probability at least $1-2\sqrt{\beta'+4\beta}$, no program in $\mathcal{L}$ of description length at most $H^*-2\log(1/\beta')$ given as input $h(P)$ and P outputs an element in $\mathsf{Rec}(R)$.*

Proof. The proof follows from the generic statement given by Theorem 5 where the random variable B is equal to the prompt, $B=P$, the training process $\mathcal{T}$, is the (α,β)-PIL mechanism, and $Y=h(P)$ where $h\sim\mathcal{T}(S)$. □

We note that, while the information in an individual prompt is ultimately what matters for bounding the probability $h(P)$ generates something substantially similar to a copyrighted work, the above result does not prevent a more adversarial model user from extracting more information about a work via repeated prompting. By contrast, a PIL training process ensures that the information the model gives is bounded regardless of how many prompts are made. While PIL generation is sufficient for preventing honest users from accidentally duplicating copyrighted work, a PIL training process provides more robust protection from users who, for example, want to use the model to extract copyrighted work that is intended to be kept secret; for example, the model may have been trained on content that is broadly only available after paying a fee. A PIL training process ensures that the only users who can do this are those which already have sufficient information about the copyrighted work. Finally, we also note that a PIL training process may provide a more rigorous baseline for establishing that a training process is "substantially transformative". Recent court opinions have focused on this aspect as instrumental in determining whether training on copyrighted data is fair use (e.g. *Bartz v. Anthropic* and *Kadrey v. Meta*).

C.3 Guarantees for Short Prompts

In the following, $\mathcal{P}$ is some prespecified domain for prompt elements (for example, the set of all ASCII characters or the set of all allowed words).

Theorem 7. *Let $\beta' \in (0,1]$ and $K \in [n]$. Assume that for any fixed prompt $p' \in \mathcal{P}^*$, the aggregate process of drawing $h \sim \mathcal{T}$ and $Y \sim h(p')$ is $(\alpha(k), \beta, k)$-PIL for any $k \in [K]$. For any realization of $S = \{x_1, ..., x_n\}$ define the random variable $H^* = \max_{\substack{\sigma \subseteq [n] \\ |\sigma| \leq K}} \{H^{\beta'}_{\mathsf{Rec}}(R | X_{\notin \sigma} = x_{\notin \sigma}) - \alpha(|\sigma|)\} - 1$ and assume $\mathcal{T}$ is run on a uniformly random permutation of S. Then for any prompt p of length at most $\frac{1}{\log(|\mathcal{P}|)}\big(H^* - 2\log(1/\beta')\big)$, with probability $1 - 2\sqrt{\beta' + 4\beta}$ it holds that $h(p) \notin \mathsf{Rec}(R)$.*

Proof. The proof follows directly from the extraction difficulty guarantee of Theorem 6 by viewing the processing of evaluating $h(p)$ as a program of description length $|p| \cdot \log(|\mathcal{P}|)$. Note we here view the random variable P in Theorem 1 as a dummy constant. □

C.4 Example Guarantees

In this section, we provide an example scenario to illustrate how the guarantees of Theorem 1 behave with respect to scaling of the group size parameter k. The purpose of this example is to provide intuition, and is not intended to reflect a realistic data sampling scenario.

Let $\mathsf{Rec}(x) = \{x\}$. Let each record $U = \{R_i\}_{i \in [N]}$ be sampled from $\mathsf{Unif}(\{0,1\}^d)$, $d \geq N$, except that each $j \in \{2, ..., N\}$, R_j has its j'th bit replaced with the j'th bit of R_1. Samp then creates a dataset by picking some $m \in [n]$, and puts m copies of R_1 into S, and fills the remaining $n - m$ slots with records $R_2, ..., R_{n-m+1}$. The result is a dataset which contains m copies of R_1 and $n - m$ other datapoints which contain a small amount of information about R_1.

Now consider the case where the training procedure $\mathcal{T}$ is $(\alpha(k), \beta, k)$-PIL for any $k \in [n]$, with $\alpha(k) = 16k^2 + 4k\log(4k\sqrt{\pi}/\beta)$. Concretely, this could be obtained via a 16-zCDP training process (see Lemma 2). Now consider some realization of m. Let us assume for simplicity the prompt made to the generative model is empty and the parameter β' in Theorem 1 is 0. For any fixed value of m, we have

$$H^* \geq H^0_{\mathsf{Rec}}(R | X_{\{n-m+1,...,n\}}) - \alpha(m) - 1 \geq d - m - 16m^2 - 4m\log(4m\sqrt{\pi}/\beta) - 1.$$

Thus, by Theorem 1, the model does not regurgitate R_1 with probability at least $1 - \beta$ conditional on the realization of m satisfying $d - m - 16m^2 - 4m\log(4m\sqrt{\pi}/\beta) - 1 \geq 2\sqrt{2\beta}$ (i.e. sufficient deduplication). By contrast, an analysis that uses k large enough to handle all dataset entries that depend on R_1 (rather than relying on information decomposition), would only show that reconstruction is prevented w.h.p. when $d - \alpha(n) - 1 \geq 2\sqrt{2\beta}$, regardless of the setting of m.

