The Access–Similarity Lens: An Operational Copyright Framework for Generative Models

Amit Saha
Georgia Institute of Technology
asaha92@gatech.edu

Yinan Huang
Georgia Institute of Technology
yhuang903@gatech.edu

Pan Li
Georgia Institute of Technology
pli@gatech.edu

Eli Chien
National Taiwan University
elichientwn@gmail.com

Abstract

Training generative models on massive datasets has raised concerns that training on copyrighted works may constitute infringement. Recent legal analyses argue that if a model retains *intrinsic, work-specific, training-induced* information about a specific copyrighted work, the model itself is a "copy" of the work, in the sense of copyright law. Motivated by U.S. copyright doctrine, we develop an operational and auditable framework for quantifying claims of model copying that focuses on two necessary evidentiary conditions: Access, that the model was trained on the work, and Similarity, that protected expression from the work is reconstructible from the model. Our framework formalizes these notions through worst-case adversarial games, parameterized by auxiliary information and model access, and limits the success of adversaries performing membership inference for Access and data reconstruction for Similarity. We show that our framework is achievable, controls the risk of generating copyrighted content, and that existing copyright protections fail to provide meaningful guarantees under our framework. We also empirically evaluate our framework on image diffusion models and language models. Altogether, our framework provides operational, quantitative, and auditable evidence to inform copyright litigation for generative models.

1 Introduction

Generative AI models have achieved remarkable success, driven by training on extensive and diverse datasets spanning various domains, including images, text, code, music, and more [1, 2, 3, 4, 5, 6, 7]. However, this success has raised concerns about whether training models on copyrighted data constitutes copyright infringement [8, 9, 10, 11]. Such concerns have even resulted in litigation, such as the recent lawsuit filed by The New York Times against OpenAI for *copying* millions of the Times's copyrighted works[1]. Excluding all copyrighted content is considered impractical [12], and training on copyrighted material does not necessarily result in infringement [13, 14]. This highlights the necessity for understanding the conditions under which a generative model infringes on the copyright of its training data.

Determining infringement is complex, as the outcome of claims frequently relies on factors that vary widely between cases, such as *de minimis* copying, fair use, and market harm; analyzing such factors is widely considered algorithmically intractable [13]. Irrespective of other factors involved in an infringement claim, a necessary condition for an infringement claim is *proof of copying* [14].

[1] https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

Preprint.

In the context of generative AI, sufficient *proof of copying* substantively indicates that an artifact of the model is influenced by the copyrighted work. One direction, motivated by previous work [13] and litigation, treats model output as an artifact. However, model generations sensitively depend on user-provided input and can be considered model misuse, a defense employed by OpenAI in copyright litigation[2]. A second, stricter view, promoted by recent technical and legal analyses [14], treats the trained model itself as the relevant artifact, and asks whether the model's weights encode protected expression from a particular work. Under this perspective, providing evidence for copying requires showing that the model retains *intrinsic, work-specific, training-induced* information about the work. Showing this then constitutes meaningful evidence that the *model itself* is a "copy" of the copyrighted work, in the sense of copyright law.

Consequently, focusing on this second view, we consider necessary evidentiary conditions for a model to constitute a copy of copyrighted content by the Ninth Circuit[3]. The plaintiff (those owning the copyright) must demonstrate Access, that the defendant's model was trained on some form of the copyrighted content, and Similarity, that the model contains reconstructible expression highly similar to the copyrighted content.

Our Contributions. We develop a framework for quantifying model-level copying risk via evidential criteria for Access and Similarity. We formalize these notions by studying the properties of a meaningful accusation of (i) accessing or (ii) being able to reconstruct copyrighted data. A reasonable definition of copyright protection should therefore limit the ability of an adversary to (i) infer whether a given copyrighted sample was used in training, or (ii) generate a reconstruction that is similar to the copyrighted sample. We show that our notions of protection are auditable, achievable, that the popular Near Access-Freeness framework [15] can leak copyrighted data, and that our framework protects against the generation of copyrighted content. We support our theoretical analysis with an empirical evaluation of our framework on image diffusion models and language models, supporting the use of our framework to inform copyright litigation.

2 Related Work

Understanding copyright for generative AI. Legal literature has studied how U.S. copyright law applies to generative AI. In particular, the works of Lee et al. [8], Cooper and Grimmelmann [14] emphasize a model-centric view of infringement, hinging on whether the trained model contains extractable, work-specific protected expression induced by training. We view this perspective as a key factor motivating our framework's design.

Provable copyright protection. Recent work proposes preventative notions of copyright protection by formalizing conditions under which a model is unlikely to produce copied generations. Vyas et al. [15] propose *Near Access-Freeness* (NAF), which limits the distance between the output distributions of models trained with copyrighted data to counterfactual models trained without this data. Existing critiques point out NAF's misinterpretations of copyright law [13] and construct counterexamples, a line of work which we extend [16]. Additionally, Cohen [16] introduces *blameless copy protection*, which aims to limit copying risk for non-malicious users . We show a similar guarantee using our framework with fewer assumptions. While our work has implications towards provable copyright protection (see Sections 4.2 and 4.3), our framework is explicitly distinct from output distributions, and focuses instead on a model-centric view of infringement. Furthermore, previous copyright frameworks lack an *auditability* property [15, 16], which our proposed framework enjoys.

In the interest of brevity, we discuss additional related work in Appendix A.

3 A Framework for Access and Similarity

In this section, we discuss the motivation and setting of our work, the framework, and its key properties. We summarize the operational form of our framework in Fig. 1.

[2] https://openai.com/index/openai-and-journalism/

[3] https://www.ce9.uscourts.gov/jury-instructions/node/261

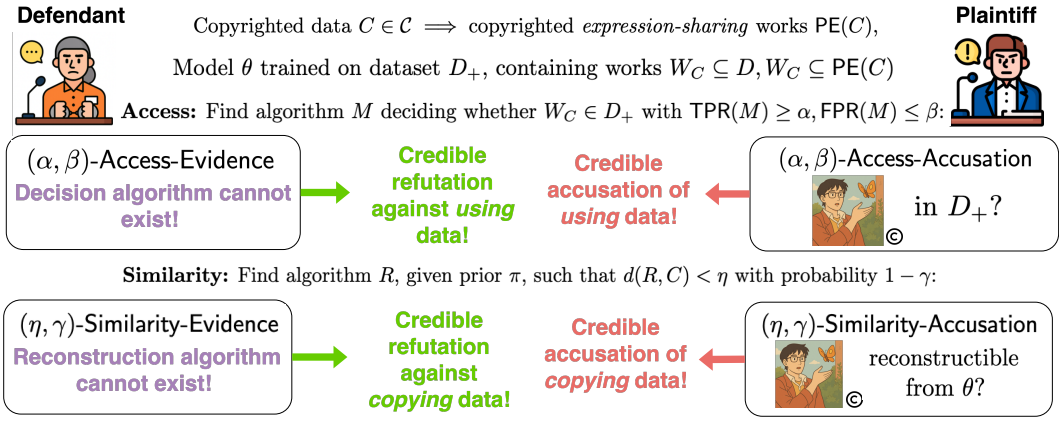


Figure 1: Operational form of copyright evidence framework for establishing Access and Similarity, for a model trained on set $\mathsf{PE}(C)$ which shares *expression* with copyrighted data C (see Section 3).

3.1 Legal Preliminaries and Problem Setting

Before we outline our framework, we refine the translation between the legal concepts we aim to address, the process of training a model, and the intended goals of our framework. **Notation.** Let $\mathcal{W}$ be the set of all works (e.g. images, documents), and let $\mathcal{C}$ be the set of copyrighted works. We generally consider $C \in \mathcal{C}$. We denote the training algorithm $\mathsf{Train} : 2^{\mathcal{W}} \to \Theta$ mapping a dataset $D_+ \subseteq \mathcal{W}$ to a model parameter $\theta \in \Theta$. We use p_θ to denote the model's output distribution.

From works to protected expression. We consider the formal notion of a work, which is the result of *specific expression* resulting in the work's creation. A copyright applies to the specific expression associated with a work, not with its ideas, which are explicitly uncopyrightable [17]. We refer to these expression-defining elements, which may be copyrighted, as *protected expression*.

Now, consider a copyrighted work $C \in \mathcal{C} \subseteq \mathcal{W}$. The works sharing protected expression with C will form a class of protected expression $\mathsf{PE}(C)$. Intuitively, $\mathsf{PE}(C)$ includes exact copies and works preserving protected expression, but excludes works that share only un-copyrightable components. Assume that we train on a dataset D_+ satisfying $D_+ \cap \mathsf{PE}(C) \neq \emptyset$. We define this intersection below, which describes the set of elements in D_+ sharing protected expression with C. Naturally, we also define $D_- = D_+ \setminus W_C$, the training elements that do not share protected expression with C.

Definition 3.1 (Protected Expression Set)**.** We say that W_C is the *protected expression set* with respect to some copyrighted $C \in \mathcal{C}$ and dataset D_+ with $W_C = D_+ \cap \mathsf{PE}(C) \neq \emptyset$.

Training a copying model. Consider the case where we train a model $\theta \leftarrow \mathsf{Train}(D_+)$. Critically, θ depends on the copyrighted data C through expression-sharing works $W_C \subseteq D_+$. In particular, note that W_C *may not contain* C, making the dependence of θ on the data C challenging to ascertain.

The litigation process. If the copyright holder suspects that the model weights θ infringe on their copyright, then they will enter litigation as the plaintiff against the model provider as a defendant. In this setting, it is challenging for both the defendant and the plaintiff to make meaningful statements about Access or Similarity. For Access, a naïve approach is to examine D_+ to find elements of W_C. However, D_+ may be extremely large, and the expression-sharing elements W_C may not be known a priori, making this challenging [8]. For Similarity, examining the high-dimensional generative distribution p_θ and weights θ themselves is computationally infeasible. With this in mind, our framework isolates the elements of a meaningful accusation by a plaintiff and uses these to produce meaningful protections for the defendant, providing evidence for or against Access and Similarity.

3.2 Evidence-based Criteria for Access

To establish Access, the plaintiff must demonstrate that the defendant's training data involved W_C. This leads to an adversarial setting between the plaintiff and the defendant: the defendant trains a generative model, while the plaintiff seeks to determine whether the model has utilized the copyrighted data based on the model and the dataset. We formalize this game below.

Definition 3.2 (Access inference game)**.** The defendant flips a fair coin b. If the outcome is heads $(b = 1)$, the defendant returns $\theta \leftarrow \mathsf{Train}(D_+)$, which involves copyrighted samples W_C. Otherwise, the defendant returns $\theta \leftarrow \mathsf{Train}(D_-)$. The plaintiff employs a decision algorithm M to deduce the value of b using the following information: (1) the dataset D_- and the copyrighted data C, and (2) access to θ. M outputs $\widehat{b} = M(D_-, C, \theta) \in \{0, 1\}$.

This can also be viewed as the membership inference game [18], although we assume a strictly *weaker* threat model, where none of the elements of W_C are explicitly known to the plaintiff. In this setting, Definition 3.2 defines a decision problem whose difficulty is an intrinsic property of Train and the pair (D_+, D_-). Let $\mathcal{M}$ denote a class of decision algorithms for the Access inference game. For brevity, we will often elide the dependence of M on (D_-, C) and write $M(\theta)$.

Success Criterion in the Access **Inference Game.** We evaluate the plaintiff's utility by the tradeoff between detecting access when it occurred and avoiding false accusations when it did not. Any M induces the true positive rate and false positive rate

$$\mathsf{TPR}(M) = \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}\big(M(\theta) = 1\big), \quad \mathsf{FPR}(M) = \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_-)}\big(M(\theta) = 1\big),$$

where probabilities are taken over all randomness in Train, in the plaintiff's access to θ, and in M. Although there exist other metrics for evaluating success in decision algorithms (e.g. the ROC curve), such metrics equally prioritize true positives and false negatives: to have more fine-grained control on M, we instead consider $\mathsf{TPR}(M)$ and $\mathsf{FPR}(M)$.

We first focus on the Access inference game from the plaintiff's perspective. Here, the plaintiff's accusation takes the form of a decision algorithm M. A nontrivial accusation as evidence in litigation should achieve non-negligible detection power while controlling false accusations. Therefore, we propose the following definition, which parametrizes the significance level of an accusation.

Definition 3.3 ((α, β)-Access-Accusation)**.** A plaintiff's accusation, or equivalently their decision algorithm $M \in \mathcal{M}$, satisfies (α, β)-Access-Accusation with respect to the copyrighted data C, works W_C, and datasets (D_+, D_-) if

$$\mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}\big(M(\theta) = 1\big) \geq \alpha, \quad \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_-)}\big(M(\theta) = 1\big) \leq \beta.$$

If the plaintiff's decision algorithm satisfies (α, β)-Access-Accusation with large α and small β, then the accusation has a high probability of discerning whether W_C was used to train the model.

Conversely, the defendant, who wishes to disprove Access, seeks to increase the intrinsic difficulty of this decision problem, so that any accusation with a controlled false positive rate necessarily has low power. With this in mind, we propose the following definition.

Definition 3.4 ((α, β)-Access-Evidence)**.** A generative model θ satisfies (α, β)-Access-Evidence with respect to copyrighted data C, works W_C, and dataset D_- if

$$\sup_{M \in \mathcal{M}_\beta} \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}\big(M(\theta) = 1\big) \leq \alpha,$$

where we define $\mathcal{M}_\beta = \big\{M \in \mathcal{M} : \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_-)}\big(M(\theta) = 1\big) \leq \beta\big\}$.

Similar to the plaintiff's setting, β controls the probability of a false accusation of Access. Taking β smaller restricts attention to highly conservative predictors, reflecting the fact that unreliable accusations should not be considered meaningful. α upper bounds the probability that any such conservative predictor can correctly detect access when it occurred. A small α certifies that the plaintiff's optimal decision algorithm $M \in \mathcal{M}$ cannot reliably establish Access. Guarantees in (α, β)-Access-Evidence are independent of the existence of a particular algorithm $M \in \mathcal{M}$, and consequently provide a *model-level, attack-independent* measure of evidence towards refuting Access.

3.3 Evidence-based Criteria for Similarity

We now turn to Similarity. Whereas Access concerns whether the defendant's training process used data containing protected expression from a work, Similarity concerns whether protected expression from that work is *recoverable* from the trained model. Analogously, we consider a reconstruction game in which the plaintiff attempts to reconstruct copyrighted content through access to the defendant's trained model.

Definition 3.5 (Similarity reconstruction game)**.** The defendant returns a model $\theta \leftarrow \mathsf{Train}(D_+)$. The plaintiff is given the reference dataset D_-, a prior distribution π over copyrighted content $C \in \mathcal{C}$, and access to θ. The plaintiff chooses a reconstruction algorithm $R \in \mathcal{R}$ and outputs a candidate reconstruction $\widehat{C} = R(D_-, \pi, \theta)$.

In this setting, we provide additional information on C in the form of the prior π. In this formulation, the reconstructibility of copyrighted content is determined by Train, the datasets (D_+, D_-), and the prior π. We also suppress the dependence of $R \in \mathcal{R}$ on (D_-, π) and write $R(\theta)$.

