# A Unified Framework for Adversary-Aware Differential Privacy Bounds

**Marika Swanberg*** Google Research
**Meenatchi Sundaram Muthu Selva Annamalai*** University College London
**Jamie Hayes** Google DeepMind
**Borja Balle†** Optiak
**Adam Smith‡** Boston University

## Abstract

Differential Privacy (DP) bounds the privacy leakage of a mechanism against worst-case membership inference, but the precise tradeoff between complex adversarial models and DP protections remains poorly understood. In this paper, we present a unified framework that generalizes the patchwork of existing bounds across membership inference, attribute inference, and data reconstruction attacks. Crucially, our framework is the first to evaluate attacks that target multiple individuals simultaneously and measure success beyond exact matches under a single cohesive bound. Our bounds capture this broad family of previously unexplored attack settings by relying solely on the privacy parameters and the adversary's baseline success rate (i.e. its prior without access to the mechanism's output), and they remain valid when only an upper confidence bound on this rate is available. To illustrate this, we compare our high-probability guarantees to empirical attacks in two novel settings: extracting multiple non-uniform secrets (passwords and PII) from DP-finetuned language models, and reconstructing tabular data from noisy marginals. Ultimately, this framework provides a rigorous theoretical foundation to investigate the risk landscape of DP algorithms in new adversarial settings.

## 1 Introduction

Differential privacy (DP) [26] has emerged as a standard for designing mechanisms that process sensitive data in a variety of deployment contexts [2, 6, 49, 51, 61]. Unlike ad-hoc methods, DP provably bounds the amount of information that a mechanism reveals about any individual using the privacy parameter(s), with larger parameters typically giving weaker privacy guarantees.

A central challenge in DP deployments is how to interpret the privacy protections of the chosen privacy parameters in context. The DP guarantee naturally bounds the success of a worst-case membership inference (MI) adversary, who attempts to distinguish between the outputs of the mechanism on two inputs that differ on a single user's data [37]. While there are numerous good reasons to use this threat model in *defining* DP[4], it provides little insight into *accurately quantifying* the risk under more challenging adversarial modeling assumptions. Consider a DP deployment that uses $\varepsilon = 1$. The DP bound tells us that worst-case MI succeeds with probability at most $0.73$[5]. Now suppose the primary concern for this deployment isn't membership inference but rather the total number of users' data that

*Equal contribution

†Work done while at Google DeepMind

‡Work done in part while at Google Deepmind.

[4]To name a few: worst-case MI *is* the correct threat model for some mechanisms like randomized response, the guarantee is dataset-agnostic, it implies protections against less-informed adversaries, it allows us to compare different mechanisms, etc.

[5]In general the probability of success against worst-case MIA for any $\varepsilon$-DP mechanism is at most $\frac{e^\varepsilon}{e^\varepsilon+1}$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

| Reference | Prior Knowledge | | Attack Goal | | | No. of Targets | | Success Metric | |
|---|---|---|---|---|---|---|---|---|---|
| | Uniform | Non-Uniform | MI | AI | Recon | Single | Multiple | Simple | Complex |
| Balle et al. [8] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| Stock et al. [65] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ |
| Hayes et al. [31] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| Cummings et al. [19] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| Steinke et al. [64] | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ |
| Mahloujifar et al. [46] | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ |
| Cherubin et al. [14] | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ |
| Kulynych et al. [42] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| **Ours** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparing the applicability of our framework with prior work in various settings (see Section 2.1). For multiple targets, our bounds apply to success metrics that count the number of individually successful targets, e.g., exact match or approximate reconstruction within a per-record distance threshold (formalized as *decomposable* metrics in Definition 3.1).

could be partially reconstructed, possibly using prior information about the user population. Until now, practitioners have not had any means to *translate* the $0.73$ probability of worst-case MI into context-dependent privacy risks. This problem is magnified for privacy parameters that provide only "vacuous guarantees" against worst-case MI but which may empirically protect against the risks that are relevant to the deployment.