C.5 Additional Properties of PIL

C.5.1 Guarantees for Non-permuted Datasets

Similar guarantees to Theorem 5 can be obtained even if the dataset is not permuted uniformly at random (i.e. the index of the datapoint in the dataset matters).

Theorem 8. *Let $\beta' \in [0,1]$ and $K \in [n]$. Assume that $\mathcal{M}$ is $(\alpha(k), \beta, k)$-PIL for each $k \in [K]$. Let $\mathcal{A}$ be a possibly randomized algorithm and B a random variable. For any realization of $S = \{x_1, ..., x_n\}$ and $B = b$ and define the random variable H^* as $H^* = \max_{\substack{\sigma \subseteq [n] \\ |\sigma| \le K}} \{H_{\mathsf{Rec}}^{\beta'}(R | X_{\notin\sigma} = x_{\notin\sigma}, B = b) - \alpha(|\sigma|)\} - K\log(en/K) - 1$. Then,*

1. ***Regurgitation Prevention:*** $\mathbb{P}\left[\mathcal{A}(Y, B, X_{\notin\sigma}) \in \mathsf{Rec}(R) \mid H^* \ge 2\log(1/\beta')\right] \le 2\sqrt{\beta' + 4\beta}$.
2. ***Extraction Difficulty:*** *Let $\mathcal{L} = \mathcal{L}(\mathcal{D})$ be a programming language. With probability at least $1 - 2\sqrt{\beta' + 6\beta}$, no program in $\mathcal{L}$ of description length at most $H^* - 3\log(1/\beta')$ given as input Y and B outputs an element in $\mathsf{Rec}(R)$.*

Proof. The proof follows similarly to the proof of Theorem 5, however when the dataset is not permuted uniformly at random, we have the following decomposition instead of equation 5 (still conditioning on a realization of $X_{\notin\sigma}$ and B),

$$\begin{aligned} I_\infty^{2\beta}(R; Y | X_{\neq\sigma} = x_{\neq\sigma}) &\le \max_{\sigma'} \{I_\infty^\beta(R; Y | \sigma = \sigma', X_{\neq\sigma} = x_{\neq\sigma})\} + I_\infty^\beta(R; \sigma | X_{\neq\sigma} = x_{\neq\sigma}) \\ &\le \alpha + K\log(en/K) + \log(1/\beta'). \end{aligned}$$

Above, we use the fact that σ takes on $\sum_{k=1}^K \binom{n}{k} \le (en/K)^K$, which can be specified with $K\log(en/K)$ bits. The rest of the proof then follows similarly to the randomly permuted case. □

C.5.2 Unconditional Guarantees via Dataset Information

We have the following alternative phrasing of Theorem 5, which bounds unconditional reconstruction probability using the information in the dataset and side information. We present this phrasing primarily for intuition.

Theorem 9. *Let $\beta' \in [0,1]$ and $k \in [n]$. Assume that $\mathcal{M}$ is (α, β, k)-PIL. Let $\mathcal{A}$ be a possibly randomized algorithm and B a random variable. For any realization of $S = \{x_1, ..., x_n\}$ and $B = b$ and define the random variable H^* as $H^* = \max_{\substack{\sigma \subseteq [n] \\ |\sigma| = k}} \{H_{\mathsf{Rec}}^{\beta'}(R) - I_\infty^\beta(R | X_{\notin\sigma}, B) - \alpha\} - 1$. If $\mathcal{M}$ is run on a uniformly random permutation of S then,*

1. ***Regurgitation Prevention:*** $\mathbb{P}\left[\mathcal{A}(Y, B, X_{\notin\sigma}) \in \mathsf{Rec}(R))\right] \le 2\sqrt{\beta' + 2\beta}$.
2. ***Extraction Difficulty:*** *Let $\mathcal{L} = \mathcal{L}(\mathcal{D})$ be a programming language. With probability at least $1 - 2\sqrt{\beta' + 4\beta}$, no program in $\mathcal{L}$ of description length at most $H^* - 2\log(1/\beta')$ given as input Y and B outputs an element in $\mathsf{Rec}(R)$.*

Proof. We have the following decomposition in place of equation 5,

$$\begin{aligned} I_\infty^{2\beta}(R; Y, X_{\notin\sigma}, B) &\le I_\infty^{2\beta}(R; Y, \sigma, X_{\notin\sigma}, B) \\ &\le \max_{\sigma', x_{\notin\sigma}, b} \{I_\infty^\beta(R; Y | \sigma = \sigma', X_{\notin\sigma} = x_{\notin\sigma}, B = b)\} \\ &\quad + \max_{x_{\notin\sigma}, b} \{I_\infty^0(R; \sigma | X_{\notin\sigma} = x_{\notin\sigma}, B = b)\} + I_\infty^\beta(R; X_{\notin\sigma}, B) \\ &\le \alpha + I_\infty^\beta(R; X_{\notin\sigma}, B). \end{aligned}$$