Success Criterion in the Similarity **Reconstruction Game.** To evaluate a reconstruction attempt, we require a notion of when an output should be regarded as "similar" to a copyrighted work. Fix a similarity function $d(\cdot, \cdot)$ and a threshold $\eta \geq 0$. For a given $R \in \mathcal{R}$, the plaintiff succeeds whenever $d\big(C,\, R(D_-, \pi, \theta)\big) < \eta$. Previous work on data reconstruction attack also tracks other quantities, such as $\mathbb{E}\left[d(C,\, R(\theta))\right]$ [19]. However, we are particularly interested in the worst-case probability of reconstruction, so we directly control the probability of achieving similarity below a threshold.

Distance functions and protected expression. A distance function is necessary to evaluate the expression-preserving similarity of a reconstructed sample. Given the highly nonlinear nature of protected expression, it is challenging to summarize it via a single distance measure [20]. Thus, d should be selected to reflect a domain-specific notion of similarity most aligned with the modality.

Prior information via π. The task of reconstructing C given D_+ depends on the information available to the plaintiff's algorithm about C. We encode this information through a prior π, in line with previous work on data reconstruction [19, 21]. The concentration of π around C determines the hardness of reconstruction. For example, if $\pi = \delta_C$, then $R \in \mathcal{R}$ can simply return C even without θ. In contrast, if $\pi = \mathsf{Uniform}(\mathcal{W})$, reconstruction is the most challenging, as π provides no additional information regarding C. In practice, π ought to reflect some, but not all, information about C; for example, π may correspond to the set of all images with a certain caption or lexical description.

We now formalize evidence for Similarity from the perspectives of both plaintiff and defendant.

Definition 3.6 ((η, γ)-Similarity-Accusation)**.** A plaintiff's accusation, or equivalently their reconstruction algorithm $R \in \mathcal{R}$, satisfies (η, γ)-Similarity-Accusation with respect to the copyrighted prior π, similarity function $d(\cdot, \cdot)$, and dataset D_+ if

$$\mathbb{P}_{C \sim \pi,\, \theta \leftarrow \mathsf{Train}(D_+)} \Big(d\big(C, R(\theta)\big) \leq \eta \Big) \geq \gamma.$$

If the plaintiff can exhibit a reconstruction algorithm satisfying (η, γ)-Similarity-Accusation with large γ at a stringent threshold η, then the accusation provides concrete evidence that protected expression is reconstructible from the trained model under the chosen evidentiary standard (π, d).

As before, the defendant aims to disprove Similarity. A reasonable guarantee should make reconstruction intrinsically challenging, ensuring a low probability of producing an output similar to the copyrighted work as follows.

Definition 3.7 ((η, γ)-Similarity-Evidence)**.** A generative model θ satisfies (η, γ)-Similarity-Evidence with respect to the dataset D_+, copyrighted prior π, and similarity function $d(\cdot, \cdot)$ if

$$\sup_{R \in \mathcal{R}} \mathbb{P}_{C \sim \pi,\, \theta \leftarrow \mathsf{Train}(D_+)} \Big(d\big(C, R(\theta)\big) \leq \eta \Big) \leq \gamma.$$

3.4 Implications for Auditability and Consistency

These definitions, while motivated by legal doctrine [8, 14], should also satisfy key properties making them useful in copyright litigation. One such property is *auditability*, aligning with the necessity for veracity in the litigation setting [22]. In particular, when provided a guarantee for Access-Evidence or Similarity-Evidence, the plaintiff should be able to empirically confirm these guarantees.

Proposition 3.8 (Auditability)**.** *Suppose that a generative model θ satisfies* (α, β)-Access-Evidence *and* (η, γ)-Similarity-Evidence *with respect to the copyrighted data C, prior π, similarity function $d(\cdot, \cdot)$, and datasets (D_+, D_-). Then, the following hold.*

1. *Any algorithm $M \in \mathcal{M}$ satisfying* $(\widetilde{\alpha}, \widetilde{\beta})$-Access-Accusation *with $\widetilde{\beta} \leq \beta$ must have $\widetilde{\alpha} \leq \alpha$.*

2. *Any algorithm $R \in \mathcal{R}$ satisfying $(\widetilde{\eta}, \widetilde{\gamma})$-*Similarity-Accusation *with $\widetilde{\eta} \leq \eta$ must have $\widetilde{\gamma} \leq \gamma$.*

We provide a proof in Appendix D. In particular, Proposition 3.8 implies that for the same evidentiary standard, the plaintiff is unable to substantively accuse the defendant with confidence that exceeds the restriction on Access inference or Similarity reconstruction guarantees. Consequently, refuting Evidence guarantees merely requires one to construct algorithms satisfying (α, β)-Access-Accusation and (η, γ)-Similarity-Accusation at higher significance levels than the claimed level of protection implies. In contrast, other notions of provable copyright protection [15, 16] lack this auditability property, allowing potentially *malicious* model providers to falsify their claimed levels of protection, lowering the trust in the litigation process.

4 Mechanisms and Implications for Provable Copyright Protection

Having defined our notions of Access and Similarity, we ask three natural questions. Firstly, are these notions achievable? We show that differential privacy is a *more restrictive* notion of copyright safety than our framework, and therefore serves as a conservative sufficient condition for both Access-Evidence and Similarity-Evidence. Secondly, do they afford broader protections outside of litigation? We show that our framework provides guarantees against the generation of copyrighted content, even by *adversarial* users. Third, what do they say about existing copyright safety frameworks? We examine the *Near Access-Freeness* framework proposed by Vyas et al. [15], which provides guarantees in terms of the safety of output distributions rather than internal model content, and show that models satisfying NAF can nonetheless be substantively accused of Access and Similarity.

4.1 Achieving Algorithms and Reconstruction Probability

Given our notions of protection, a model provider may seek to satisfy Access-Evidence and Similarity-Evidence, to prevent substantive accusations of either accessing or embedding protected expression in their model in order to avoid litigation. Previous work has posited that *differential privacy* (DP) can be a sufficient condition for copyright protection [23, 24, 25], and notions of copyright protection hold under differential privacy [16]. We formalize this notion for our framework in Proposition 4.1.

Proposition 4.1 (Privacy implies Evidence)**.** *Suppose $\theta \leftarrow \mathsf{Train}(D_+)$ is (ϵ, δ)-DP with respect to the N-element addition/removal relation. Then, with respect to the dataset D_+ and* any *work $C \in D_+$ with $|W_C| \leq N$ for any $\beta, \eta \in [0, 1]$, we satisfy (α, β)-*Access-Evidence *with $\alpha \geq e^{\epsilon}\beta + \delta$ and (η, γ)-*Similarity-Evidence *with respect to an* arbitrary prior and distance (π, d) *with*

$$\gamma = e^{\epsilon}\kappa(\pi, d) + \delta, \quad \kappa(\pi, d) = \sup_{w \in \mathcal{W}} \mathbb{P}_{C \sim \pi}(d(w, C) \leq \eta).$$

Intuitively, for any non-trivial prior π and distance function, there will be exponentially many distinguishable reconstructions at a certain distance threshold η. Under this assumption, it turns out that the probability of reconstructing a fixed sample is *exponentially* small.

Theorem 4.2 (Privacy Implies $e^{-\Omega(d)}$ Attack Success)**.** *Assume the same setting as Proposition 4.1. Suppose that (π, d) satisfy the entropy condition $H^{\eta}_{rec}(\pi, d) = -\log \kappa(\pi, d) = \Omega(d)$. Then $\gamma = e^{-\Omega(d)} + \delta$, so the probability of reconstruction is exponentially small in d.*

The conditions of Theorem 4.2 are natural, and are satisfied by a wide variety of (π, d) pairs, including Gaussian priors with ℓ_2 distance and shared-prefix priors over bit strings. We prove these results, as well as Proposition 4.1 and Theorem 4.2 in Appendix E.

We emphasize that differential privacy provides stronger protection than either of Access-Evidence or Similarity-Evidence require, in that (i) the protection applies to any datapoint $w \in D$, rather than only to copyrighted training points, and (ii) that the bounds hold uniformly for any $\beta, \eta \in [0, 1]$ as described in Proposition 4.1. It is also known from the literature that ensuring privacy guarantees can sacrifice non-negligible model utility [26, 27, 28]. Consequently, we stress that our guarantees do not reduce to privacy, and believe there exist algorithms better tailored to satisfy either Access-Evidence or Similarity-Evidence, respectively, that may better preserve model utility than DP.

4.2 Limiting the Generation of Copyrighted Works

While our framework focuses on preventing the manifestation of work-specific information in a model, it also provides protection for output distributions. Such protection prevents users from accidentally generating copyrighted content, independent of litigation procedures against the defendant.

To formalize this idea, we view a user as a reconstruction algorithm $R_{\text{user}}(D_-, \pi, p_\theta)$ who only has *query-level* access to p_θ. When (η, γ)-Similarity-Evidence is satisfied, the probability that even an adversarial R_{user} can generate content similar to C is necessarily limited, as seen in Theorem 4.3.

Theorem 4.3 (Limited Risk of Copying)**.** *If a generative model θ satisfies (η, γ)-*Similarity-Evidence *with respect to $\pi, d(\cdot, \cdot), D_+$, then*

$$\mathbb{P}_{C \sim \pi, \theta \leftarrow \mathsf{Train}(D_+)} \left(d(C, R_{\text{user}}(D_-, \pi, p_\theta)) \leq \eta \right) \leq \gamma.$$

We provide a proof in Appendix F. In particular, our guarantees parallel those of Cohen [16], who assume the user is non-adversarial, but also provide the user with more information about C via an auxiliary information construction.

4.3 NAF Models are Vulnerable to Adversaries with Query Access

We demonstrate that the NAF criterion fails to provide appropriate copyright evidence with respect to Access and Similarity. First, we briefly review the framework developed by Vyas et al. [15]. We provide a more complete recollection of the k-NAF framework in Appendix B.

Let $C \in \mathcal{C}$ be a copyrighted work, and consider a function $\mathsf{safe}_C : \mathcal{C} \to \Theta$ mapping copyrighted works to safe models. Abusing notation, $\mathsf{safe}_C(\cdot|x)$ is any model trained without access to C, and x is the prompt. Near Access-Freeness requires that the divergence between the generative distribution $p_\theta(\cdot|x)$ and $\mathsf{safe}_C(\cdot|x)$ be bounded.

Definition 4.4 ([15])**.** We say $p_\theta(\cdot|x)$ is k_x-NAF with respect to copyrighted data C, prompt x, and $\mathsf{safe}_C(\cdot|x)$ if $\Delta(p_\theta(\cdot|x), \mathsf{safe}_C(\cdot|x)) \leq k_x$, where Δ is some divergence. If $k \geq k_x$ for all prompts x, then $p_\theta(\cdot|x)$ is said to be k-NAF with respect to Δ.

In Vyas et al. [15] and our subsequent discussion, we focus on the case where the divergence is the Rényi divergence of order infinity (i.e., $\Delta = \max_{y \in \mathsf{Supp}(\rho)} \log(\rho(y)/\mu(y))$ for two distributions ρ, μ). Consequently, each query to $p_\theta(\cdot|x)$ can leak up to k_x bits of information. Formally, they show that if $p_\theta(\cdot|x)$ is k_x-NAF on prompt x with respect to $\mathcal{C}$ and safe, then $p_\theta(\cdot|x) \leq 2^{k_x} \cdot \mathsf{safe}_C(\cdot|x)$.

If both k_x and $\mathsf{safe}_C(\cdot|x)$ are small on any events that generate works within $\mathsf{PE}(C)$, this guarantee tightly controls the probability of generating a sample with the protected expression. Nevertheless, when $k_x > 0$, each query can still leak k_x bits of information about the training data. This observation has been used to show k-NAF models are vulnerable to untargeted training data extraction [16]. We generalize these examples to construct adversaries that can efficiently infer access to and reconstruct *specific instances* of copyrighted data, culminating in the following theorem.

Theorem 4.5 (Substantive Accusations against k-NAF)**.** *Fix any constant $k > 0$ and* arbitrary *parameters $\beta, \delta \in (0, 1)$. There exist a copyrighted work $C^\star$, datasets D_- and $D_+ = D_- \cup \{C^\star\}$, and algorithm* Train *inducing a generative distribution p_θ, such that p_θ is k-*NAF *and the following hold.*

1. *There exists a decision algorithm $M \in \mathcal{M}$ satisfying $(1 - \delta, \beta)$-*Access-Accusation.
2. *There exists a reconstruction algorithm $R \in \mathcal{R}$ satisfying $(0, 1 - \delta)$-*Similarity-Accusation.

We provide a proof in Appendix C. Our theorem demonstrates that the k-NAF criterion cannot protect against substantive accusations of Access and Similarity, based on the existence of arbitrarily strong attackers and Proposition 3.8. In contrast, our proposed notions of Access-Evidence and Similarity-Evidence in Definitions 3.4 and 3.7 have precise operational meaning, and are grounded in a formal adversarial formulation between plaintiff and defendant.

5 Experiments and Analysis

We evaluate our framework on both language and image generation tasks under the finetuning setting, with the goal of simulating how the defendant and plaintiff may show or refute Access and Similarity in practice. We provide additional experimental details in Appendix G.

Datasets and models. First, we fine-tune and evaluate Llama2-7B [29] on abstracts from math papers (MathAbstracts) [30] and writing prompts (WritingPrompts) [31] that have previously been used for copyright evaluation [32]. Secondly, we fine-tune and evaluate Stable Diffusion v1.4 [3] on Pokemon caption-image datasets (Pokémon) [33] and a subset of images from LAION (LAION-MI) [34].

Membership inference attacks. To test Access-Evidence directly, we evaluate reference-free membership inference attacks that are strong for each modality and have demonstrated state-of-the-art performance. For diffusion and language models, we use the Proximal Initialization Attack (PIA) and MinK%++ respectively [35, 36]. We measure the performance of MIA by the TPR at low FPR.

Data reconstruction attacks. In order to test Similarity-Evidence, we evaluate data reconstruction attacks for both modalities. For diffusion and language modeling, we adopt a reconstruction attack from Carlini et al. [37], and Nasr et al. [38] respectively. In addition to the reconstruction algorithm, evaluating Similarity-Evidence requires a distance function. For images, we use CLIP similarity to capture semantic alignment [39], and DreamSim for perceptual similarity [40]. Similarly, for text, we report ROUGE-L scores [41] and BERTScore [42] to capture token overlap and semantic similarity respectively. We scale all metrics into $[0, 1]$ such that decreasing values indicate increasing similarity.

Utility Metrics. We use standard metrics to evaluate finetuned model utility. For diffusion models, we adopt KID, CLIPScore, and CLIP-IQA [43, 39, 44]. For language models, we adopt perplexity under the external model Mistral-7B ($\mathrm{PPL}_{\mathrm{ext}}$), and numerical LLM-as-a-judge fluency (FLU) evaluation with PrometheusV2 [45, 46, 47].

Copyright protection strategies. We evaluate algorithms proposed to satisfy the k-NAF guarantees proposed by [15]. We focus on the CP-k algorithm, which relies on a rejection sampling approach to satisfy the k-NAF guarantee. Given a data partition $D_+ = D_1 \cup D_2$, CP-k takes three models: a draft model p, trained on D_+, and q_1, q_2 trained on D_1, D_2 respectively. Then, for each generated output $y \sim p$ with prompt z, the CP-k algorithm will release sample y only if the maximum log-likelihood ratio $\max_{i=\{1,2\}} \log(p(y|z)/q_i(y|z)) \leq k$. Setting a lower k results in a smaller acceptance probability α_k and a better NAF guarantee for moderate α_k [15]. In our experiments, we study different acceptance probabilities $\alpha_k \in [0, 1]$ and examine how the CP-k algorithm affects the performance of MIA and DRA. Notably, a smaller α_k also implies higher computational complexity, as we need to sample $1/\alpha_k$ more times in expectation for an accepted output.