Practitioners in this situation often rely on empirical privacy attacks to gauge risk. However, this presents several challenges. First, empirical attacks merely establish lower bounds on privacy leakage; future, more sophisticated attacks could easily invalidate current risk assessments. Furthermore, state-of-the-art attacks are often highly tuned to specific mechanisms, hyperparameter settings, or data distributions, making them difficult to generalize. For example, even within DP-SGD, epsilons ranging from $10^2$ to $10^9$ have thwarted state-of-the-art reconstruction attacks [8, 10, 56]. Unless a practitioner's exact deployment and threat model mirrors a published experiment, empirical findings offer modest actionable guidance. Furthermore, as Cummings et al. [19] argue, many attack settings that are relevant to practitioners have not been studied yet. In the end, the attacks' findings are boiled down to general rules of thumb like the "undocumented but still widely used goal for DP ML models of achieving an $\varepsilon \leq 10$" [57].

### 1.1 Our Contributions

In this work, we formalize precise, adversary-aware DP upper bounds that quantify privacy risks across previously uncharacterized threat models. While prior bounds have largely treated membership inference, attribute inference, and exact reconstruction as isolated scenarios, our framework mathematically bridges these paradigms. Specifically, we provide the machinery to bound attacks along three simultaneous dimensions: the scale of the attack (targeting multiple individuals), the skew of the adversary's prior knowledge (handling non-uniform distributions), and the strictness of the success criteria (accommodating approximate, rather than purely exact, data recovery). At a technical level, our bounds achieve this flexibility by depending exclusively on the privacy parameters and the adversary's baseline probability of making a successful guess without access to the mechanism's output. Importantly, this baseline need not be known exactly: our bounds remain valid when it is replaced by a conservative upper estimate, such as a confidence bound computed from auxiliary data. More specifically:

**Unifying and Improving Previous Work.** Our bounds unify the patchwork of bounds focused on various settings of membership inference (MI) [14, 46, 64], attribute inference (AI) [14], and reconstruction [8, 18, 19, 31, 42, 65]. Technically, we build on two standard tools: the DP bound on prior-to-posterior odds, and the coupling and linear-programming arguments of one-run auditing [64]. Our key conceptual step, following Narcissus resiliency [18], is to compare an adversary's success to its baseline success on a fresh sample from the prior *for the specific attack attempt it makes*.

ReRo [8, 31] instead compares it to the single most likely record. Prior one-run analyses assume uniform membership bits and exact matches. Extending them to arbitrary priors and approximate success criteria required reworking the coupling and the treatment of $\delta$ (Appendix B). Beyond generalizing prior bounds, the result is often tighter: our bound strictly improves on the Narcissus resiliency bound for DP mechanisms; it is substantially tighter than KGK+ [42] for non-uniform priors (Figure 1); it closely matches MMC [46] on uniform priors without requiring ordered canaries or exact index matching; and it recovers ReRo as a special case while avoiding ReRo's dependence on the most likely record. Detailed comparisons are in Appendix G.

**Covering New Settings.** The new bounds' generality allows them to be applied in previously unexplored settings. Table 1 summarizes several dimensions in which one can compare existing bounds to ours; the table's terminology is detailed in Section 2.1.

**Empirical Case Studies.** To illustrate the versatility of our bounds and compare them to the success of concrete attacks, we study two specific settings that have been unexplored in prior work deriving upper bounds: extracting non-uniform secrets (passwords and PII) from DP-finetuned language models (Section 4) and reconstructing tabular data from noisy marginals (Appendix F).

**Robustness to Prior Misspecification.** Since the true data distribution is rarely known in practice, we show that our bounds compose cleanly with upper confidence bounds on the adversary's baseline success rate, yielding certificates that hold with high probability even when the prior is only estimated (Section 3.5).

## 2 Background

In this section, we cover relevant background topics of DP and threat modeling terms. Informally, DP bounds the influence that any single user has on the output distribution of a mechanism. Throughout this paper, we assume the "replace-one" adjacency where exactly one record is replaced with another record.