Now applying this in conjunction with Lemma 4 we obtain,

$$\begin{aligned} \mathbb{P}_{y\sim Y, s\sim S, b\sim B}\left[H_{\mathsf{Rec}}^{\sqrt{\beta'+4\beta}}(R | Y = y, \sigma = \sigma', B = b) \ge H_{\mathsf{Rec}}^{\beta'}(R) - I_\infty^{2\beta}(R; X_{\notin\sigma}, B) - \alpha - 1\right] \\ \ge 1 - \sqrt{\beta' + 4\beta}. \end{aligned}$$

The rest of the proof now follows similarly to the proof of Theorem 5 □

C.6 Alternative Information Measures

Shannon's mutual information is the de facto information measure and has indeed been used to quantify to what degree a model has "memorized" training data in several works (Brown et al., 2021; Feldman et al., 2025). While this has proven fruitful for understanding generalization, Shannon information is not well suited to characterizing reconstruction, even for the simple case of exact reconstruction. Specifically, Shannon information is a measure of the expected change in Shannon entropy, and Shannon entropy does not accurately track reconstruction potential. For example, consider a random variable X with support $\{0,1\}^d$, and two possible distributions for X, ρ_1 and ρ_2. Let ρ_1 have $\rho_1(\{0\}^d) = \frac{1}{2}$ and $\rho_1(x) = \frac{1}{2(2^d-1)}$ for every other point. We define ρ_2 as the uniform distribution over a subset of $\{0,1\}^d$ of size $2^{d/2}$. Under either distribution $H(X) = \frac{d}{2} + O(1)$, however under distribution ρ_1 one can guess X (i.e. perfect reconstruction) and succeed with probability $1/2$, and under distribution ρ_2 no guess succeeds with probability more than $2^{-d/2}$. That is, despite similar Shannon entropy, there is a dramatic difference in the ability of an adversary to reconstruct X. Similar observations were made in the context of guessing secrets in Alvim et al. (2020).

Another information measure which has been considered for analyzing reconstruction potential is the Bayes capacity (Biswas et al., 2024). The Bayes leakage between random variables X and Y is $\mathop{\mathbb{E}}_{y\sim Y}\left[2^{H_\infty(X|Y=y)-H_\infty(X)}\right]$ (i.e. the expected change in the best-guess probability for X). The Bayes capacity is the maximum such value over all priors on X. We find approximate max information a better fit for multiple reasons. In contrast to Bayes capacity, I_∞^β readily provide high probability guarantees against reconstruction, whereas Bayes capacity information only provides such guarantees in expectation. Second, on a mechanistic level, approximate max information has a number of desirable properties such as symmetry and composibility. By contrast, Bayes capacity is asymmetric and deriving composition-like guarantees generally requires distribution specific analysis.

Finally, another choice which has been considered in the literature is the Fisher information. Like the previous information measures, this also suffers from only characterizing the expected increase in reconstruction potential under the output of the data analysis mechanism. Further however, it is not clear how to provide guarantees for more complex instantiations of Rec, and previous work has focused only mean squared error related reconstruction distance (Hannun et al., 2021; Guo et al., 2022).

D Supplement to Section 3

D.1 Implementing PIL Algorithms

Differentially Private Mechanisms are PIL Mechanisms.

Lemma 5. *(Restatement) Let $\epsilon > 0$ and $\delta \in [0,1]$. $\mathcal{M} : \mathcal{X}^n \mapsto \mathcal{Y}$ be an (ϵ,δ)-DP algorithm. Then $\mathcal{M}$ satisfies (ϵ,δ)-PIL.*

Proof. This largely follows from lemmas in Rogers et al. (2016). Let $x_{1:n-1}$ be any fixing of the first $n-1$ inputs of $\mathcal{M}$ and let X be some random variable used as the input at position n. Let $Y = \mathcal{M}(x_{n-1}, X)$. First, by (Rogers et al., 2016, Claim 3.3) the distribution over (Y,X) is (ϵ,δ) indistinguishable from the distribution of $Y \times X$. That is, for any $\mathcal{O} \subset \mathcal{Y}\times\mathcal{X}$, $\mathbb{P}[(X,Y)\in\mathcal{O}] \le e^\epsilon \mathbb{P}[X\times Y \in \mathcal{O}] + \delta$. Rearranging this expression shows this bounds the approximate max information, $I_\infty^\delta(Y;X) \le \epsilon$. Since the above holds for any distribution over X and any index for which X is used as input (i.e. we have not used that X was in the last position), this proves that $\mathcal{M}$ is (ϵ,δ)-PIL. □