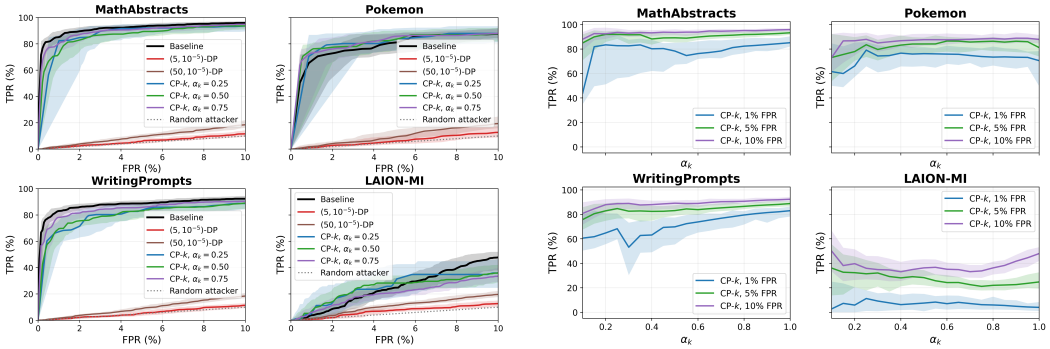
As described in Section 4.1, we train (ϵ, δ)-DP models with DP-Adam to satisfy both (α, β)-Access-Evidence and (η, γ)-Similarity-Evidence respectively. We study $\epsilon \in \{5, 10, 20, 50\}$ and $\delta = 10^{-5}$.

We discuss experiments involving additional distance metrics and utility metrics in Appendix H.

5.1 Results

Evaluating the Access Inference Game. We begin by examining the Access inference game by instantiating decision algorithms against models trained on copyrighted data. We present Fig. 2a, which provides Access evaluation metrics through TPR at a low FPR over multiple models and data modalities for each tested dataset. Models satisfying (α, β)-Access-Evidence and (η, γ)-Similarity-Evidence reduce the ability of the plaintiff to meaningfully accuse a model provider of Access: such a reduction holds between privacy parameters $\epsilon \in \{5, 50\}$. This is reflected by a lower TPR at fixed FPR across data modalities and datasets, compared to baselines, and underscores our framework's capability in avoiding litigation and reducing intrinsic model impact.

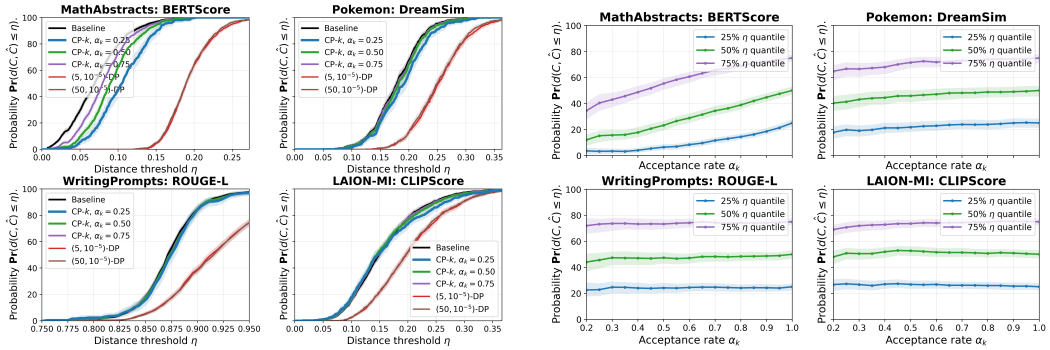
Evaluating the Similarity Reconstruction Game. We now examine our second setting by studying reconstruction algorithms against models trained on copyrighted data. We summarize results in Fig. 3a. Models trained to satisfy (η, γ)-Similarity-Evidence via differential privacy substantially reduce reconstruction success across modalities and datasets, indicating that the plaintiff's ability to produce a meaningful accusation of Similarity is weakened under our evidentiary criterion. This supports our claim that the appropriate notion of copying is model-centric, and is governed by whether training



(a) TPR at fixed, low FPR.

(b) TPR versus CP-k acceptance rate α_k.

Figure 2: Models satisfying (α, β)-Access-Evidence substantially reduce TPR at fixed, low FPR across text and image datasets, whereas CP-k offers only marginal protection. Furthermore, more restrictive thresholds of α_k do not meaningfully impact the protection offered by CP-k.



(a) Similarity reconstruction success.

(b) Similarity attack success at fixed thresholds.

Figure 3: Estimated probability that an attack reconstructs content within threshold η of the copyrighted target across text and image datasets. Models trained to satisfy (η, γ)-Similarity-Evidence via DP exhibit markedly lower success probabilities compared to models satisfying k-NAF guarantees.

induces recoverable, work-specific information in the model, which (η, γ)-Similarity-Evidence allows us to directly control.

The Role of α_k. In contrast to the discussion above, models trained with k-NAF through the CP-k algorithm do not substantially reduce the confidence level of accusations for either Access or Similarity. In fact, for CP-k, we find that lowering the acceptance rate α_k (i.e. providing a more restrictive k-threshold) does *not* substantially reduce the success of membership inference or data reconstruction, as seen in Figs. 2b and 3b. These results hold for FPR $\in \{1\%, 5\%, 10\%\}$, suggesting that NAF is unable to provide meaningful guarantees for Access-Evidence or Similarity-Evidence.

Model Utility. For the settings we consider, training with (ϵ, δ)-DP may result in somewhat lower utility compared to baselines and mechanisms satisfying k-NAF, as in Table 1. Nevertheless, DP-trained models maintain reasonable generation quality *while* substantially reducing Access and Similarity attack success. Thus, in our experiments, DP trades a limited amount of utility for a considerably stronger form of evidence-based protection than output-based protections, such as k-NAF.

Taking these findings together, we believe that mechanisms that control output distributions can still permit training-induced, work-specific information to persist in model parameters, which does not prevent substantive accusations of Access or Similarity.

Conclusion. We introduced the Access-Similarity framework, an operational and auditable approach to quantifying copyright evidence for generative models. By formalizing Access and Similarity as adversarial games between plaintiff and defendant, our framework provides quantitative, auditable

Table 1: Utility evaluation across language and image benchmarks. Across modalities, DP-trained models have somewhat lower, but still competitive, utility relative to CP-k while providing stronger evidence-based protection in our Access and Similarity games.

	MathAbstracts		**WritingPrompts**		**Pokémon**			**LAION-MI**		
Model	$\text{PPL}_{\text{ext}} \downarrow$	FLU $\uparrow$	$\text{PPL}_{\text{ext}} \downarrow$	FLU $\uparrow$	$10^4 \cdot \text{KID} \downarrow$	CLIPScore $\uparrow$	CLIP-IQA $\uparrow$	$10^4 \cdot \text{KID} \downarrow$	CLIPScore $\uparrow$	CLIP-IQA $\uparrow$
Baseline	6.29	3.07	7.06	3.86	6.32	0.337	0.812	0.769	0.356	0.653
CP-k, $\alpha_k = 25\%$	6.33	3.07	6.81	4.24	7.30	0.320	0.799	1.80	0.339	0.649
DP, $\varepsilon = 50$	2.49	3.00	4.90	3.87	6.63	0.347	0.756	0.832	0.328	0.599
DP, $\varepsilon = 5$	2.46	3.01	4.93	3.58	7.02	0.330	0.729	0.90	0.301	0.589

guarantees grounded in U.S. copyright doctrine. The usability of our framework is validated by experiments on realistic datasets and multiple modalities. Our framework provides a principled foundation for copyright adjudication as litigation for generative models continues to develop.

References

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[7] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[8] Katherine Lee, A Feder Cooper, and James Grimmelmann. Talkin"bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023.

[9] Giorgio Franceschelli and Mirco Musolesi. Copyright in generative deep learning. *Data & Policy*, 4:e17, 2022.

[10] Pamela Samuelson. Generative ai meets copyright. *Science*, 381(6654):158–161, 2023.

[11] Christopher T. Zirpoli. Generative artificial intelligence and copyright law. Legal Sidebar LSB10922, Congressional Research Service, February 2023.

[12] Masahiro Kaneko and Timothy Baldwin. Investigating how pre-training data leakage affects models' reproduction and detection capabilities. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23545–23555, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1201. URL `https://aclanthology.org/2025.emnlp-main.1201/`.

[13] Niva Elkin-Koren, Uri Hacohen, Roi Livni, and Shay Moran. Can copyright be reduced to privacy? In *5th Symposium on Foundations of Responsible Computing (FORC 2024)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2024.

[14] A. Feder Cooper and James Grimmelmann. The files are in the computer: On copyright, memorization, and generative ai, 2025. URL https://arxiv.org/abs/2404.12590.

[15] Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pages 35277–35299. PMLR, 2023.

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

[17] U.S. Copyright Office. Copyright law of the united states and related laws contained in title 17 of the united states code. Technical report, U.S. Copyright Office, December 2025. URL https://www.copyright.gov/title17/title17.pdf. PDF compilation; includes amendments enacted through December 18, 2025.

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

[19] Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, pages 8056–8071. PMLR, 2022.

[20] Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing human ingenuity: A quantitative framework for copyright law's substantial similarity. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pages 37–49, 2022.

[21] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.

[22] Yanming Li, Seifeddine Ghozzi, Cédric Eichler, Nicolas Anciaux, Alexandra Bensamoun, and Lorena Gonzalez Manzano. Data provenance auditing of fine-tuned large language models with a text-preserving technique, 2025. URL https://arxiv.org/abs/2510.09655.

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

[24] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use, 2023. URL https://arxiv.org/abs/2303.15715.

[25] Roi Livni, Shay Moran, Kobbi Nissim, and Chirag Pabbaraju. Credit attribution and stable compression, 2024. URL https://arxiv.org/abs/2406.15916.

[26] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[27] Eli Chien, Wei-Ning Chen, Chao Pan, Pan Li, Ayfer Ozgur, and Olgica Milenkovic. Differentially private decoupled graph convolutions for multigranular topology protection. *Advances in Neural Information Processing Systems*, 36:45381–45401, 2023.

[28] Eli Chien, Yuzheng Hu, Ryan McKenna, Shanshan Wu, Zheng Xu, and Peter Kairouz. Maple: Metadata augmented private language evolution. *arXiv preprint arXiv:2603.19258*, 2026.

[29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian

Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[30] Yifan Zhang, Yifan Luo, Yang Yuan, and Andrew C Yao. Autonomous data selection with zero-shot generative classifiers for mathematical texts, 2025. URL https://arxiv.org/abs/2402.07625.

[31] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation, 2018. URL https://arxiv.org/abs/1805.04833.

[32] Javier Abad, Konstantin Donhauser, Francesco Pinto, and Fanny Yang. Copyright-protected language generation via adaptive model fusion. *arXiv preprint arXiv:2412.06619*, 2024.

[33] diffusers. pokemon-gpt4-captions. https://huggingface.co/datasets/diffusers/pokemon-gpt4-captions, 2024. Hugging Face dataset.

[34] Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemysław Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models, 2023. URL https://arxiv.org/abs/2306.12983.

[35] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization, 2023. URL https://arxiv.org/abs/2305.18355.

[36] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models, 2025. URL https://arxiv.org/abs/2404.02936.

[37] 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, 2023. URL https://arxiv.org/abs/2301.13188.

[38] Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations (ICLR) 2025*, 2025. URL https://openreview.net/forum?id=vjel3nWP2a. Poster.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[40] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in neural information processing systems*, 2023.

[41] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *The Eighth International Conference on Learning Representations*, 2020.

[43] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Demystifying MMD GANs. In *Proceedings of the Fifth International Conference on Learning Representations*, 2017.

[44] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.

[45] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, 2021.

[46] Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

[47] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, 2024.

[48] Hiroaki Chiba-Okabe and Weijie J Su. Tackling copyright issues in ai image generation through originality estimation and genericization. *Scientific Reports*, 15(1):10621, 2025.

[49] Timothy Chu, Zhao Song, and Chiwun Yang. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17871–17879, 2024.

[50] Jiachen T Wang, Zhun Deng, Hiroaki Chiba-Okabe, Boaz Barak, and Weijie J Su. An economic solution to copyright challenges of generative ai. *arXiv preprint arXiv:2404.13964*, 2024.

[51] Junwei Deng and Jiaqi Ma. Computational copyright: Towards a royalty model for AI music generation platforms. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024. URL https://openreview.net/forum?id=CIQxqQvkEE.

[52] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

[53] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

[54] Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

[55] Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1813–1830, 2024.

[56] Michael-Andrei Panaitescu-Liess, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. Can watermarking large language models prevent copyrighted text generation and hide training data? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25002–25009, 2025.

[57] Xiang Li, Qianli Shen, and Kenji Kawaguchi. Va3: Virtually assured amplification attack on probabilistic copyright protection for text-to-image generative models, 2024. URL https://arxiv.org/abs/2312.00057.

[58] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

[59] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd international conference on artificial intelligence and statistics*, pages 1226–1235. PMLR, 2019.

[60] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[61] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=Q42f0dfjECO`.

[62] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[64] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[65] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

A Additional Related Works

Blameless Copy Protection [16]. We see this work as an important orthogonal perspective to the perspective we take in Section 1. In particular, Cohen [16]'s framework rests on the principle that safe models must avoid enabling *blameless users*, those who do not themselves induce infringement, from inadvertently reproducing copyrighted content. To this end, Cohen introduces *clean-room copy protection*, under which a training algorithm is (κ, β)-clean if any user whose probability of copying in a counterfactual clean-room environment is at most β faces at most κ probability of copying in reality. Like our work, Cohen proves that NAF fails to satisfy this criterion (tainted models can enable verbatim reproduction of training data), and that differential privacy provides a sufficient condition for the proposed guarantee. While certainly related to our general goal of studying copyright infringement, our framework is distinct from this setting. Whereas Cohen's framework is *output-centric*: asking whether a user's generated output infringes, ours is *model-centric*, asking whether protected expression is intrinsically encoded in the model's weights and recoverable by an adversarial plaintiff. Consequently, the two frameworks thus address different aspects of the copyright problem. Nevertheless, we acknowledge the importance of the perspective provided by their work.

Quantifying Evidential Notions for Copyright. Previous work has attempted to formalize notions of evidence in copyright. Scheffler et al. [20] addresses a distinct problem: determining "substantial similarity" in legal contexts. The authors propose a complexity-theoretic similarity test based on the description length required to derive one specific work from another. Their framework is designed to measure the similarity between two specific samples, regardless of whether they are AI-generated. In contrast, we directly regulate the ability of the generative model to generate similar output before its actual generation process, so our work is orthogonal to their approach. Chiba-Okabe and Su [48] proposes a distance-based originality measure and the corresponding generalization for reducing the risk of copyright infringement. Their work can be viewed as another attempt that tries to provide evidence regarding Similarity, but it remains meaningfully different from our proposal of leveraging the Similarity reconstruction game, and does not attempt to address Access of the copyrighted data. Chu et al. [49] proposes measuring the degree of copyright infringement by comparing the average loss on copyrighted versus non-copyrighted data in the training set, and aims to mitigate the risk of generating copyrighted content by increasing this loss gap during training. While potentially beneficial, the loss gap is a heuristic and generally does not offer a rigorous guarantee of copyright protection. Besides trying to provide theoretical measures related to AI copyright infringement, there are works that instead focus on designing platforms for distributing revenues to copyrighted content holders based on Shapely values [50] or other data attribution techniques [51], which are quite interesting but remain orthogonal to our work.