**Definition 2.1** (Approximate DP [26])**.** *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies $(\varepsilon, \delta)$-DP if, for any two adjacent datasets $\mathbf{x}, \mathbf{x}' \in \mathcal{D}$ and $S \subseteq \mathcal{R}$, it holds:*

$$\Pr[\mathcal{M}(\mathbf{x}) \in S] \leq \mathrm{e}^{\varepsilon} \Pr[\mathcal{M}(\mathbf{x}') \in S] + \delta$$

There are many other variants of DP, e.g., the pure DP variant where $\delta = 0$ and the more recent Rényi DP ($(\alpha, \gamma)$-RDP) [50], $f$-DP [25], and $\mu$-GDP [25] variants, which characterize more complex relationships between the output distributions.

### 2.1 Threat Modeling

Several prior works have focused on systematizing and proposing a taxonomy of privacy attacks [19, 59, 60]. To categorize the privacy risks evaluated in Table 1, we adopt the taxonomy of Cummings et al. [19]. **Attack Goals** range from membership inference (MI) [62], to attribute inference (AI) [30], and full data reconstruction (Recon) [24]. **Prior Knowledge** dictates whether the adversary assumes uniform probabilities [46, 64] or exploits realistic, non-uniform data distributions [21, 29]. **Number of Targets** distinguishes between isolating a single target versus simultaneously attacking multiple records [4, 14–16, 21, 28]. **Success Metrics** denote whether the attacker must achieve an exact match (Simple) or if they succeed via approximate reconstruction distances (Complex) [8, 11, 31].

## 3 Our Privacy Leakage Framework

In this section, we bound the success of an adversary in a *generic attack game*. We begin by specifying the game and the setting it represents. Then in Theorem 3.2, we state our bound for pure DP mechanisms in this generic setting. After discussing multiple interpretations of the bound, we describe how to encode specific attack settings in the parameters of our framework. Next, we extend our bounds to approximate DP mechanisms and discuss the nuances that arise in this challenging setting. We then show that our bounds remain valid when the prior is only known approximately. Lastly, we compare our bounds with the current state-of-the-art bounds to show how our bounds are tighter and more flexible than prior work.

### 3.1 Generic Attack Game

The *generic attack game* in Algorithm 1 describes the family of attacks that our bounds cover. The game works as follows. First, $n$ targets $X$ are sampled from a given product distribution $\mathcal{D}$. The targets $X$ are fed to the mechanism $\mathcal{M}$, which produces an output $a$. The adversary $\mathcal{A}$ uses the mechanism output $a$, the data generating distribution $\mathcal{D}$, and the mechanism specification $\mathcal{M}$ to produce $k$ attack attempts $\mathbf{z}$. Lastly, the quality of the attack attempt $\mathbf{z}$ is measured against the true dataset $X$ according to some success metric $\mathcal{L}$.

**Algorithm 1** Generic Attack Game

**Input:** Mechanism $\mathcal{M}$, Adversary $\mathcal{A}$, Data distribution $\mathcal{D}$, Success metric $\mathcal{L}$
**Output:** Attack success measure

1: Sample targets $(X_1, \ldots, X_n) \sim \mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_n$
2: Compute mechanism output $a \leftarrow \mathcal{M}(X_1, \ldots, X_n)$
3: Attack attempt $(\mathbf{z}_1, \ldots, \mathbf{z}_k) = \mathcal{A}(a, \mathcal{D}, \mathcal{M})$
4: **return** Attack success $\mathcal{L}(X, \mathbf{z})$

We call $\mathcal{D}$ the adversary's *prior distribution* over the records, and the prior $\mathcal{D}_i$ for person $i \in [n]$ encodes the knowledge that the adversary has for that target person's data before seeing the mechanism output. More generally, our framework can be applied to whatever *privacy unit* the mechanism $\mathcal{M}$ is designed to protect, such as, individual records, users, or households. Crucially, $X$ does not necessarily represent full data records. Instead, $\mathcal{D}$ is defined over the unknown domain of the attack, whether that is membership bits, specific missing attributes, or full records. Our bounds handle arbitrary priors $\mathcal{D}_1, \ldots, \mathcal{D}_n$ for each target, as long as they are *independent*.