Lemma 6. *(More general statement of Lemma 2) Let $\xi,\rho \ge 0$. If $\mathcal{M} : \mathcal{X}^n \mapsto \mathcal{Y}$ is an (ξ,ρ)-zCDP algorithm, then $\mathcal{M}$ satisfies $(\xi\cdot k\sum_{i=1}^k \frac{1}{i} + 2\rho k^2 + \sqrt{\rho k^2 \cdot \log\frac{\sqrt{\pi\rho k^2}}{\delta}}, \delta, k)$-PIL for any $k \in [n]$ and $\delta > 0$.*

Proof. This proof largely follows from results in Bun & Steinke (2016). First, by (Bun & Steinke, 2016, Proposition 5.3), a mechanism $\mathcal{M}$ satisfying (ξ,ρ)-zCDP also satisfies $(\xi k\sum_{i=1}^k \frac{1}{i}, \rho k^2)$-

zCDP for groups of size k. Then, via (Bun & Steinke, 2016, Lemma 3.6), we have that mechanism $\mathcal{M}$ provides $(\xi \cdot k \sum_{i=1}^{k} \frac{1}{i} + \rho k^2 + \sqrt{4\rho k^2 \cdot \log \frac{\sqrt{\pi \rho k^2}}{\delta}}, \delta)$-DP for groups of size k, for any $\delta > 0$. Applying the same argument as in the proof of Lemma 1, we have the claim. □

PIL Mechanisms via Conditionally Bounded Description Length.

Lemma 7. *(Restatement) Let $k \in [n]$. Let $\mathcal{M} : \mathcal{X}^n \mapsto \mathcal{Y}$ be a (possibly randomized) mechanism such that for any fixing of $n - k$ of its inputs and any fixing the random seed of $\mathcal{M}$, the induced range (with respect to the remaining inputs) of $\mathcal{M}$ is discrete with cardinality at most M. Then for any $\beta \in (0, 1]$, $\mathcal{M}$ satisfies $(\log(M/\beta), \beta, k)$-PIL.*

Proof. Consider the (possibly randomized) channel formed by $\mathcal{M}$ after fixing to any value any $n-k$ of its inputs. To prove $\mathcal{M}$ is $(\log(M/\beta), \beta, k)$-PIL, it suffices to show that this channel has β-AMI capacity $\log(M/\beta)$ for any such fixing. This fact follows directly from (Dwork et al., 2015, Theorem 30), restated below. □

Theorem 10 (Theorem 30 of Dwork et al. (2015))**.** *Let $\mathcal{A}$ be an algorithm with randomized description length k taking as input an $n-$element dataset and outputting a value in $\mathcal{Y}$. Then for every $\beta > 0$, $I_\infty^\beta(\mathcal{A}, n) \leq \log(|\mathcal{Y}|/\beta)$.*

Composition and Subsampling. It is straightforward to show PIL mechanisms compose additively in their parameters.

Lemma 8. *Let $\mathcal{M}_1, ..., \mathcal{M}_T$ be a sequence of possibly adaptive mechanisms such that mechanism $\mathcal{M}_t$ is (α_t, β_t, k)-PIL. Then the aggregate mechanism formed by applying each mechanism to S is $(\sum_{t=1}^{T} \alpha_t, \sum_{t=1}^{T} \beta_t, k)$-PIL.*

Proof. This follows directly from the composition properties of approximate max information capacity, see Lemma 12 in Appendix F. □

Alternatively, we can obtain more nuanced results when not every datapoint is used in each PIL mechanism.

Lemma 9. *Let $\mathcal{M}_1, ..., \mathcal{M}_T$ be a sequence of possibly adaptive mechanisms acting on subsets of dataset $S \in \mathcal{X}^n$, $S_1, ..., S_T$, which may be selected randomly and adaptively. Assume that each $\mathcal{M}_t$ is (α, β, k)-PIL. Then if there exists $r \in [T]$ such that $\forall G \subset S$ s.t. $|G| = k$, $\mathbb{P}\left[\sum_{t=1}^{T} \mathbf{1}_{\{|S_t \cap G| > 0\}} \geq r\right] \leq \beta'$, then the aggregate mechanism formed by sequentially applying $\mathcal{M}_1, ..., \mathcal{M}_T$ is $(r\alpha, r\beta + \beta', k)$-PIL.*

Proof. Fix some G and condition on the event that no datapoint in G is fed to more than k mechanisms, which by assumption happens with probability at least $1 - \beta'$. Conditional on this event, the $(k\beta)$-AMI capacity of the channel w.r.t. the inputs in G is $k\alpha$ by composition (only mechanisms which act on the datapoint leak information about it). Finally, we obtain a $k\beta + \beta'$ approximate max information bound after accounting for the small probability event that a datapoint can be in more than k mechanisms. □

For example, when sampling uniformly random subsets of size m from the data for each mechanism, it is known that the number of times any point will be used in the mechanisms concentrates tightly around the expectation, Tm/n. We can use this to instantiate the result with uniformly random batches.