Watermarking approaches. Another line of research focuses on watermarking generative models, that is, injecting detectable signals into generated samples to enable identification of whether a sample originates from a specific model [52, 53, 54, 55]. Although watermarking was not originally designed to address copyright concerns, recent empirical studies have shown that watermarking language models can reduce the generation of copyrighted content and mitigate membership inference attacks on copyrighted training data [56]. Nevertheless, watermarking alone does not provide a formal framework for measuring copyright infringement, nor does it offer a rigorous guarantee of copyright protection.

B Previous Notions of Copyright Protection

In this section, we discuss technical details related to previous works on provable copyright protection. In particular, we focus on the NAF framework's theoretical basis, algorithms, and guarantees [15].

B.1 Near Access-Freeness

We independently provide the same exposition discussed in the main paper for completeness. Recall that $\mathcal{W}$ denotes the set of all works and $\mathcal{C} \subseteq \mathcal{W}$ denotes the set of copyrighted works. Fix a copyrighted work $C \in \mathcal{C}$, and let $\mathsf{PE}(C)$ denote the class of works sharing protected expression with C. For a training dataset D_+, recall that

$$W_C = D_+ \cap \mathsf{PE}(C), \qquad D_- = D_+ \setminus W_C.$$

Thus, D_- is the counterfactual training dataset obtained by removing all training works that share protected expression with C.

Let $\mathsf{Train} : 2^{\mathcal{W}} \to \Theta$ denote a training algorithm and let $p_\theta(\cdot \mid z)$ denote the conditional output distribution of a model $\theta \in \Theta$ on prompt z. A *safe map* associates to each copyrighted work $C \in \mathcal{C}$ a conditional distribution $\mathsf{safe}_C(\cdot \mid z)$ induced by a model trained without access to works in W_C. For example, we can take

$$\theta_C^- \leftarrow \mathsf{Train}(D_-), \qquad \mathsf{safe}_C(\cdot \mid z) = p_{\theta_C^-}(\cdot \mid z).$$

We focus on the guarantees proved by Vyas et al. [15] for the max-divergence,

$$\Delta_{\max}(p\|q) = \sup_y \log \frac{p(y)}{q(y)},$$

but note that analogous statements can be made for other divergences, such as Δ_{KL}. We say that p_θ is k_z-NAF on prompt z, with respect to copyrighted work C and safe distribution safe_C, if

$$\Delta_{\max}(p_\theta(\cdot \mid z) \,\|\, \mathsf{safe}_C(\cdot \mid z)) \le k_z.$$

If there exists $k \ge k_z$ for all prompts z, then we say that p_θ is k-NAF. This definition yields an operational event-level guarantee. For every copyrighted work $C \in \mathcal{C}$ and every event E in the output space,

$$p_\theta(E \mid z) \le 2^{k_z}\, \mathsf{safe}_C(E \mid z). \tag{1}$$

Indeed, $\Delta_{\max}(p\|q) \le k$ implies $p(y) \le 2^k q(y)$ pointwise, and summing over $y \in E$ gives Eq. (1).

We next discuss a rejection-sampling mechanism, CP-k, proposed by [15] to actually instantiate a model satisfying a k-NAF guarantee. We summarize this procedure in Algorithm 1. Let $\theta \leftarrow \mathsf{Train}(D_+)$ be a draft model trained on the full dataset, and let $p_\theta(\cdot \mid z)$ denote its output distribution. Let $D_+ = D_1 \sqcup D_2$ be a partition of the data, and for $i \in \{1, 2\}$ let

$$\theta_i \leftarrow \mathsf{Train}(D_i), \qquad q_i(\cdot \mid z) = p_{\theta_i}(\cdot \mid z),$$

For a fixed copyrighted work C, the sharding construction is interpreted as safe when at least one shard excludes all works in W_C; that is, when $D_i \cap W_C = \emptyset$ for some $i \in \{1, 2\}$. In that case, the corresponding shard model q_i is trained without access to protected-expression-sharing works for C and may serve as safe_C.

The CP-k sampler draws from the draft model and accepts only samples whose likelihood ratio against each shard model is sufficiently small. With threshold k_0, at each iteration one draws $y \sim p_\theta(\cdot \mid z)$ and accepts it if

$$\max_{i \in \{1,2\}} \log \frac{p_\theta(y \mid z)}{q_i(y \mid z)} \le k_0.$$

Now, let

$$\alpha_k(z) = \mathbb{P}_{y \sim p_\theta(\cdot|z)}\left[\max_{i \in \{1,2\}} \log \frac{p_\theta(y \mid z)}{q_i(y \mid z)} \le k\right]$$

denote the single-shot acceptance probability, which is nondecreasing in k. Let $p_{\theta,k}(\cdot \mid z)$ denote the distribution of the accepted sample. A result of [15] implies that $p_{\theta,k}$ satisfies a NAF guarantee with

$$k_z = k_0 + \log\left(\frac{1}{\alpha_{k_0}(z)}\right).$$

Algorithm 1 CP-k sampling procedure [15]

Input: Dataset D_+, training algorithm Train, prompt z, threshold k_0, shard partition $D_+ = D_1 \sqcup D_2$.
Train draft model $\theta \leftarrow \mathsf{Train}(D_+)$ and set $p_\theta(\cdot \mid z)$
Train shard models $\theta_i \leftarrow \mathsf{Train}(D_i)$ and set $q_i(\cdot \mid z) = p_{\theta_i}(\cdot \mid z)$ for $i \in \{1, 2\}$
repeat
  if $\max_{i\in\{1,2\}} \log \frac{p_\theta(y|z)}{q_i(y|z)} \leq k_0$ **then**
    break
  end
until *Sample* $y \sim p_\theta(\cdot \mid z)$;
return *Sample* y

Equivalently, increasing the rejection threshold increases the acceptance probability and reduces the additive rejection-sampling slack $\log(1/\alpha_{k_0}(z))$, while also increasing the direct likelihood-ratio term k_0. In practice, we target a desired acceptance rate α_k and choose k by estimating the corresponding empirical quantile of

$$\max_{i\in\{1,2\}} \log \frac{p_\theta(y \mid z)}{q_i(y \mid z)} \qquad \text{for } y \sim p_\theta(\cdot \mid z),$$

an approach taken by previous work [57].

C Counterexamples against the *Near Access-Freeness* Framework

We restate Theorem 4.5, and note that the algorithms employed in this theorem are polynomial in the input parameters.

Theorem C.1 (Substantive Accusations against k-NAF)**.** *Fix any constant $k > 0$ and* arbitrary *parameters $\beta, , \delta \in (0, 1)$. There exist a copyrighted work $C^\star$, datasets D_- and $D_+ = D_- \cup \{C^\star\}$, and algorithm* Train *inducing a generative distribution p_θ, such that p_θ is k-*NAF *and the following hold.*

1. *There exists a decision algorithm $M \in \mathcal{M}$ satisfying $(1 - \delta, \beta)$-*Access-Accusation*.*
2. *There exists a reconstruction algorithm $R \in \mathcal{R}$ satisfying $(0, 1 - \delta)$-*Similarity-Accusation*.*

Both M and R require polynomially many queries to p_θ.

The construction we use adapts portions of the bit-leakage construction of Cohen [16].

Proof. Fix arbitrary parameters $k > 0$, $\delta, \beta \in (0, 1)$. We construct a training algorithm whose output distributions satisfy k-NAF, but for which there are black-box algorithms satisfying $(1-\delta, \beta)$-Access-Accusation and $(0, 1 - \delta)$-Similarity-Accusation.

Construction. Let $m \in \mathbb{N}$. Let the work domain be $\mathcal{W} = \{0, 1\}^m$. We model the auxiliary information available to the plaintiff as a public bit string $a \in \{0, 1\}^\ell$, for some $\ell \in \mathbb{N}$. This string indexes a nonempty subclass $\mathcal{C}_a \subseteq \mathcal{W}$ of candidate works consistent with the auxiliary information. Equivalently, a should be interpreted as a finite description of the side information that narrows the plaintiff's prior to $\mathcal{C}_a$, but does not by itself identify the protected work.

Fix a copyrighted work $C^\star \in \mathcal{C}_a$, and let D_- be any dataset satisfying $D_- \cap \mathcal{C}_a = \emptyset$. Set

$$D_+ = D_- \cup \{C^\star\}.$$

The prompt space contains triples $x = (a, s, i)$, where $s \in \mathbb{N}$ and $i \in [m]$, and the output alphabet is $\mathcal{Y} = \{0, 1\}$. Let $\rho = \min\{1, 2^k - 1\}$, and define a safe model by

$$\mathsf{safe}_{C^\star}(0 \mid x) = \mathsf{safe}_{C^\star}(1 \mid x) = \frac{1}{2}$$

for every prompt x. Clearly, this model is not biased towards generating any particular sample. Consequently, under the safe model, generating any m-bit work has probability 2^{-m}, satisfying the desideratum for the safe model to have low probability of generating copyrighted content [15]. We now define Train. Given a dataset D, if $D \cap \mathcal{C}_a = \emptyset$, then the returned model p_D satisfies

$$p_D(\cdot \mid x) = \mathrm{Unif}(\mathcal{Y})$$

for every prompt x. If $D \cap \mathcal{C}_a \neq \emptyset$, let C_D denote the first element of $D \cap \mathcal{C}_a$ under a fixed canonical ordering, and write $C_D = (C_{D,1}, \ldots, C_{D,m})$. For prompts $x = (a, s, i)$, define

$$p_D(1 \mid a, s, i) = \frac{1}{2} + \rho \left(C_{D,i} - \frac{1}{2} \right),$$

and set $p_D(\cdot \mid x) = \mathrm{Unif}(\mathcal{Y})$ for all remaining prompts x. We verify that every model returned by this training algorithm satisfies k-NAF with respect to the safe model. If $p_D(\cdot \mid x)$ is uniform, then

$$\max_{y \in \{0,1\}} \frac{p_D(y \mid x)}{\mathsf{safe}_{C^\star}(y \mid x)} = 1.$$

Otherwise, for some coordinate i,

$$\max_{y \in \{0,1\}} \frac{p_D(y \mid x)}{\mathsf{safe}_{C^\star}(y \mid x)} = 1 + \rho \leq 2^k.$$

Hence, for every prompt x, $\Delta_{\max}(p_D(\cdot \mid x) \,\|\, \mathsf{safe}_{C^\star}(\cdot \mid x)) \leq k$, so the model is k-NAF.

Access algorithm. We next construct an access algorithm M. Let $r = \left\lceil \frac{\log_2(1/\beta)}{m} \right\rceil$, and let n be any odd integer satisfying

$$n \geq \frac{2}{\rho^2} \log \frac{rm}{\delta}.$$

For each $s \in [r]$ and $i \in [m]$, the algorithm queries the prompt (a, s, i) exactly n times and decodes the coordinate by majority vote, obtaining $\widehat{C}_{s,i}$. Let $\widehat{C}_s = (\widehat{C}_{s,1}, \ldots, \widehat{C}_{s,m})$. Then, let M outputs $M(D_-, C^\star, \theta) = 1$ if and only if $\widehat{C}_1 = \cdots = \widehat{C}_r = C^\star$.

Under $\theta \leftarrow \mathsf{Train}(D_+)$, the model encodes $C^\star$. For each pair (s, i), the queried Bernoulli random variables have mean $1/2 + \rho/2$ if $C^\star_i = 1$, and mean $1/2 - \rho/2$ if $C^\star_i = 0$. Therefore, Hoeffding's inequality gives

$$\Pr[\widehat{C}_{s,i} \neq C^\star_i] \leq \exp\left(-\frac{n\rho^2}{2}\right).$$

By a union bound over all rm decoded coordinates,

$$\Pr_{\theta \leftarrow \mathsf{Train}(D_+)}[M(D_-, C^\star, \theta) = 1] \geq 1 - rm \exp\left(-\frac{n\rho^2}{2}\right) \geq 1 - \delta.$$

Under $\theta \leftarrow \mathsf{Train}(D_-)$, every query response is an independent fair bit. Since n is odd, each majority vote is also a fair bit, and the vectors $\widehat{C}_1, \ldots, \widehat{C}_r$ are independent and uniform on $\{0,1\}^m$. Hence

$$\Pr_{\theta \leftarrow \mathsf{Train}(D_-)}[M(D_-, C^\star, \theta) = 1] = 2^{-mr} \leq \beta.$$

Thus M satisfies $(1 - \delta, \beta)$-Access-Accusation.

Reconstruction algorithm. We now construct a reconstruction algorithm R. Let n_R be any odd integer satisfying

$$n_R \geq \frac{2}{\rho^2} \log \frac{m}{\delta}.$$

Given D_-, any prior π supported on $\mathcal{C}_a$, and black-box access to θ, the algorithm R queries $(a, 1, i)$ exactly n_R times for each $i \in [m]$, decodes each coordinate by majority vote, and outputs the resulting string $\widehat{C} \in \{0,1\}^m$. Fix any $C \in \mathcal{C}_a$. If $\theta \leftarrow \mathsf{Train}(D_- \cup \{C\})$, then $(D_- \cup \{C\}) \cap \mathcal{C}_a = \{C\}$, so the model encodes C. Applying the same Hoeffding and union-bound argument over the m coordinates gives

$$\Pr_{\theta \leftarrow \mathsf{Train}(D_- \cup \{C\})}[R(D_-, \pi, \theta) = C] \geq 1 - m \exp\left(-\frac{n_R\rho^2}{2}\right) \geq 1 - \delta.$$

Since this bound holds pointwise for every $C \in \mathcal{C}_a$, averaging over any prior π supported on $\mathcal{C}_a$ yields

$$\Pr_{C \sim \pi,\ \theta \leftarrow \mathsf{Train}(D_- \cup \{C\})}[R(D_-, \pi, \theta) = C] \geq 1 - \delta.$$

Consequently, for any distance d satisfying $d(C, C) = 0$,

$$\Pr_{C \sim \pi,\ \theta \leftarrow \mathsf{Train}(D_- \cup \{C\})}\big[d\big(C, R(D_-, \pi, \theta)\big) \leq 0\big] \geq 1 - \delta.$$

Thus R satisfies $(0, 1 - \delta)$-Similarity-Accusation.

Finally, the query complexities are rmn for M and mn_R for R. Since $\rho = \min\{1, 2^k - 1\}$ and $2^k - 1 = \Omega(k)$ for $0 < k \leq 1$, both query complexities are polynomial in the input parameters, which completes the construction. $\square$

D Proof of Auditability Property in Proposition 3.8

We repeat Proposition 3.8 below for clarity.