**On the assumption of independent privacy units** Analyzing semantic privacy guarantees under correlated data is a foundational challenge in practical differential privacy deployments and is not unique to our framework. While DP itself is a distribution-free property of the algorithm and does not strictly require independent records, semantic privacy guarantees degrade significantly under correlation if the sensitivity analysis assumes changing one record does not affect others [33]. Our framework directly inherits this property: if the underlying neighboring relations or privacy parameters explicitly account for known correlations upfront (e.g., via group privacy or correlated-neighbor definitions), our bounds apply directly. Conversely, when unmodeled correlations exist across privacy units, quantifying an individual's realized semantic risk requires explicitly modeling those dependencies—exactly as is required for evaluating realized privacy loss in any standard differential privacy mechanism.

The *attack success metric* $\mathcal{L}$ encodes the goal of the adversary. The metric $\mathcal{L} : \mathcal{X} \times \mathcal{Z} \to \{0, \ldots, n\}$ takes a fixed list of targets $X$ of size $n$ and a list of *guesses* or *attack attempts* $\mathbf{z}$ of size $k$, and reports the number of targets in $X$ that were successfully "attacked" by the adversary. Our bounds handle arbitrary success metrics, with the requirement that $\mathcal{L}$ is *decomposable*, which we define next.

**Definition 3.1** (Decomposable metric)**.** *Let $\mathcal{X}$ and $\mathcal{Z}$ be the data and attack domains, respectively. A metric $\mathcal{L} : \mathcal{X} \times \mathcal{Z} \to \{0, \ldots, n\}$ is* decomposable *into $\ell_1 \cdots \ell_n$ if*

$$\mathcal{L}(\mathbf{x}, \mathbf{z}) = \sum_{i \in [n]} \ell_i(\mathbf{x}_i, \mathbf{z}).$$

In Section 3.3 we discuss how to encode common attack goals—including MI, AI, and reconstruction—in the success metric. Although we do not explicitly state it, our generic attack game allows any number of fixed "non-target" records $\mathbf{x}_{fixed}$ to be input to the mechanism in addition to the sampled target records. We assume that the adversary has full knowledge of these records—and can even select them. As a result, the attack success is not measured with respect to the non-target records. We omit the fixed records in our attack game and bounds, but one can imagine the mechanism $\mathcal{M}$ takes these as a hard-coded input.

### 3.2 Bounding Pure DP Mechanisms

We now present our bound for $\varepsilon$-DP mechanisms. For a private dataset that is drawn from the prior distribution, and given any fixed mechanism output $a$, we present a high probability bound on the

attack success as a function of: (1) the mechanism's DP parameters, and (2) the adversary's success probability if they had only used prior information. Formally:

**Theorem 3.2.** *Let $\mathcal{M}$ satisfy $(\varepsilon, 0)$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an adversary and let $\mathcal{L}$ be decomposable into $\ell_1 \cdots \ell_n$. Then, for all $v \in \mathbb{R}$ and for all mechanism outputs $a \in Supp(\mathcal{M})$:*

$$\Pr_{X\sim\mathcal{D}}\left[\mathcal{L}(X, \mathcal{A}(a)) \geq v | \mathcal{M}(X) = a\right] \leq \Pr_{S_i \sim \mathsf{Bernoulli}(\beta_i(\mathbf{z},\varepsilon))}\left[\sum_{i\in[n]} S_i \geq v\right],$$

*where $\mathbf{z} = \mathcal{A}(a)$ and $\beta_i(\mathbf{z}, \varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X\sim\mathcal{D}_i}[\ell_i(X,\mathbf{z})=1]}}$*

To put this bound into context, consider a perfectly private $(0, 0)$-DP mechanism. The bound tells us that, for every fixed mechanism output $a$ and attacker output $\mathbf{z}$, the number of successfully attacked targets is distributed like a sum of Bernoullis with a flipping probabilities $\Pr_{X\sim\mathcal{D}_i}[\ell_i(X, \mathbf{z}) = 1]$ for $i = 1, \ldots, n$. As expected, this is exactly the adversary's prior success probability for $\mathbf{z}$ on a fresh sample of targets from the prior.