Lemma 10. *Let $T \in \mathbb{Z}^+$ and $m \in [n/2]$ and assume $Tm/n \geq 5\log(1/\beta)/24$. Let $\mathcal{M}_1, ..., \mathcal{M}_T$ each be (α, β)-PIL and let $S_1, ..., S_T$ each be subsets of S of size m sampled uniformly at random. Then the aggregate mechanism formed by sequentially applying each $\mathcal{M}_t$ to each S_t and all previous outputs is $(3\frac{Tm}{n}\alpha, 3\frac{Tm}{n}\beta + \beta)$-PIL.*

Proof. Consider any datapoint $x \in S$ and let p be the number of subsamples that x appears in. By Bernstein's bound we have that,

$$\mathbb{P}\left[p \geq 3\frac{Tm}{n}\right] \leq \exp\left(-\frac{12Tm}{5n(1-m/n)}\right) \leq \exp\left(-\frac{24Tm}{5n}\right) \leq \beta.$$

Above we have used the assumed lower bound on Tm/n. The lemma statement then follows from Lemma 9. □

D.2 Releasing Exact Histograms via a Conditionally Bounded Description Length Mechanism

Let $\mathcal{X} = [N]$ and consider the problem of releasing a histogram of the dataset $S = x_1, ..., x_n \in \mathcal{X}^n$ such that element $j \in [N]$ of the histogram H has $H_j = |\{i : x_i = j\}$. Using a conditionally bounded description length mechanism, we can release the exact histogram by trading off α and k. In contrast to general bounded description length mechanisms (i.e. $k = n$), this can lead to dramatically smaller α when k is small.

Lemma 11. *Let $\beta \in (0, 1]$ and $k \in [n]$. There exists a deterministic mechanism, $\mathcal{M}$, which is (α, β, k)-PIL with $\alpha = k\log(N) + \log(1/\beta)$.*

Proof. Fix any subset $\sigma \subseteq [n]$ such that $|\sigma| = k$. Consider the partial histogram formed over the set $\tilde{S} = \{x_i : i \in [n] \setminus \sigma\}$; that is, $\tilde{H}_j = |\{i \in [n] \setminus \sigma : x_i = j\}|$. The final histogram can be determined with $\tilde{H}_j$ and $k\log(N)$ bits specifying the bin of the remaining k dataset elements. The claim then follows from Lemma 3. □

E Supplement to Section 4

E.1 Proof of Theorem 2

By part (ii) of Theorem 1 with the random variable P set as a dummy constant, no program of description length at most $H^{\beta'}_{\mathsf{Rec}}(X') - \alpha - 1 - 2\log(1/\beta')$ reconstructs X' with probability more than $2\sqrt{\beta' + 4\beta}$. Observe $\min_{p\in\mathcal{P}^*}\{|p|\cdot\log(|\mathcal{P}|)+d_{\mathcal{F}}(y,h(p))\cdot\log(|\mathcal{F}|)|\}$ is the description length of the program which, given h as input, reconstructs X'. Specifically, the program can be specified by specifying the prompt with $|p|\log(|\mathcal{P}|)$ bits and the edit operations with $d_{\mathcal{F}}(y, h(p)) \cdot \log(|\mathcal{F}|)|$ bits. The lower bound on m^* then follows from the program length guarantee of Theorem 1.

E.2 Memorization Testing with our Data Extractor

In this section, we describe how our data extraction procedure can serve as a memorization test. While our extraction procedure bears certain similarities to the process of Schwarzschild et al. (2024), we here show how our test fits into the framework of Cohen et al. (2024). We note strictly Cohen et al. (2024) discusses extraction quality, but the idea is morally equivalent to ideas in existing memorization tests; that is, they compute a score which reflects the difficulty of extracting the datapoint x from the output, h, of the data analysis procedure (in this case, the model training process), and then normalize this quantity using the difficulty of generating x without h.

To see how our test fits into this framework, first recall the prompt alphabet $\mathcal{P}$ and family of edit operations, $\mathcal{F}$, detailed in Section 4.1. Consider the family of programs which take as input a reconstruction target x and optionally a generative model h. Each program in the language is specified by some prompt in $\mathcal{P}^*$ and series of edit operations in $\mathcal{F}$. When h is not provided to the program, the program sequentially applies the edit operations to a default starting value, such as the empty string, ϕ. When h is received as input, the program first generates $y = h(p)$, where p is the specified prompt, and then outputs the result of sequentially applying the specified edit operations of $\mathcal{F}$ to y.

When such a program successfully extracts x, we can define its degree of memorization (or extraction quality) as $1 - \frac{m^*}{d_{\mathcal{F}}(x,\phi)}$, where m^* and $d_{\mathcal{F}}$ are as defined in Section 4.1. Here, scores closer to 1 indicate a higher degree of memorization.