Proposition D.1 (Auditability)**.** *Suppose that a generative model θ satisfies (α, β)-*Access-Evidence *and (η, γ)-*Similarity-Evidence *with respect to the copyrighted data C, prior π, similarity function $d(\cdot, \cdot)$, and datasets (D_+, D_-). Then, the following hold.*

1. *Any algorithm $M \in \mathcal{M}$ satisfying $(\widetilde{\alpha}, \widetilde{\beta})$-*Access-Accusation *with $\widetilde{\beta} \leq \beta$ must have $\widetilde{\alpha} \leq \alpha$.*
2. *Any algorithm $R \in \mathcal{R}$ satisfying $(\widetilde{\eta}, \widetilde{\gamma})$-*Similarity-Accusation *with $\widetilde{\eta} \leq \eta$ must have $\widetilde{\gamma} \leq \gamma$.*

Proof. Recall the definition of (α, β)-Access-Evidence with respect to C, copyrighted set W_C, and datasets (D_+, D_-). We have

$$\sup_{M \in \mathcal{M}_\beta} \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(M(\theta) = 1) \leq \alpha.$$

where $\mathcal{M}_\beta = \left\{M \in \mathcal{M} : \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_-)}\big(M(\theta) = 1\big) \leq \beta\right\}$. Suppose that any particular algorithm $M^* \in \mathcal{M}$ satisfies $(\widetilde{\alpha}, \widetilde{\beta})$-Access-Accusation with respect to the same parameters with $\widetilde{\beta} \leq \beta$. Then $M^* \in \mathcal{M}_\beta$, so it follows that

$$\widetilde{\alpha} = \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(M^*(\theta) = 1) \leq \sup_{M \in \mathcal{M}_\beta} \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(M(\theta) = 1) \leq \alpha,$$

so $\widetilde{\alpha} \leq \alpha$, as desired. Similarly, if θ satisfies (η, γ)-Similarity-Evidence with respect to prior π, copyrighted set W_C, datasets (D_+, D_-), and distance function $d(\cdot, \cdot)$. Then

$$\sup_{R \in \mathcal{R}} \mathbb{P}_{C \sim \pi, \theta \leftarrow \mathsf{Train}(D_+)}(d(C, R(\theta) \leq \eta) \leq \gamma.$$

By essentially the same argument, if any particular algorithm $R^* \in \mathcal{R}$ satisfies $(\widetilde{\eta}, \widetilde{\gamma})$-Similarity-Accusation with respect to the same parameters, we have

$$\widetilde{\gamma} \leq \mathbb{P}_{C \sim \pi, \theta \leftarrow \mathsf{Train}(D_+)}(d(C, R^*(\theta) \leq \widetilde{\eta}) \leq \mathbb{P}_{C \sim \pi, \theta \leftarrow \mathsf{Train}(D_+)}(d(C, R^*(\theta) \leq \eta),$$

by the definition of $(\widetilde{\eta}, \widetilde{\gamma})$-Similarity-Accusation and the observation that $\{d(C, R^*(\theta) \leq \widetilde{\eta}\} \subseteq \{d(C, R^*(\theta) \leq \eta\}$, which holds since $\widetilde{\eta} \leq \eta$. Since $R^* \in \mathcal{R}$, it follows directly that

$$\widetilde{\gamma} \leq \mathbb{P}_{C \sim \pi, \theta \leftarrow \mathsf{Train}(D_+)}(d(C, R^*(\theta) \leq \eta) \leq \sup_{R \in \mathcal{R}} \mathbb{P}_{C \sim \pi, \theta \leftarrow \mathsf{Train}(D_+)}(d(C, R(\theta) \leq \eta) \leq \gamma,$$

so $\widetilde{\gamma} \leq \gamma$, what we wanted to show.

□

E Proof of Achievability, Exponentially Small Reconstruction

E.1 Proof of Achievability

We restate Proposition 4.1.

Proposition E.1 (Privacy implies Evidence)**.** *Suppose $\theta \leftarrow \mathsf{Train}(D_+)$ is (ϵ, δ)-DP with respect to the N-element addition/removal relation. Then, with respect to the dataset D_+ and* any *work $C \in D_+$ with $|W_C| \leq N$ for any $\beta, \eta \in [0,1]$, we satisfy (α, β)-*Access-Evidence *with $\alpha \geq e^\epsilon \beta + \delta$ and (η, γ)-*Similarity-Evidence *with respect to an* arbitrary prior and distance (π, d) *with*

$$\gamma = e^\epsilon \kappa(\pi, d) + \delta, \quad \kappa(\pi, d) = \sup_{w \in \mathcal{W}} \mathbb{P}_{C \sim \pi}(d(w, C) \leq \eta).$$

Proof. Before proceeding, we fix $C \in D$ arbitrarily. We consider W_C as defined in the main text, and denote $D_+ = D$ and $D_- = D_+ \setminus W_C$. We begin by showing the (α, β)-Access-Evidence bound. By the definition of (ϵ, δ) differential privacy under the N-element addition/removal relation, we have for any measurable event $S \in \Theta$,

$$\mathbb{P}(\mathsf{Train}(D_- \cup W_C) \in S) = \mathbb{P}(\mathsf{Train}(D_+) \in S) \leq e^\epsilon \cdot \mathbb{P}(\mathsf{Train}(D_-) \in S) + \delta$$

when $|W_C| \leq N$, and the randomness in the probabilities is taken over the randomness in the training algorithm (for example, from stochastic gradient descent). Consider a decision algorithm $M \in \mathcal{M}$, and consider the measurable event $E_M \subseteq \Theta$

$$E_M = \{\theta \in \Theta : M(\theta) = 1\},$$

that is, the set of parameters $\theta \in \Theta$ such that the decision algorithm M suggests that θ was trained using W_C on D_+. For this particular event on a fixed M, the event bound implies

$$\mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(E_M) \leq e^\epsilon \cdot \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_-)}(E_M) + \delta.$$

Fix $\beta \in [0,1]$. By the definition of (α, β)-Access-Evidence, we are only interested in predictors $M \in \mathcal{M}_\beta$ satisfying

$$\mathcal{M}_\beta = \left\{M \in \mathcal{M} : \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_-)}\big(M(\theta) = 1\big) \leq \beta\right\}.$$

Then, for any $M \in \mathcal{M}_\beta$, we have $\mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_-)}\big(M(\theta) = 1\big) \leq \beta$. Substituting into the event bound, we have

$$\mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(E_M) = \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(M(\theta) = 1) \leq e^\epsilon \beta + \delta.$$

Taking $\alpha \geq e^\epsilon \beta + \delta$ implies that $\mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(E_M) \leq \alpha$. Since this holds for any $M \in \mathcal{M}_\beta$, it follows that

$$\sup_{M \in \mathcal{M}_\beta} \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(M(\theta) = 1) \leq \alpha,$$

so θ satisfies (α, β)-Access-Evidence with respect to any dataset D_+ and $C \in D$ with a copyrighted set W_C, as long as $|W_C| \leq N$. Now, we consider the (η, γ)-Similarity-Evidence guarantee. Fix $\eta \in [0,1]$, and define κ as

$$\kappa \;=\; \sup_{w \in \mathcal{W}} \mathbb{P}_{C \sim \pi}\big(d(w, C) < \eta\big).$$

Let $R \in \mathcal{R}$, and define the measurable success event $A \subseteq (\Theta \times \mathcal{W})$

$$A \;=\; \big\{(\theta, C) \in \Theta \times \mathcal{W} : d(R(\theta), C) \leq \eta\big\}.$$

Equivalently, for each fixed $C \in \mathcal{W}$, define the measurable slice

$$A_c \;=\; \big\{\theta \in \Theta : d(R(\theta), C) \leq \eta\big\}.$$

By the definition of (η, γ)-Similarity-Evidence, it suffices to bound $\mathbb{P}_{C \sim \pi,\ \theta \leftarrow \mathsf{Train}(D_+)}\big(d(R(\theta), C)$. In particular, we can rewrite this probability in terms of an expectation, as

$$\mathbb{P}_{C \sim \pi,\ \theta \leftarrow \mathsf{Train}(D_+)}\big(d(R(\theta), C) \leq \eta\big) \;=\; \mathbb{E}_{C \sim \pi}\Big[\mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(A_C)\Big].$$

Fix an arbitrary reference work $C_0 \in \mathcal{W}$, and let $D_0 = D_- \cup W_{C_0}$. For each fixed $C \in \mathcal{W}$, note that $D_+ = D_- \cup W_C$ and D_0 differ by at most N elements (under the addition/removal relation), so by (ϵ, δ)-differential privacy we have, for the measurable event $A_c \subseteq \Theta$,

$$\mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_+)}(A_c) \;\leq\; e^\epsilon \cdot \mathbb{P}_{\theta \leftarrow \mathsf{Train}(D_0)}(A_c) + \delta.$$

Taking expectation over $C \sim \pi$ and using linearity of expectation yields

$$\mathbb{E}_{C\sim\pi}\Big[\mathbb{P}_{\theta\leftarrow\mathsf{Train}(D_-\cup W_C)}(A_C)\Big] \;\le\; e^{\epsilon}\cdot\mathbb{E}_{C\sim\pi}\Big[\mathbb{P}_{\theta\leftarrow\mathsf{Train}(D_0)}(A_C)\Big]+\delta.$$

We now bound the remaining term. Since $\theta \leftarrow \mathsf{Train}(D_0)$ is independent of $C \sim \pi$, we may write

$$\mathbb{E}_{C\sim\pi}\Big[\mathbb{P}_{\theta\leftarrow\mathsf{Train}(D_0)}(A_C)\Big] = \mathbb{E}_{\theta\leftarrow\mathsf{Train}(D_0)}\Big[\mathbb{P}_{C\sim\pi}\big(d(R(\theta),C)<\eta\big)\Big].$$

For each fixed θ, the inner probability is bounded by κ by definition of κ (taking $w = R(\theta)$), i.e.

$$\mathbb{P}_{C\sim\pi}\big(d(R(\theta),C)\le\eta\big)\le\kappa.$$

Therefore,

$$\mathbb{E}_{C\sim\pi}\Big[\mathbb{P}_{\theta\leftarrow\mathsf{Train}(D_0)}(A_C)\Big]\le\mathbb{E}_{\theta\leftarrow\mathsf{Train}(D_0)}[\kappa]=\kappa.$$

Combining the bounds, we conclude that for any reconstruction algorithm $R \in \mathcal{R}$,

$$\mathbb{P}_{C\sim\pi,\ \theta\leftarrow\mathsf{Train}(D_+)}\big(d(R(\theta),C)\le\eta\big)\le e^{\epsilon}\kappa+\delta.$$

Thus, taking $\gamma = e^{\epsilon}\kappa + \delta$ implies that θ satisfies (η,γ)-Similarity-Evidence, with respect to D and any $C \in D$, completing the proof.

□

E.2 Proofs for Exponentially Small Probability of Reconstruction

Theorem E.2 (Privacy Implies $e^{-\Omega(d)}$ Attack Success)**.** *Assume the same setting as Proposition 4.1. Suppose that (π, d) satisfy*

$$H^{\eta}_{\mathrm{rec}}(\pi,d) = -\log\kappa(\pi,d) = \Omega(d), \qquad \kappa(\pi,d) = \sup_{w\in\mathcal{W}}\mathbb{P}_{C\sim\pi}\big(d(w,C)<\eta\big).$$

Then the reconstruction probability is exponentially small in d. In particular, with $\epsilon = \mathcal{O}(1)$, we have $\gamma = e^{-\Omega(d)} + \delta$.

Proof. By Proposition 4.1, for any reconstruction algorithm $R \in \mathcal{R}$,

$$\mathbb{P}_{C\sim\pi,\ \theta\leftarrow\mathsf{Train}(D_+)}\big(d(R(\theta),C)\le\eta\big)\le e^{\epsilon}\kappa(\pi,d)+\delta.$$

By definition of the reconstruction entropy,

$$\kappa(\pi,d) = \exp\big(-H^{\eta}_{\mathrm{rec}}(\pi,d)\big).$$

Substituting this identity into the above expression gives

$$\mathbb{P}_{C\sim\pi,\ \theta\leftarrow\mathsf{Train}(D_+)}\big(d(R(\theta),C)<\eta\big)\le\exp(\epsilon-H^{\eta}_{\mathrm{rec}}(\pi,d))+\delta.$$

If

$$H^{\eta}_{\mathrm{rec}}(\pi,d)\ge\epsilon+\rho d,$$

then

$$\exp(\epsilon-H^{\eta}_{\mathrm{rec}}(\pi,d))\le e^{-\rho d}.$$

Therefore, for every reconstruction algorithm $R \in \mathcal{R}$,

$$\mathbb{P}_{C\sim\pi,\ \theta\leftarrow\mathsf{Train}(D_+)}\big(d(R(\theta),C)<\eta\big)\le e^{-\rho d}+\delta.$$

Taking the supremum over $R \in \mathcal{R}$ shows that the model satisfies (η,γ)-Similarity-Evidence with

$$\gamma\le e^{-\rho d}+\delta.$$

In particular, if $H^{\eta}_{\mathrm{rec}}(\pi,d)-\epsilon=\Omega(d)$, then $\gamma = e^{-\Omega(d)} + \delta$, as claimed. □

Proposition E.3 (Gaussian priors with ℓ_2 reconstruction)**.** *Let $\pi = \mathcal{N}(0,\sigma^2 I_d)$ over $\mathbb{R}^d$, and consider the distance function $d(w,C) = \|w - C\|_2$. Fix $a \in (0,1)$ and set $\eta = a\sigma\sqrt{d}$. Then*

$$H^{\eta}_{\mathrm{rec}}(\pi,\|\cdot\|_2)\ge\frac{d}{2}\big(a^2-1-\log a^2\big).$$

*Consequently, if $\frac{d}{2}\big(a^2-1-\log a^2\big)\ge\epsilon+\rho d$, then an (ϵ,δ)-DP training algorithm satisfies (η,γ)-*Similarity-Evidence *with $\gamma \le e^{-\rho d} + \delta$. In particular, for any fixed $a \in (0,1)$ and $\epsilon = \mathcal{O}(1)$, we have $\gamma = e^{-\Omega(d)} + \delta$.*

Proof. We first bound the prior-only reconstruction probability

$$\kappa(\pi, \|\cdot\|_2) = \sup_{w\in\mathbb{R}^d} \mathbb{P}_{C\sim\mathcal{N}(0,\sigma^2 I_d)}\big(\|C - w\|_2 < \eta\big).$$

The Gaussian density is radially symmetric and maximized at the origin. Therefore, among all Euclidean balls of a fixed radius, the Gaussian measure is maximized by the ball centered at the origin. Hence

$$\kappa(\pi, \|\cdot\|_2) \le \mathbb{P}_{C\sim\mathcal{N}(0,\sigma^2 I_d)}\big(\|C\|_2 < \eta\big).$$

Equivalently, if $Z \sim \mathcal{N}(0, I_d)$, then

$$\mathbb{P}(\|C\|_2 < \eta) = \mathbb{P}\big(\|Z\|_2 < a\sqrt{d}\big) = \mathbb{P}\big(\chi^2_d < a^2 d\big).$$

Let $X = \chi^2_d$. For any $\lambda > 0$, Markov's inequality gives

$$\mathbb{P}(X \le a^2 d) = \mathbb{P}(e^{-\lambda X} \ge e^{-\lambda a^2 d}) \le e^{\lambda a^2 d}\mathbb{E}e^{-\lambda X}.$$

Since $X \sim \chi^2_d$,

$$\mathbb{E}e^{-\lambda X} = (1 + 2\lambda)^{-d/2}.$$

Therefore

$$\mathbb{P}(X \le a^2 d) \le \exp\left(\lambda a^2 d - \frac{d}{2}\log(1 + 2\lambda)\right).$$

Optimizing over $\lambda > 0$, choose

$$1 + 2\lambda = \frac{1}{a^2}, \qquad \lambda = \frac{1 - a^2}{2a^2} > 0.$$

Substituting this value yields

$$\mathbb{P}(X \le a^2 d) \le \exp\left(\frac{d}{2}(1 - a^2) + \frac{d}{2}\log a^2\right) = \exp\left(-\frac{d}{2}(a^2 - 1 - \log a^2)\right).$$

Thus

$$\kappa(\pi, \|\cdot\|_2) \le \exp\left(-\frac{d}{2}(a^2 - 1 - \log a^2)\right),$$

and hence

$$H^\eta_{\mathrm{rec}}(\pi, \|\cdot\|_2) = -\log \kappa(\pi, \|\cdot\|_2) \ge \frac{d}{2}(a^2 - 1 - \log a^2).$$

Applying Theorem 4.2 gives the claimed (η, γ)-Similarity-Evidence bound. □

Proposition E.4 (Shared-prefix priors with Hamming reconstruction)**.** *Let $\mathcal{W} = \{0,1\}^d$ and let d_H denote Hamming distance. Suppose the prior π fixes the first s bits of C and leaves the remaining $m = d - s$ bits uniformly random. Equivalently,*

$$C = (u, Z), \qquad u \in \{0,1\}^s, \qquad Z \sim \mathrm{Unif}(\{0,1\}^m).$$

Let $\eta = \alpha d$, and suppose $0 < \frac{\eta}{m} < \frac{1}{2}$. Then

$$H^\eta_{\mathrm{rec}}(\pi, d_H) \ge m\left(\log 2 - h\Big(\frac{\eta}{m}\Big)\right),$$

*where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function. Consequently, if $m\left(\log 2 - h\big(\frac{\eta}{m}\big)\right) \ge \epsilon + \rho d$, then an (ϵ, δ)-DP training algorithm satisfies (η, γ)-*Similarity-Evidence *with*

$$\gamma \le e^{-\rho d} + \delta.$$

Proof. We bound

$$\kappa(\pi, d_H) = \sup_{w\in\{0,1\}^d} \mathbb{P}_{C\sim\pi}\big(d_H(w, C) < \eta\big).$$

Write every candidate reconstruction as $w = (w_{\mathrm{pre}}, w_{\mathrm{free}})$, where $w_{\mathrm{pre}} \in \{0,1\}^s$ corresponds to the fixed prefix coordinates and $w_{\mathrm{free}} \in \{0,1\}^m$ corresponds to the remaining coordinates. Since the prefix of C is fixed to u, any mismatch between w_{pre} and u only increases the Hamming distance.