Now consider $(\varepsilon, 0)$-DP mechanisms for $\varepsilon > 0$. The bound tells us exactly how far the posterior success rate on each target can be from that of the perfectly private mechanism. Rather than a posterior of $\Pr_{X\sim\mathcal{D}_i}[\ell_i(X, \mathbf{z}) = 1]$, which would indicate the adversary gains no information from the mechanism output, the posterior essentially has an $e^\varepsilon$ multiplicative dependence.[6] For decomposable success metrics, we can sum the posterior probabilities to get the resulting bound over the total number of successfully attacked targets.

The dependence on the prior attack success probability captures the phenomenon that for natural (non-uniform) data distributions, some targets are easier to "attack" using only prior knowledge, and some success measures are more permissive than others (for example approximate rather than exact reconstruction measures). Furthermore, it captures the tradeoff between the *inherent difficulty* of the attack and the protections guaranteed by DP–something which previously was only stated in vague terms and not quantified clearly.[7] In Proposition A.2 we show Theorem 3.2 is tight for distributions that are sufficiently close to uniform. The analysis of Keinan et al. [39] on which types of DP mechanisms yield tight auditing bounds in this "one run" framework, could likely be extended to our more general bounds, although the focus of our work is not auditing.

There are a number of ways to view and use this bound. As in the *privacy auditing* literature [5, 23, 35, 64], one may interpret the bound as a hypothesis test to detect violations of the DP guarantee for a particular mechanism that indicate it was not implemented correctly. If a real attack is more successful than $v$ with high probability then one can confidently refute the hypothesis that the mechanism is differentially private with the claimed privacy parameters. In this work, we are more interested in characterizing the privacy leakage of faithfully implemented DP mechanisms. To that end, we measure $v$ for different probabilities and compare that to the success of empirical attacks.

### 3.3 Encoding Attacks

Our bound is flexible enough to encode most attack settings from the taxonomy developed by Cummings et al. [19].

**Attack targets and goals.** The target distribution $\mathcal{D}$ represents the adversary's prior over the unknown part of each data record. We can encode MI by setting $X_i \sim \mathsf{Bernoulli}(0.5)$ to denote the membership bit of record $i \in [n]$ (as in [64]). The fixed data records in the MI game are encoded in $\mathcal{M}$ and are thus known to the adversary. For full reconstruction, $\mathcal{D}_i$ becomes the data-generating distribution for person $i$, while AI interpolates these by conditioning $\mathcal{D}_i$ on known attributes.

[6] This $e^\varepsilon$ is the dependence used in Reconstruction Robustness, though the actual $\beta$ in Theorem 3.2 is significantly tighter for large priors and large epsilons.

[7] For example, when discussing why reconstruction attacks fail against DP mechanisms with large $\varepsilon$s [57] states "There is a natural intuition for why larger $\varepsilon$s provide effective protection in these works–the attacks generally consider an adversary attempting to answer a high-dimensional question (e.g., reconstructing a full training example) with only limited information about the dataset (e.g., distributional)." We can now formalize this intuition.

**Success metrics.** Our attack success metric captures exact match $\ell_i(X_i, \mathbf{z}) = \mathbb{I}\{X_i = \mathbf{z}_i\}$ or complex individual-level approximate reconstruction by thresholding distance $\ell_i(\mathbf{x}_i, \mathbf{z}) = \mathbb{I}\{(\min_j ||\mathbf{x}_i - \mathbf{z}_j||_2) \leq \tau\}$ as done in prior work [8, 31] .

**Prior knowledge.** In Theorem 3.2, the target generation distribution $\mathcal{D}_i$ for each person $i \in [n]$ can be set *arbitrarily*, as long as it is independent from others' target distributions. The bound itself only depends on the prior probability of success for a particular attack attempt $\mathbf{z}$: $\Pr_{X\sim\mathcal{D}_i}[\ell_i(X, \mathbf{z}) = 1]$, that is, the probability that a freshly sampled record $X \sim \mathcal{D}_i$ would be "close to" the attack attempt $\mathbf{z}$. This probability need not be known exactly: by Corollary 3.4, any upper confidence bound on it suffices.