E.3 Data Extraction Algorithm

Algorithm 1 details the algorithm we implement to obtain an upper bound on m^* defined in equation 1. The procedure works by searching over programs (i.e. prompts and edit operations) of increasing length, to find the shortest program that reconstructs the target. ExtData uses GreedyCoordinateGradient of Zou et al. (2023) (Algorithm 2) as a subroutine to optimize the fixed length prompt to reconstruct the target. We note that Schwarzschild et al. (2024) demonstrate that GreedyCoordinateGradient effectively finds short prompts that extract training data, but not arbitrary sequences of tokens. Then, the procedure takes the optimized prompt returned from Algorithm 2 and finds the shortest sequence of edits that transforms the output of the prompt into the reconstruction target. Since GreedyCoordinateGradient is not guaranteed to exactly find the shortest program to reconstruct the target, the resulting estimate for m^* that Algorithm 1 returns is an upper bound for the shortest program length.

Algorithm 1 Data Extraction Procedure: $\mathsf{ExtData}_{\mathcal{P},\mathcal{F}}(y, h)$

Require: Reconstruction target $y \in \mathcal{X}$, model h, program vocabulary $\mathcal{P}$, edit operations $\mathcal{F}$

1: Length threshold: $L = \frac{|y|}{\log(|\mathcal{P}|)}$
2: Define $\mathcal{L}(y|p;h)$ as the autoregressive next token prediction loss over y given p as the context
3: **for** $l = 1, ..., L$ **do**
4: $\quad p^{(l)} = \mathsf{GCG}(\mathcal{L}, h, \mathcal{P}, y, l, 100, 500)$
5: $\quad F^{(l)} = \arg\min_{F \in \mathcal{F}^D} \{d_{\mathcal{F}}(y, h(p^{(l)}))\}$, where $D = D_{\mathcal{F}}(\mathcal{X})$
6: $\quad m^{(l)} = l \cdot \log(|\mathcal{P}|) + |F^{(l)}| \cdot \log(|\mathcal{F}|)|$
7: **end for**
8: **Output:** $\min_{l \in [L]} \{m^{(l)}\}$

Algorithm 2 GreedyCoordinateGradient: $\mathsf{GCG}(\mathcal{L}, h, \mathcal{P}, y, l, B, n)$ Zou et al. (2023)

Require: Loss $\mathcal{L}$, model h, program vocabulary $\mathcal{P}$, reconstruction target $y \in \mathcal{X}$, prompt length l, batch size B, number of steps n
Initialize prompt p to a random list of l tokens from $\mathcal{P}$

$E = h$'s embedding matrix
for n times **do**
$\quad$ **for** $i = 0, ..., l$ **do**
$\quad\quad \sigma_i = \text{Top-k}(-\nabla_{e_{x_i}} \mathcal{L}(y|p))$
$\quad$ **end for**
$\quad$ **for** $b = 1, ..., B$ **do**
$\quad\quad \tilde{p}^{(b)} = p$
$\quad\quad \tilde{p}^{(b)} \sim \text{Uniform}(\sigma_i), i \sim \text{Uniform}([1, ..., l])$
$\quad$ **end for**
$\quad p = \tilde{p}^{(b^*)}$ where $b^* = \arg\min_b \{\mathcal{L}(y|\tilde{p}^{(b)})\}$
end for
Output: p

E.4 Additional Experimental Details

In our experiments, the edit distance in ExtData is taken to be the Levenshtein distance between the output tokens generated by the model and the target tokens (which we call the token-level edit distance). We also conducted experiments using the Levenshtein distance between the output string generated by the model and the target string (which we call the string-level edit distance). In both cases, by viewing each operation as a token/character swap and adding an additional element to the vocabulary for deletions, we can bound $|\mathcal{F}| \leq |\mathcal{V}| + 1$ and thus in line 6 of Algorithm 1 we have $m^{(l)} = l \cdot \log(|\mathcal{P}|) + d_{\mathcal{F}}(y, h(p)) \cdot \log((|\mathcal{V}| + 1) \cdot L)$, where $\mathcal{V}$ is the domain in which the edits occur. Since the edit distance was penalized less for the string-level edit distance, it resulted in ExtData finding longer prompts and, in turn, returning longer program lengths, see Figure 1. For MiniPrompt (Algorithm 3), since edit distances are not accounted for, we have that the program length is $l \cdot \log(|\mathcal{P}|)$, where l is the length of the shortest prompt that outputs the target.

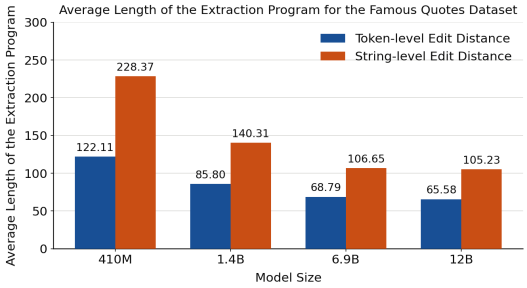

Figure 1: **Average length of the extraction program for EXTDATA on the Famous Quotes dataset run on both the token-level and string-level edit distances.** For the token-level edit distance $V = 50,254$ and for the string-level edit distance $V = 1,023$. Token-level edit distance resulted in shorter program lengths across all model sizes.