Therefore the supremum is attained, or at least upper bounded, by considering candidates with $w_{\mathrm{pre}} = u$. Hence

$$\kappa(\pi, d_H) \leq \sup_{v \in \{0,1\}^m} \mathbb{P}_{Z \sim \mathsf{Unif}(\{0,1\}^m)}\big(d_H(v, Z) < \eta\big).$$

For every fixed $v \in \{0,1\}^m$, the number of strings $z \in \{0,1\}^m$ satisfying $d_H(v,z) < \eta$ is at most

$$\sum_{i=0}^{\lfloor \eta \rfloor} \binom{m}{i}.$$

Therefore

$$\mathbb{P}_{Z \sim \mathrm{Unif}(\{0,1\}^m)}\big(d_H(v, Z) < \eta\big) \leq 2^{-m} \sum_{i=0}^{\lfloor \eta \rfloor} \binom{m}{i}.$$

Using the standard Hamming-ball bound, for $0 < \eta/m < 1/2$,

$$\sum_{i=0}^{\lfloor \eta \rfloor} \binom{m}{i} \leq \exp\left(m h\Big(\frac{\eta}{m}\Big)\right).$$

Thus

$$\kappa(\pi, d_H) \leq \exp\left(-m \log 2 + m h\Big(\frac{\eta}{m}\Big)\right) = \exp\left(-m\left(\log 2 - h\Big(\frac{\eta}{m}\Big)\right)\right).$$

Taking negative logarithms gives

$$H^{\eta}_{\mathrm{rec}}(\pi, d_H) = -\log \kappa(\pi, d_H) \geq m\left(\log 2 - h\Big(\frac{\eta}{m}\Big)\right).$$

Applying Theorem 4.2 gives the stated (η, γ)-Similarity-Evidence guarantee. □

F Proof of Limiting the Generation of Copyrighted Work in Theorem 4.3

We briefly restate Theorem 4.3.

Proposition F.1. *If a generative model satisfies* (η, γ)-Similarity-Evidence *with respect to* $\pi, d(\cdot,\cdot), D_+$*, then*

$$\mathbb{P}_{C\sim\pi,\theta\leftarrow\mathsf{Train}(D_+)}\left(d(C, R_{\text{user}}(D_-, \pi, p_\theta) \leq \eta\right) \leq \gamma.$$

Proof. The proof is fairly direct. Since the generative model satisfies (η, γ)-Similarity-Evidence with respect to $\pi, d(\cdot,\cdot), D_+$, we know that for any reconstruction algorithm $R \in \mathcal{R}$, we have

$$\mathbb{P}_{C\sim\pi,\theta\leftarrow\mathsf{Train}(D_+)}\left(d(C, R(\theta)) < \eta\right) \leq \gamma. \tag{2}$$

Since $R_{\text{user}}(D_-, \pi, p_\theta)$ only has query access to the model p_θ, they have access to less information than the class of adversaries bounded in (η, γ)-Similarity-Evidence. Then, the guarantee in Eq. (2) holds for the user $R_{\text{user}}(D_-, \pi, p_\theta)$. This implies that the user is unable to reconstruct the copyrighted sample C w.r.t distance $d(\cdot,\cdot)$ at threshold η, which is what we wanted to show. □

G Experiment Settings

We provide supplementary details regarding the models, datasets, implementation specifics, and evaluation protocols used in the empirical evaluation presented in Section 5. Our experiments involve a total of 28 models trained on a total of 4 datasets between 2 modalities. We discuss all these models below.

G.1 Models

In this section, we review the details of the Stable Diffusion v1.4 [3] and Llama2-7B [29] models used in our experiments.

Diffusion Models. Stable Diffusion v1.4 [3] is a latent diffusion model for text-to-image generation, which couples a pretrained variational autoencoder that maps images to a lower-dimensional latent space with a denoising diffusion model that operates in that latent space, and a text encoder that conditions generation on a prompt.

Language Models. Llama2-7B [29] is a decoder-only Transformer language model with approximately seven billion parameters. Given a sequence of tokens, the model predicts the next-token distribution using stacked self-attention and feed-forward layers with residual connections and normalization. Text is represented using a subword tokenizer, and generation proceeds autoregressively by sampling or decoding from the next-token distribution and appending tokens iteratively.

G.2 Datasets

We conduct experiments on the following datasets. All datasets are accessible via Huggingface. We partition each dataset into member set and nonmember sets, where the member set is of size 1000. These datasets are further partitioned into equally sized halves in the CP-k procedure, which requires training models on disjoint subsets of the training data (see Appendix B for further details on this sharding procedure).

MathAbstracts. AutoMathText [30] is a text dataset of mathematical paper title–abstract pairs used for conditional generation, where the title is the prompt and the abstract is the target. The papers are sourced from various websites that collect mathematical literature, including arXiv. We refer to this set as MathAbstracts, in line with previous work using this dataset [32].

WritingPrompts. WritingPrompts [31] is a corpus of prompt–story pairs for conditional long-form generation, where a short writing prompt is used to generate an associated story, archived from title-story pairs taken from a Reddit forum.

Pokémon. Pokémon is a caption–image dataset of Pokémon-style images paired with text captions, typically used for text-to-image diffusion fine-tuning [33].

LAION-MI. LAION-MI [34] is a held-out LAION-derived image dataset constructed for membership-inference evaluation, containing member and non-member examples designed to better reflect a realistic membership inference setting. We only fine-tune diffusion models on non-member samples (i.e. those not contained in the original training set of the Stable Diffusion model).

G.3 Training Algorithms and Finetuning

In this section, we recall the details of the training procedures employed in Section 5.

G.3.1 DP-Adam and Private Training

To train diffusion models with (ϵ, δ)-DP, we employ DP-Adam, a variant of the original DP-SGD method introduced in Abadi et al. [26] utilizing Adam updates rather than stochastic gradient updates. We note that the mechanism for maintaining privacy (i.e. adding noise and clipping gradients) remains the same between both samplers, as post-processing guarantees that the privacy loss is the same. We use the Opacus library [58] with Rènyi privacy accounting [59]. As discussed in the main text, we train (ϵ, δ)-DP models with $\epsilon \in \{5, 10, 20, 50\}$ and $\delta = 10^{-5}$.

G.3.2 Finetuning Settings

In the case of Stable Diffusion, we finetune all U-Net parameters. For Llama2-7b, we finetune attention parameters with low-rank adaptation [60]. We apply DP-SGD on the LoRA weights directly, which has been known to result in better privacy-utility tradeoff [61].

We formally list all finetuning hyperparameters below, with diffusion finetuning parameters in Tables 2 and 3 and language model finetuning parameters in Tables 4 and 5.

Table 2: Hyperparameters for finetuning Stable Diffusion (used for baselines, CP-k).

Category	Parameter	Value
Finetuning	Learning rate	$5 \cdot 10^{-5}$
	Batch size	16
	Epochs	50

Table 3: Hyperparameters for finetuning Stable Diffusion with DP-Adam.

Category	Parameter	Value
Finetuning	Learning rate	$5 \cdot 10^{-5}$
	Batch size	64
	Epochs	25

Table 4: Hyperparameters for finetuning Llama2-7B (used for baselines, CP-k).

Category	Parameter	Value
Finetuning	Learning rate	$1 \cdot 10^{-4}$
	Sequence length	2048
	Batch size	4
	Epochs	25
	Rank	64
	α	32

Table 5: Hyperparameters for finetuning Llama2-7B with DP-Adam.

Category	Parameter	Value
Finetuning	Learning rate	$1 \cdot 10^{-4}$
	Sequence length	2048
	Batch size	8
	Epochs	10
	Rank	64
	α	32

G.4 Samplers

We provide hyperparameters for the sampling strategies used during inference for image generation in Table 6 and text generation in Table 7. Note that stochasticity in the generation process is *necessary*, so that the log-probability ratio used in CP-k does not diverge when the token sampled is not the token generated by the sharded model.

Table 6: Hyperparameters for sampling from diffusion models.

Category	Parameter	Value
Sampling	Sampler	DDPM
	Generation Steps	20
	Resolution	512×512
	Guidance Scale	7.5

Table 7: Hyperparameters for sampling from language models.

Category	Parameter	Value
Sampling	Max New Tokens	512
	Decoding Method	Sampling
	Temperature	1.0
	Top-P	0.95
	Top-K	50

G.5 Attack Methods

In this section, we discuss the attack methods (i.e. membership inference algorithms, data reconstruction attack algorithms) used in Section 5. Precise parameter choices are specified in Section G.5.4.

G.5.1 Membership Inference against Diffusion Models

As discussed in Section 5, we use proximal initialization attacks (PIA) [35] to evaluate the performance of models in the Access game. We briefly describe this attack here for completeness.

Fix a real sample y_0. First, we obtain the model's own noise estimate at $t = 0$, given by $\varepsilon_0 = \varepsilon_\theta(y_0, 0)$. Then, we estimate the noised input at any later timestep t via the deterministic forward map

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon_0.$$

A second query yields $\varepsilon_\theta(y_t, t)$, and the attack score is measured by the ℓ_p norm difference given by

$$R_{t,p}(y_0) = \big\|\varepsilon_0 - \varepsilon_\theta(y_t, t)\big\|_p.$$

Since training samples tend to reproduce the model's proximal initialization more faithfully, smaller values of $R_{t,p}$ indicate higher likelihood of membership in the training set [35]. Hence, the formal attack may be written as

$$f(y_0) = \mathbf{1}[R_{t,p} < \tau],$$

where τ is some threshold adjusted based on the desired FPR. In our experiments, we choose t to maximize the AUROC of the attack curve for each relevant attack, since the plaintiff aims to show the most vulnerability in the defendant's model and need not fix the parameters of their attacks across defendants. In general, regardless of the choice of model, this maximizer was given by $t \approx 200$.

G.5.2 Membership Inference against Language Models

To evaluate membership inference in the Access game for auto-regressive language models, we use the Min-K%++ detector of Zhang et al. [36], which is a strengthened baseline for pre-training data detection that operates under grey-box access to the target model, requiring only token logits or token probabilities.

Fix a tokenized input sequence $x_{1:T}$, and let $p_\theta(\cdot \mid x_{<t})$ denote the model's conditional distribution over the vocabulary $\mathcal{V}$ at position t. Min-K%++ assigns a token-wise score by calibrating the log probability of the realized token x_t using statistics of the full categorical distribution. In particular, define the random variable $\ell_t = \log p_\theta(x_t \mid x_{<t})$ with mean and variance

$$\mu_t = \mathbb{E}_{w \sim p_\theta(\cdot \mid x_{<t})}\big[\log p_\theta(w \mid x_{<t})\big], \quad \sigma_t^2 = \mathsf{Var}_{w \sim p_\theta(\cdot \mid x_{<t})}\big[\log p_\theta(w \mid x_{<t})\big],$$

and set the calibrated token score to be

$$s_t = \frac{\ell_t - \mu_t}{\sigma_t}.$$

Intuitively, this normalization measures whether the realized token x_t is assigned unusually large probability mass relative to the model's entire next-token distribution at that prefix, rather than relying on the absolute value of ℓ_t alone [36]. To obtain a sentence-level statistic, Min-K%++ follows Min-K% in selecting the $k\%$ of tokens with the smallest token scores and averaging over them [36]. Writing $S_k(x)$ for the resulting score, we define

$$S_k(x_{1:T}) = \frac{1}{|I_k|} \sum_{t \in I_k} s_t, \qquad I_k \in \arg\min_{I \subseteq [T],\, |I| = \lceil kT/100 \rceil} \sum_{t \in I} s_t.$$

Since training examples tend to exhibit fewer low-scoring "outlier" positions, larger values of S_k provide evidence of membership [36]. The associated threshold test is

$$f(x_{1:T}) = \mathbf{1}\big[S_k(x_{1:T}) > \tau\big],$$

where τ is selected to meet a desired false positive rate.

In our experiments, we select k by sweeping over $k \in \{10, 20, \ldots, 100\}$ and reporting the best-performing choice for each relevant attack setting, mirroring the evaluation protocol used for Min-K%++ when a fixed validation set is unavailable [36]. We then choose τ according to the desired operating point on the ROC curve.

G.5.3 Data Reconstruction Attacks against Diffusion and Language Models

As described in Section 5, we evaluate the Similarity game using reconstruction attacks against both image and language models. For diffusion models, we use a modified version of the data reconstruction attack of Carlini et al. [37]; for language models, we use a reconstruction procedure adapted from the scalable extraction methodology of Nasr et al. [38]. In both cases, the goal is to recover a fixed target sample $y^\star$ from model generations from a known prompt $z^\star$.

First, fix a target sample $y^\star$ and let $z^\star$ denote its associated prompt, which we treat as public context. We draw n independent samples $y^i \sim p(\cdot \mid z^\star), i \in [n]$. We then rank these candidates using the modality-specific membership score from the corresponding Access evaluation.

In particular, for diffusion models, each candidate image is assigned the PIA score $R_{t,p}(y^i)$, computed as in Section G.5.1, and candidates are sorted in ascending order of $R_{t,p}$. For language models, each completion is assigned the Min-K%++ score $S_k(y^i)$, computed as in Section G.5.2, and candidates are sorted in descending order of S_k. This ranking step prioritizes the generation most likely to reflect memorized training content, following the role of membership scoring in scalable extraction [38].