### 3.4 Bounding $f$-DP

Generalizing our pure-DP to approximate notions of DP comes with a number of challenges. Approximate DP mechanisms suffer from potentially unbounded privacy loss, so incorporating the mechanism output requires more care. We apply the proof techniques from Steinke et al. [64] and extend their bounds in two key ways: by allowing for general success metrics and arbitrary data distributions. We state and prove approx-DP bounds in Appendix B, which we further generalize to $f$-DP in Appendix C by invoking the primal-dual perspective of $f$-DP [25].

We begin by stating an $f$-DP generalization of our pure DP bound which incurs an additive factor of $n \cdot \delta_f(\varepsilon)$ in expectation, where $\delta_f(\varepsilon) = 1 + f^*(-e^\varepsilon)$ and $f^*$ is the convex conjugate of $f$. Here, we trade the simplicity of the bound with the computational overhead of having to estimate the probability over multiple runs of the mechanism.

**Theorem 3.3.** *Let $\mathcal{M}$ satisfy $f$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an attacker and let $\mathcal{L}$ be decomposable into $\ell_1 \cdots \ell_n$. For all $\varepsilon > 0$, $v \in \mathbb{R}$:*

$$\Pr_{\substack{X\sim\mathcal{D}\\ a\sim\mathcal{M}(X)}}[\mathcal{L}(X, \mathcal{A}(a)) \geq v] \leq \Pr_{\substack{X\sim\mathcal{D}\\ a\sim\mathcal{M}(X)\\ S_i\sim\mathsf{Bernoulli}(\beta_i(\mathcal{A}(a),\varepsilon))}}\left[\sum_{i\in[n]} S_i \geq v\right] + n \cdot \delta_f(\varepsilon),$$

*where* $\beta_i(\mathcal{A}(a), \varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X\sim\mathcal{D}_i}[\ell_i(X,\mathcal{A}(a))=1]}}$.

The approx-DP analog of this bound is Theorem B.2. Note that since $\delta_f$ is typically much smaller than $1/n$, the additive term brought on by $f$-DP is fairly small, however the flipping probabilities $\beta_1, \ldots, \beta_n$ are now random variables, which must be estimated from multiple runs of the mechanism on fresh data samples. In Appendix C, we provide a version of this bound which uses the worst-case prior and avoids a dependence on the mechanism output on the RHS.

### 3.5 Robustness to Prior Misspecification

Our bounds depend on the prior $\mathcal{D}$ only through the baseline success probabilities $p_i(\mathbf{z}) := \Pr_{X\sim\mathcal{D}_i}[\ell_i(X, \mathbf{z}) = 1]$, and they are monotone in these quantities, since $\beta(p, \varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + 1/p}$ is increasing in $p$. A practitioner who cannot specify $\mathcal{D}$ exactly therefore only needs *upper* bounds on these probabilities. For example, these can be confidence bounds estimated from auxiliary data.

**Corollary 3.4.** *Let $\mathcal{M}$, $\mathcal{D}$, $\mathcal{A}$, and $\mathcal{L}$ be as in Theorem 3.2. Let $\hat{p}_1, \ldots, \hat{p}_n$ be estimates computed independently of $X$ and of the randomness of $\mathcal{M}$, such that for some $\eta \in (0, 1)$,*

$$\Pr_{\hat{p}}\left[\hat{p}_i(\mathbf{z}) \geq p_i(\mathbf{z}) \;\; \forall i \in [n], \; \forall \mathbf{z}\right] \geq 1 - \eta.$$

*Then, with probability at least $1 - \eta$ over $\hat{p}$, for all $v \in \mathbb{R}$ and all $a \in Supp(\mathcal{M})$,*

$$\Pr_{X\sim\mathcal{D}}\left[\mathcal{L}(X, \mathcal{A}(a)) \geq v \mid \mathcal{M}(X) = a\right] \leq \Pr_{\hat{S}_i\sim\mathsf{Bernoulli}(\beta(\hat{p}_i(\mathbf{z}),\varepsilon))}\left[\textstyle\sum_{i\in[n]} \hat{S}_i \geq v\right], \qquad \mathbf{z} = \mathcal{A}(a).$$