Model Size	EXTDATA	MINIPROMPT
410M	1.32	29.07
1.4B	1.22	21.16
6.9B	1.17	16.53
12B	1.11	16.59

Table 1: **Average prompt lengths for EXTDATA and MINIPROMPT on the Famous Quotes dataset.**

E.5 MEMORIZATION TESTING ALGORITHM OF SCHWARZSCHILD ET AL. (2024)

Algorithm 3 MiniPrompt: $\text{MiniPrompt}_{\mathcal{P}}(y, h, max)$ (Schwarzschild et al., 2024)

Require: Reconstruction target $y \in \mathcal{X}$, model h, program vocabulary $\mathcal{P}$, maximum prompt length max

$l = 5, l_min = 0, l_max = max$
Define $\mathcal{L}(y|p; h)$ as the autoregressive next token prediction loss over y given p as the context
$m = \infty$
while $l > l_min$ **and** $l < l_max$ **do**
  $p = \mathsf{GCG}(\mathcal{L}, h, \mathcal{P}, y, l, 100, 500)$
  if $h(p) = y$ **then**
    $l_max = l$
    $m = l \cdot \log(|\mathcal{P}|)$
    $l = l - 1$
  else
    $l_min = l$
    $l = l + 5$
  end if
end while
Output: m

F COMPOSITION OF APPROXIMATE MAX INFORMATION

Dwork et al. (2015) provided a composition theorem for random variables with bounded max information. We provide a slightly more general statement here that allows for release of the entire transcript and does not require capacity bounds. We also correct an issue with the original proof.

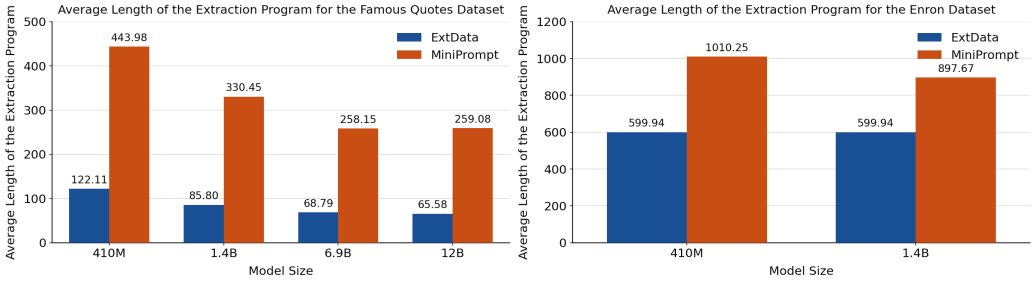

Figure 2: **Average length of the extraction program for EXTDATA and MINIPROMPT on the Famous Quotes (left) and Enron emails (right) datasets.** The added flexibility afforded to EXTDATA results in shorter program lengths than MINIPROMPT across all models on both datasets. As a result, Algorithm EXTDATA audits memorization of both datasets tighter than Algorithm MINIPROMPT.

Lemma 12. *Let X, Y, Z be random variables and denote $\mathcal{P}_Y(X)$ as the channel which maps X to Y and $\mathcal{P}_{Z,y}(X; y)$ as the channel which maps X to Z when $Y = y$. Then,*

$$I_\infty^{\beta_1+\beta_2}(X; Y, Z) \leq \max_{y\in\mathcal{Y}}\{I_\infty^{\beta_2}(Z; X|Y = y)\} + I_\infty^{\beta_1}(X; Y)$$
$$\leq \max_{y\in\mathcal{Y}}\{I_\infty^{\beta_2}(\mathcal{P}_{Z,y})\} + I_\infty^{\beta_1}(\mathcal{P}_Y).$$

Before beginning, we introduce some notation and a useful lemma. Let ρ_X denote the probability mass function of the random variable X and similarly for other random variables. We define the β-approximate max divergence between random variables A and B with supports $\mathcal{A}$ and $\mathcal{B}$ as $D_\infty^\beta(A||B) = \log\left(\max_{\mathcal{O}\subseteq\mathcal{A}:\mathbb{P}[A\in\mathcal{O}]\geq\beta} \frac{\mathbb{P}[A\in\mathcal{O}]-\beta}{\mathbb{P}[B\in\mathcal{O}]}\right)$. We will call the function $\tilde{\rho}$ a β-trimmed subdistribution distribution of the distribution ρ if it is non-negative, is upper bounded by ρ, and has total measure at least $1-\beta$. That is, $\tilde{\rho}$ is formed by removing β mass from ρ.