Finally, let $\widehat{y}$ denote the top-ranked candidate after this modality-specific ordering. We measure reconstruction success by comparing $\widehat{y}$ directly to the target $y^\star$. Let $d(\cdot, \cdot)$ denote a modality-specific distance. We say that y successfully reconstructs y^* if $d(\widehat{y}, y^\star) \leq \eta$. Consequently, the empirical reconstruction success probability is therefore estimated across target samples by

$$\mathbb{P}\big(d(\widehat{y}, y^\star) \leq \eta\big) \approx \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\big(d(\widehat{y}_j, y_j^\star) \leq \eta\big),$$

where $y_j^\star$ is the jth target sample and $\widehat{y}_j$ is the corresponding top-ranked reconstruction. In our experiments, we vary the distance metric $d(\cdot, \cdot)$ and threshold η to capture different notions of visual or textual similarity, as further discussed in Section G.7.

G.5.4 Attack Settings

We list all hyperparameters and settings for attack methods below, with membership inference settings in Table 8 and data reconstruction settings in Table 9.

Table 8: Hyperparameters for membership inference attacks against Stable Diffusion and Llama2-7B.

Model	Attack	Parameter	Value
Stable Diffusion	PIA [35]	ℓ_p norm Step t	$p = 4$ $t = 200$
Llama2-7B	MinK%++ [36]	$k\%$	$\{5, 10, 20, 30, 40, 50\}\%$

Table 9: Hyperparameters for data reconstruction attacks against Stable Diffusion and Llama2-7B.

Model	Attack	Parameter	Value
Stable Diffusion	DRA with PIA, Section G.5.3	N, generations/prompt	10
Llama2-7B	DRA with MinK%++, Section G.5.3	N, generations/prompt	10

For data reconstruction, we employ the sampling parameters discussed in Section G.4.

G.6 Implementing the CP-k Algorithm

We discuss the specifics of implementing and attacking models that utilize the CP-k algorithm. For a discussion on the CP-k algorithm itself, refer to Appendix B or the original formulation by Vyas et al. [15].

Implementation. To avoid naïvely wasting the majority of samples when the k_0-threshold is low, we instead generate many reconstructions and estimate quantiles based on the empirical distribution of log-probabilities. This lets us obtain fine-grained control over the number of samples that are accepted or rejected by using α_k, which improves experiment scalability and has been employed in previous work [57].

Measuring Log-Probabilities. For language models, log-probabilities are computed directly from the autoregressive factorization of a generated sequence. Given a prompt z and completion

$y = (y_1, \ldots, y_T)$, we evaluate

$$\log p_\theta(y \mid z) = \sum_{t=1}^{T} \log p_\theta(y_t \mid z, y_{<t}),$$

using teacher forcing under the model being scored. Thus, each generated completion receives a sequence-level score equal to the sum, or length-normalized average, of its token log-probabilities. For diffusion models, with a fixed prompt z, we accumulate the Gaussian transition log-densities assigned by the model at each denoising step, a proxy which has been used by previous work implementing the CP-k algorithm [15, 57].

In our CP-k implementation, these log-probability scores are used to compare the draft model p against shard models q_1, q_2, releasing a sample only when the corresponding log-likelihood ratio is below the CP-k threshold: see Appendix B for further discussion on the CP-k algorithm.

Membership Inference Attacks. For the CP-k sampler, we estimate the log probability ratio, $\max_{i\in\{1,2\}} \log(p(y|z), q_i(y|z))$ with p, q_i as given in Appendix B, of member and nonmember samples by running the forward diffusion process on the partially noised sample y_t as a proxy for true sample generation. When a certain sample exceeds the threshold, we simply exclude it from the attack. This mimics the realistic scenario where a defendant's model, which implements CP-k, will not release information about samples that exceed the fixed threshold k_0.

Data Reconstruction Attacks. Our data reconstruction attacks only require repeated model samples. Consequently, we estimate k_0 thresholds using the quantile approach above, and compute success rates on samples with log-probabilities that fall below this threshold.

G.7 Distance Metrics for Similarity

We evaluate a variety of metrics for Similarity, which we outline below. In all cases, we normalize each score into the range $[0, 1]$, where a smaller value indicates higher similarity, and a higher value indicates less similarity.

G.7.1 Image Generation

We discuss the following metrics used in Section 5 for evaluating image reconstruction.

Contrastive Language-Image Pre-training (CLIP) Similarity. CLIP similarity [39] measures the cosine similarity between the CLIP embeddings of two images in order to capturing high-level semantic alignment. CLIP similarity has been widely used for text-to-image evaluation and similarity assessment [37, 62].

DreamSim. DreamSim [40] is a recent perceptual similarity metric trained to align with human judgments of mid-level image similarity, capturing properties such as layout and object identity that are captured by neither pixel-level and purely semantic similarity.

We also discuss the following two metrics in Appendix H.

ℓ_2 Distance. The ℓ_2 distance is a canonical measure of near-exact similarity, such that the distance between images $x, y \in \mathbb{R}^d$ is given by $\|x - y\|_2/\sqrt{d}$. The ℓ_2 distance has been used as a standard proxy for near-duplicate detection in the context of training data extraction from diffusion models [37].

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS [63] computes a weighted ℓ_2 distance between deep feature activations of two images extracted from a pretrained network, providing a perceptually calibrated similarity measure.

G.7.2 Text Generation

We discuss the primary metrics used in our experiments below for evaluating text reconstruction.

ROUGE-L. ROUGE-L [41] measures the length of the longest common subsequence between a candidate and reference string, normalized by the lengths of both sequences. It captures sentence-level structural overlap beyond strict n-gram matching and is widely used in summarization and generation evaluation [32]

BERTScore. BERTScore [42] computes token-level similarity between a candidate and reference text using contextual embeddings from a pretrained language model: conditioned on the expressive strength of the contextual embeddings, BERTScore has demonstrated strong correlation with human judgments across a variety of text generation tasks.

We also evaluate the following additional metrics that appear only in Appendix H.

Normalized Levenshtein Distance. The Levenshtein (edit) distance counts the minimum number of character-level insertions, deletions, and substitutions required to transform one string into another. We normalize by the length of the longer string to obtain a score in $[0, 1]$. This metric is robust to minor surface-level variations and has been used for verbatim similarity assessment in copyright studies [32].

BLEU. BLEU [64] measures n-gram precision between a candidate and reference text, with a brevity penalty to discourage short outputs. BLEU is a standard metric for evaluating textual overlap and has been employed to assess verbatim reproduction in copyright evaluation of language models [32]. We measure the *smoothed* n-gram precision, which averages the BLEU score across $n \in \{1, 2, 3, 4\}$ and has been used for similarity evaluation [32].

G.7.3 Distance Metric Settings

We list models and settings for all relevant metrics below.

Table 10: Hyperparameters for distance functions used in data reconstruction evaluations. If no additional hyperparameters are required for the distance metric, we report "–".

Modality	Distance Function	Parameter	Value
Text	Normalized Levenshtein	–	–
	BLEU	–	–
	ROUGE-L	`rouge_l_use_stemmer`	`false`
	BERTScore-F1	`bertscore_model`	–
		`bertscore_lang`	`en`
Image	Normalized ℓ_2	–	–
	LPIPS	`lpips_backbone`	`alex`
	CLIP	`clip_model`	`clip-vit-base-patch32`
	DreamSim	`dreamsim_model`	`ensemble`
		`dreamsim_pretrained`	`true`

G.8 Utility Evaluations

Here, we discuss the evaluation metrics used to confirm the performance of the models studied in Section 5.

G.8.1 Evaluating Diffusion Model Utility

We evaluate diffusion model utility using KID, CLIPScore, and CLIP-IQA. KID measures whether generated images match the distributional statistics of real images in a pretrained visual feature space, making it a standard proxy for sample quality [43]. CLIPScore measures semantic alignment between generated images and their conditioning captions using CLIP embeddings, making it useful for text-to-image evaluation [62]. CLIP-IQA assesses the perceptual quality and aesthetics of generated images using CLIP-embeddings, evaluating generation fidelity beyond distributional and semantic measures [44].

G.8.2 Evaluating Language Model Utility

We evaluate language model utility using perplexity and fluency. Perplexity is computed using an external model (Mistral-7B [65]) on held-out completions, capturing whether protection mechanisms degrade the model's likelihood-based language modeling ability [4]. Fluency is measured with PrometheusV2 [45, 46, 47], which provides an automatic judge of generation quality under a

specified rubric, capturing notions of coherence. This protocol has previously been employed to assess language model utility in the context of copyright infringement study [32].

G.8.3 Utility Evaluation Settings

We list parameters used for utility evaluations in Table 11.

Table 11: Hyperparameters and evaluation models used for utility evaluation. If no additional hyperparameters are required for the metric, we report "–".

Modality	Utility Metric	Parameter	Value
Text	External Perplexity	`reference_model`	`Mistral-7B`
	External Perplexity	`evaluation_split`	`held-out completions`
	Fluency	`judge_model`	`PrometheusV2`
		`rubric`	`fluency/coherence`
Image	KID	`feature_space`	–
	CLIPScore	`embedding_model`	`clip-vit-base-patch32`
	CLIP-IQA	`embedding_model`	`clip-vit-base-patch32`

G.9 Numerical Precision

All experimentation, including model training and inference, were completed in FP16, with the exception of log-probability calculations for CP-k, which was carried out in FP32 to avoid underflow and/or overflow.

G.10 Hardware

All model training and experimentation was performed on 2 NVIDIA H200 GPU(s) with an Intel Xeon Platinum 64-core processor.

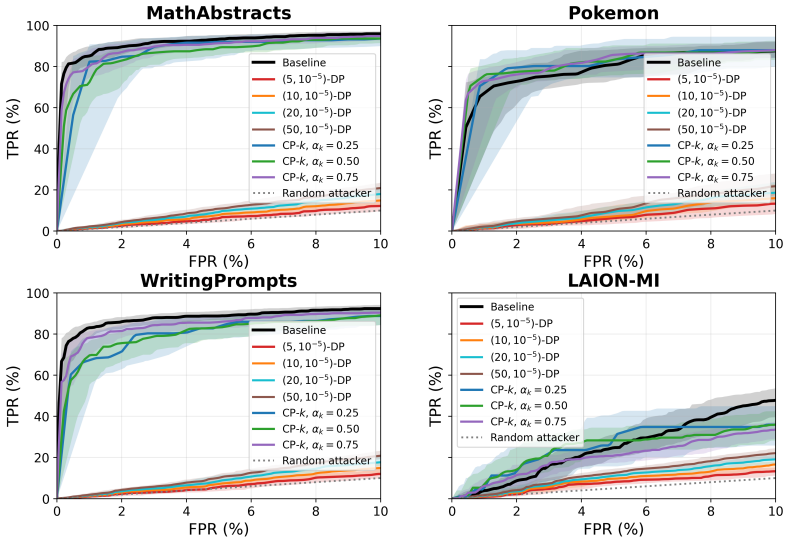
H Additional Experimental Results

We discuss additional experimental results below for the sake of completeness: this includes additional distance metrics, additional utility metrics, and additional ablations among (ϵ, δ)-DP models.

H.1 Additional Access Results

We provide additional results for Access here, including all (ϵ, δ)-DP models.

Figure 4: Full TPR-FPR tradeoff plot, including all (ϵ, δ)-DP models with $\epsilon \in \{5, 10, 20, 50\}$.



As mentioned in Section 5, models satisfying our notion of (α, β)-Access-Evidence have substantially lower vulnerability to membership inference. This is made clear by Fig. 4 across all choices of privacy parameters $\epsilon \in \{5, 10, 20, 50\}$, as indicated by the substantially less vulnerable curves present in Fig. 4 relative to the CP-k algorithm, whose curves appear to have no clear correlation with improving or reducing vulnerability to membership inference. Interestingly, LAION-MI seems to reduce the advantage of attackers considerably. We hypothesize that the somewhat unstructured nature of LAION-MI (in comparison to MathAbstracts, WritingPrompts, and Pokémon, which have shared attributes and styles) [34] is responsible for the lower power of membership inference attackers.

Table 12: Additional TPR at fixed FPR metrics for Pokémon and LAION-MI datasets.

	LAION-MI			**Pokémon**		
Model	TPR@FPR1	TPR@FPR5	TPR@FPR10	TPR@FPR1	TPR@FPR5	TPR@FPR10
Baseline	0.040 ± 0.040	0.248 ± 0.097	0.480 ± 0.072	0.705 ± 0.212	0.812 ± 0.075	0.878 ± 0.052
CP-k, $\alpha_k = 0.25$	0.112 ± 0.101	0.315 ± 0.146	0.360 ± 0.090	0.791 ± 0.154	0.835 ± 0.099	0.879 ± 0.077
CP-k, $\alpha_k = 0.50$	0.075 ± 0.084	0.283 ± 0.058	0.363 ± 0.084	0.762 ± 0.110	0.841 ± 0.088	0.877 ± 0.048
CP-k, $\alpha_k = 0.75$	0.062 ± 0.045	0.212 ± 0.049	0.338 ± 0.056	0.739 ± 0.075	0.861 ± 0.080	0.883 ± 0.035
$(5, 10^{-5})$-DP	0.008 ± 0.012	0.078 ± 0.017	0.125 ± 0.024	0.012 ± 0.025	0.060 ± 0.048	0.125 ± 0.059
$(10, 10^{-5})$-DP	0.010 ± 0.011	0.089 ± 0.017	0.150 ± 0.023	0.014 ± 0.023	0.064 ± 0.042	0.143 ± 0.072
$(20, 10^{-5})$-DP	0.013 ± 0.010	0.100 ± 0.017	0.168 ± 0.024	0.016 ± 0.018	0.083 ± 0.047	0.162 ± 0.073
$(50, 10^{-5})$-DP	0.014 ± 0.011	0.110 ± 0.018	0.189 ± 0.023	0.018 ± 0.023	0.088 ± 0.045	0.185 ± 0.066

We also report additional TPR at fixed FPR metrics in Tables 12 and 13 across modalities and datasets. As in Section 5, we observe a categorical reduction in vulnerability to Access-Accusation with differentially private models, which provide a sufficient condition for (α, β)-Access-Evidence and (η, γ)-Similarity-Evidence. Such a reduction occurs even in a more challenging membership inference setting, such as LAION-MI.

Table 13: Additional TPR at fixed FPR metrics for MathAbstracts and WritingPrompts datasets.

	MathAbstracts			WritingPrompts		
Model	TPR@FPR1	TPR@FPR5	TPR@FPR10	TPR@FPR1	TPR@FPR5	TPR@FPR10
Baseline	0.852 ± 0.052	0.933 ± 0.020	0.959 ± 0.014	0.830 ± 0.047	0.888 ± 0.021	0.924 ± 0.019
CP-k, $\alpha_k = 0.25$	0.827 ± 0.273	0.920 ± 0.032	0.936 ± 0.040	0.683 ± 0.137	0.847 ± 0.084	0.888 ± 0.044
CP-k, $\alpha_k = 0.50$	0.784 ± 0.128	0.892 ± 0.038	0.936 ± 0.026	0.698 ± 0.135	0.831 ± 0.056	0.891 ± 0.040
CP-k, $\alpha_k = 0.75$	0.812 ± 0.069	0.913 ± 0.031	0.948 ± 0.025	0.782 ± 0.079	0.859 ± 0.031	0.905 ± 0.024
$(5, 10^{-5})$-DP	0.013 ± 0.009	0.064 ± 0.024	0.116 ± 0.030	0.010 ± 0.008	0.054 ± 0.020	0.112 ± 0.024
$(10, 10^{-5})$-DP	0.015 ± 0.009	0.074 ± 0.023	0.134 ± 0.026	0.012 ± 0.008	0.065 ± 0.021	0.133 ± 0.024
$(20, 10^{-5})$-DP	0.018 ± 0.007	0.085 ± 0.022	0.155 ± 0.028	0.015 ± 0.011	0.075 ± 0.020	0.154 ± 0.023
$(50, 10^{-5})$-DP	0.020 ± 0.008	0.093 ± 0.019	0.176 ± 0.032	0.017 ± 0.010	0.087 ± 0.022	0.176 ± 0.023

H.2 Additional Similarity Results

We provide additional results relating to our empirical evaluation of the Similarity reconstruction game.