Lemma 13. *(Implicit in (Dwork & Roth, 2014, Lemma 3.17)) Let $\delta, k > 0$ and A and B be random variables over support $\mathcal{A}$. It holds that $D_\infty^\beta(A||B) \leq$ if and only if there exists a β-trimmed subdistribution of A, $\tilde{\rho}_A$, such that $\max_{a\in\mathcal{A}}\{\log(\tilde{\rho}_A(a)/\rho_B(a))\} \leq k$.*

The original statement was given in terms of constructing a new random variable within β total variation distance of A, but this statement also directly falls out of the proof.

Proof of Lemma 12. As characterized in Dwork et al. (2015), $I_\infty^\beta(A; B) = D_\infty^\beta((A, B)||A \times B)$. As such, our goal will be to find a $(\beta_1 + \beta_2)$-trimmed subdistribution of (X, Y, Z), $\tilde{\rho}_{X,Y,Z}$, such that $\tilde{\rho}_{X,Y,Z} \leq 2^{k_1+k_2}\rho_X(x)\rho_{Y,Z}(y, z)$ for all x, y, z. This will establish the claim by Lemma 13.

To start, by the definition of max information, we have that there exists a β_1-trimmed subdistribution over $\mathcal{X} \times \mathcal{Y}$ such that for any x, y one has $\tilde{\rho}_{X,Y}(x, y) \leq \min\{\rho_{X,Y}(x, y), 2^{k_1}\rho_X(x)\rho_Y(y)\}$. Define $r_1(x, y) = \tilde{\rho}_{X,Y}(x, y)/\rho_{X,Y}(x, y)$ when $\rho_{X,Y}(x, y) > 0$ and 0 otherwise. Similarly, for any $y \in \mathcal{Y}$ there exists a β_2-trimmed distribution, $\tilde{\rho}_{X,Z|y}$ such that $\tilde{\rho}_{X,Z|y}(x, y) \leq \min\{\rho_{X,Z|y}(x, z), 2^{k_2}\rho_{X|y}(x)\rho_{Z|y}(z)\}$ and define $r_{2,y}(x, z) = \tilde{\rho}_{X,Z|y}(x, y)/\rho_{X,Z|y}(x, z)$ when $\rho_{X,Z|y}(x, z) > 0$ and 0 otherwise.

Observe both r_1 and $r_{2,y}$ (for any $y \in \mathcal{Y}$) have range in $[0, 1]$, and so we will construct our trimmed subdistribution of (X, Y, Z) as,

$$\tilde{\rho}_{X,Y,Z}(x, y, z) := \rho_{X,Y,Z}(x, y, z)r_1(x, y)r_2(x, z) \leq \rho_{X,Y,Z}(x, y, z), \quad \forall x, y, z.$$

To see that $\tilde{\rho}_{X,Y,Z}$ has total measure at least $1 - \beta_1 - \beta_2$, we have,

$$\begin{aligned}
1 - \sum_{x,y,z} \tilde{\rho}_{X,Y,Z}(x,y,z) &\leq 1 - \mathbb{E}\left[r_1(X,Y) r_{2,Y}(X,Z)\right] \\
&\leq 1 - \mathop{\mathbb{E}}_{X,Y}\left[r_1(X,Y)\right] + 1 - \mathop{\mathbb{E}}_{X,Z,Y}\left[r_{2,Y}(X,Z)\right] \\
&= 1 - \sum_{x,y} \tilde{\rho}_{X,Y}(x,y) + 1 - \sum_{x,y,z} \tilde{\rho}_{X,Z|y}(x,z)\rho_Y(y) \\
&\leq 1 - \beta_1 - \beta_2.
\end{aligned}$$

The first inequality uses the fact that r_1 and $r_{2,y}$ have range in $[0,1]$. To finish the proof, it suffices to show $\tilde{\rho}_{X,Y,Z}(x,y,z) \leq 2^{k_1+k_2}\rho_X(x)\rho_{Y,Z}(y,z)$. In this regard we have,

$$\begin{aligned}
\tilde{\rho}_{X,Y,Z}(x,y,z) &= \rho_{X,Y,Z}(x,y,z) r_1(x,y) r_2(x,z) \\
&\leq \rho_{X,Y,Z}(x,y,z) \frac{2^{k_1}\rho_X(x)\rho_Y(y)}{\rho_{X,Y}(x,y)} \frac{2^{k_2}\rho_{X|y}(x)\rho_{Z|y}(z)}{\rho_{X,Z|y}(x,z)} \\
&= 2^{k_1+k_2}\rho_X(x)\rho_{Y,Z}(y,z) \frac{\rho_{X,Y,Z}(x,y,z)\rho_{X|y}(x)}{\rho_{X,Y}(x,y)\rho_{X,Z|y}(x,z)} \\
&= 2^{k_1+k_2}\rho_X(x)\rho_{Y,Z}(y,z).
\end{aligned}$$

This concludes the proof. □