H.2.1 Evaluating Similarity for Diffusion Models

We begin by presenting a full set of empirical reconstruction probabilities against various finetuned variants of Stable Diffusion for different distance metrics, including exact-matching metrics (ℓ_2), perceptual metrics (LPIPS, DreamSim), and semantic similarity metrics (CLIP Similarity).

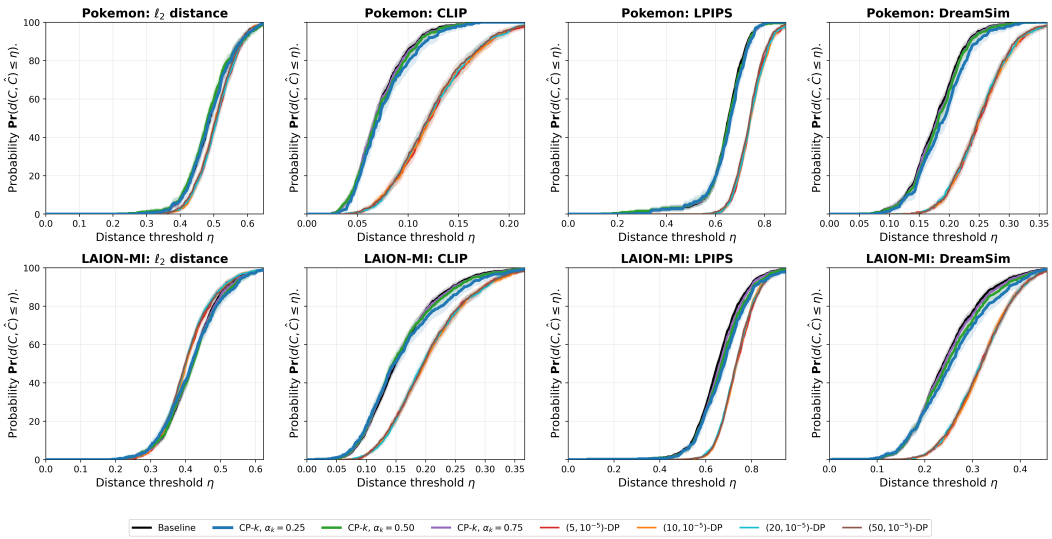


Figure 5: Full evaluation of reconstruction success within threshold η of the copyrighted target across all image datasets and all distance metrics.

As discussed in Section 5, our experiments here, as seen in Fig. 5, clearly demonstrate that models satisfying (η, γ)-Similarity-Evidence are much less vulnerable to data reconstruction attacks than both baseline models and those satisfying k-NAF. The difference in reconstructability appears to be the most clear when using appropriate, robust distance metrics, such as DreamSim or LPIPS [40, 63]. In comparison, when using near-exact distance measures such as ℓ_2, reconstruction success appears to be broadly limited. This motivates our claim that (η, γ)-Similarity-Evidence ought to require the usage of a robust distance function, in order to avoid only considering exact closeness and ignoring other perceptual and compositional notions of similarity that may be relevant to accusations of a model containing reconstructible *expression*.

As before, we also study the relationship between reconstruction success and the parameter α_k used in the CP-k algorithm. It is clear from Fig. 6 that this relationship is weak at best: vulnerability to reconstruction attack appears to persist regardless of the value of α_k, suggesting that the k-NAF algorithm provides no protection of internal model content.

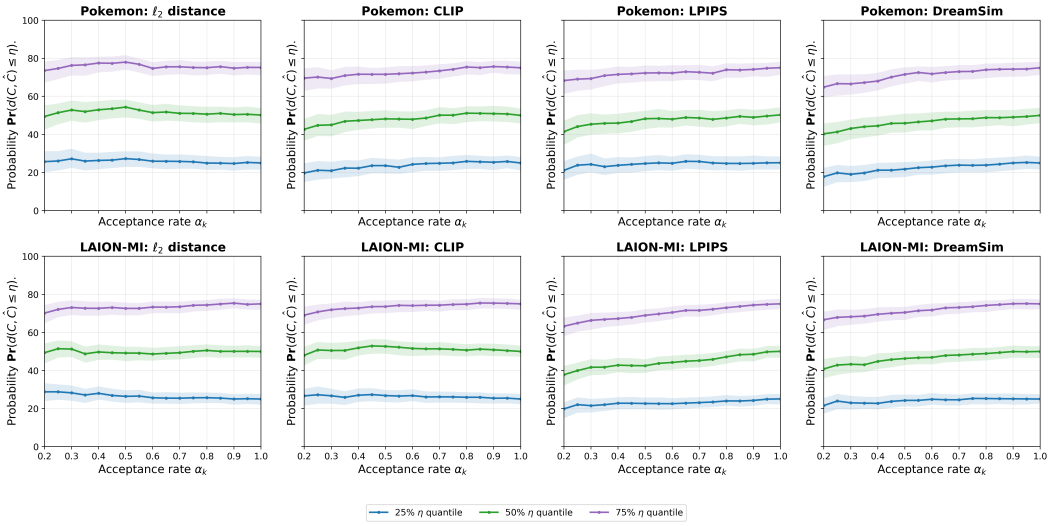


Figure 6: Similarity attack success at fixed thresholds across all image datasets and distance metrics. We observe little relation between the choice of α_k and the observed reconstruction success.

H.2.2 Evaluating Similarity for Language Models

In the same manner as the prior section, we provide a full evaluation scheme of reconstruction attacks against language models to empirically understand Similarity evaluation.

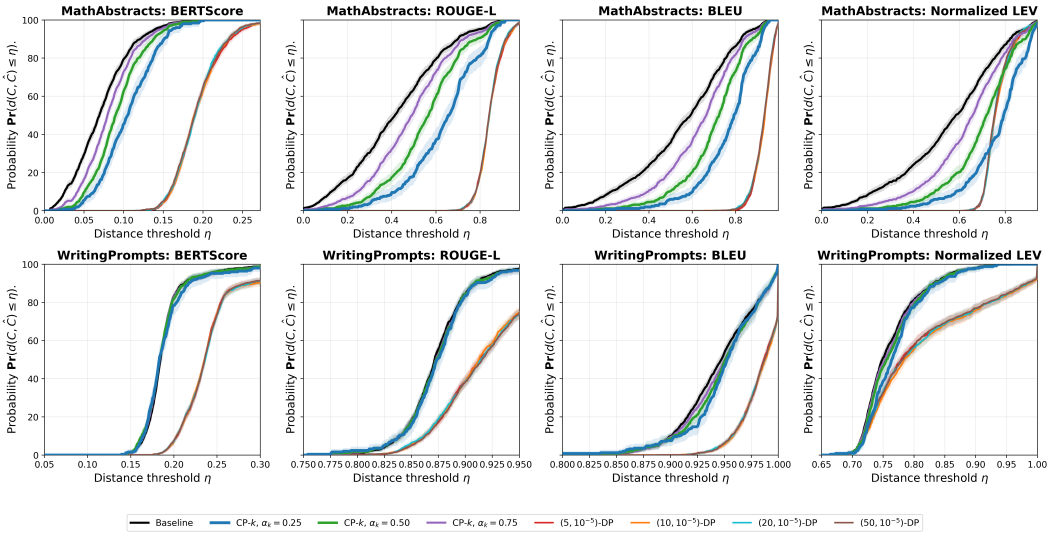


Figure 7: Full evaluation of reconstruction success for text generations from copyrighted target across all image datasets and distance metrics.

Curiously, different datasets and distance functions appear to have nontrivial effects on the reconstructibility of data. In particular, it appears that it is substantially *easier* to reconstruct data from the MathAbstracts dataset, while extracting data appears to be categorically more challenging from WritingPrompts. This is likely because of the highly specific structure of math papers and abstracts compared to the free-form nature of creative writing prompts, promoting memorization and thus reconstructability in the former case. Nevertheless, Figs. 7 and 8 still indicate that reconstruction is far easier without the protection of (η, γ)-Similarity-Evidence, suggesting that our notions of copyright protection are more suitable than NAF for a model-centric view of copying.

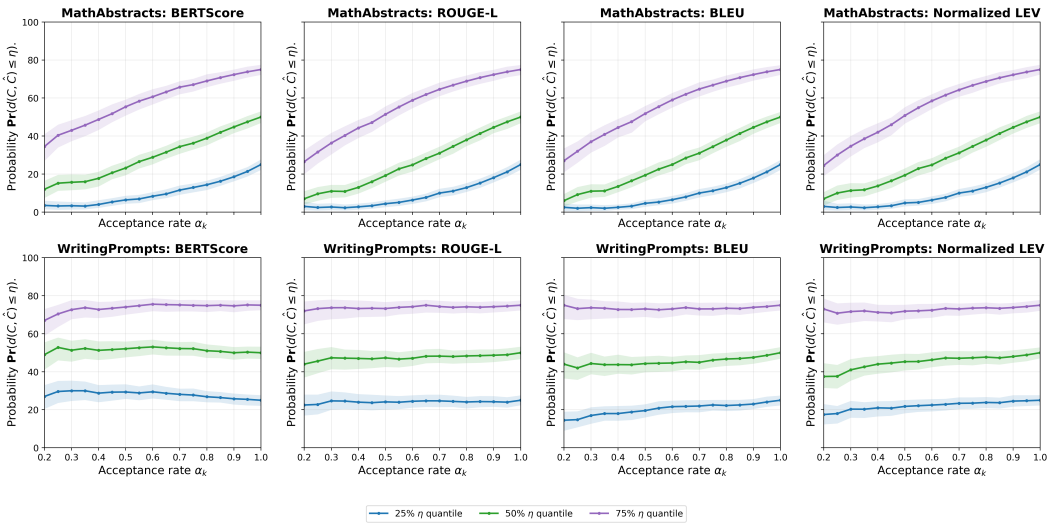


Figure 8: Estimated Similarity attack success at fixed distance thresholds. Although some reduction is observed with small α_k, it is not sufficient to prevent reconstruction relative to baselines.

H.3 Additional Utility Results

We restate utility metrics in this section for completeness, including standard deviations. We note improved quality relative to CP-k in certain metrics in Table 14, such as KID: we attribute this change due to the CP-k rejection sampling approach, which lowers sample diversity and thus penalizes diversity-based distance measures, such as KID.

For language models evaluated in Table 15, we note that external perplexity and fluency measure broad linguistic quality rather than task-specificity or memorization. Thus, non-private and CP-k models can score well by generating fluent but highly specific text, while DP training tends to regularize away training-specific information, trading some specificity for reduced memorization risk [37, 61]. We hypothesize that this results in the lower external model perplexity observed in our results.

Table 14: Diffusion model utility evaluation across Pokémon and LAION-MI.

	Pokémon			**LAION-MI**		
Model	KID ↓	CLIPScore ↑	CLIP-IQA ↑	KID ↓	CLIPScore ↑	CLIP-IQA ↑
Baseline	6.32×10^{-4}	0.337 ± 0.028	0.812 ± 0.103	7.69×10^{-5}	0.356 ± 0.035	0.653 ± 0.165
CP-k, $\alpha_k = 25\%$	7.30×10^{-4}	0.32 ± 0.030	0.799 ± 0.107	1.80×10^{-4}	0.339 ± 0.036	0.649 ± 0.112
CP-k, $\alpha_k = 50\%$	2.35×10^{-4}	0.335 ± 0.030	0.782 ± 0.105	1.44×10^{-4}	0.322 ± 0.035	0.613 ± 0.166
CP-k, $\alpha_k = 75\%$	2.38×10^{-4}	0.336 ± 0.029	0.721 ± 0.102	1.15×10^{-4}	0.338 ± 0.031	0.612 ± 0.099
DP, $\varepsilon = 50$	6.63×10^{-4}	0.347 ± 0.022	0.756 ± 0.103	8.32×10^{-5}	0.328 ± 0.035	0.599 ± 0.144
DP, $\varepsilon = 20$	6.94×10^{-4}	0.342 ± 0.13	0.734 ± 0.081	8.50×10^{-5}	0.327 ± 0.192	0.605 ± 0.148
DP, $\varepsilon = 10$	6.88×10^{-4}	0.334 ± 0.011	0.737 ± 0.101	8.60×10^{-5}	0.309 ± 0.011	0.591 ± 0.143
DP, $\varepsilon = 5$	7.02×10^{-4}	0.330 ± 0.029	0.729 ± 0.099	9.00×10^{-5}	0.301 ± 0.021	0.589 ± 0.092

Table 15: Language model utility evaluation across MathAbstracts and WritingPrompts.

	MathAbstracts		**WritingPrompts**	
Model	$\mathrm{PPL}_{\mathrm{ext}} \downarrow$	$\mathrm{FLU} \uparrow$	$\mathrm{PPL}_{\mathrm{ext}} \downarrow$	$\mathrm{FLU} \uparrow$
Baseline	6.29 ± 2.17	3.07 ± 0.85	7.06 ± 2.47	3.86 ± 1.79
CP-k, $\alpha_k = 25\%$	6.33 ± 2.23	3.07 ± 0.85	6.81 ± 2.54	4.24 ± 1.57
CP-k, $\alpha_k = 50\%$	6.41 ± 2.42	3.07 ± 0.85	6.94 ± 2.49	3.90 ± 1.74
CP-k, $\alpha_k = 75\%$	6.28 ± 2.20	3.07 ± 0.85	7.00 ± 2.46	4.00 ± 1.70
DP, $\varepsilon = 50$	2.49 ± 0.85	3.00 ± 1.01	4.90 ± 1.48	3.87 ± 1.47
DP, $\varepsilon = 20$	2.45 ± 0.79	2.95 ± 1.03	4.91 ± 1.47	3.64 ± 1.42
DP, $\varepsilon = 10$	2.47 ± 0.80	3.00 ± 1.05	4.87 ± 1.13	3.65 ± 1.43
DP, $\varepsilon = 5$	2.46 ± 0.80	3.01 ± 0.77	4.93 ± 1.52	3.58 ± 1.21

I Limitations

The key limitation of our work is that we do not focus on developing training algorithms specifically tailored to satisfy the proposed copyright criteria. Consequently, there is significant room for future work in designing training methods that better balance utility and copyright compliance, potentially improving over approaches such as differentially private training.

J Broader Impacts

Our work introduces a theoretical framework for quantifying evidence of copyright infringement by generative AI models. We believe this provides essential groundwork for integrating quantitative, legally interpretable evidence into copyright adjudication and regulatory processes.

We do not foresee significant negative societal impacts from this work, since our contribution is primarily theoretical and does not involve the release of high-risk models or data.