*The same substitution $p_i \mapsto \hat{p}_i$ preserves Theorem 3.3 and Theorems B.1 to B.3, C.3 and C.4. For Theorems B.3 and C.4, it suffices to upper bound the $n$ scalars $p_i^* = \max_{z_i} p_i(z_i)$, and $\hat{p}_i^*$ replaces $p_i^*$ in both $\beta_i$ and $\alpha$.*

In other words, overestimating how easy a target is to guess a priori can only make our bounds more conservative. A practitioner who is unsure of the prior can therefore err on the side of caution, and the resulting certificate remains valid with high probability (proof in Appendix D). The cost of this caution depends on the privacy level. For large $\varepsilon$, the bound is most sensitive to errors in small priors, so conservative estimates loosen the bound more in that regime.

### 3.6 Numerical Comparisons With Prior Work

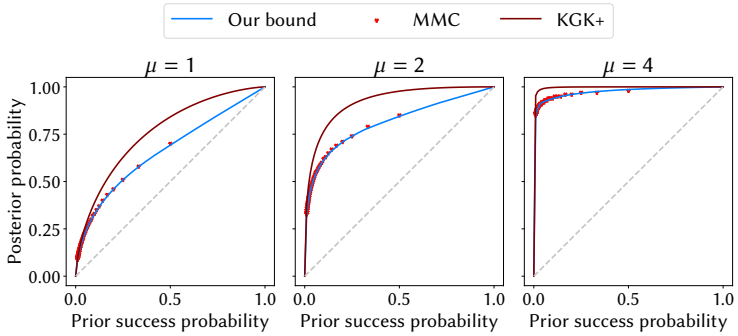



Figure 1: Compare our bounds with prior state-of-the-art bounds (KGK+ [42] and MMC [46]) when attacking a single record with varying prior at different privacy levels for $\mu$-GDP.

We run numerical experiments to showcase the versatility of our bound and compare it with the state-of-the-art bounds. In order to compare the bounds in a common attack setting, we consider the single target record ($n = 1$) and single attempt ($k = 1$) setting. To that end, we report our upper bounds on the attack success probability from Theorem C.4 for varying prior probabilities in Figure 1 and compare this to the current state-of-the-art bounds KGK+ [42] and MMC [46].

For all privacy levels, our work is substantially tighter than the current state-of-the-art bounds that supports non-uniform priors, KGK+. For uniform priors, our bounds closely align with the MMC bound. Unlike MMC, our bounds are continuous and can be extended to any given prior success probability between 0 and 1.

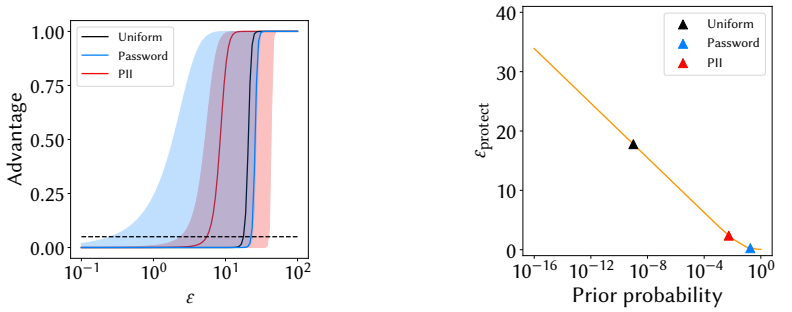
## 4 Bounding Data Extraction from LLMs

We now leverage the versatility of our bound to present novel reconstruction bounds for attack settings involving non-uniform priors and multiple target samples. Specifically, we focus on data extraction attacks in the context of LLMs [10, 11, 45, 53]. Although prior work has shown some success in effectively training LLMs on private data with DP [43, 63, 69], sensitive data can still be extracted from these models [45]. The main reason for this is that data can often be inferred from the context present in the sample [45], similar data appears elsewhere in the dataset [44], or the data itself has low entropy (e.g., 123456789 is a very common password). In other words, the *prior probability* of extracting data from language models is often non-uniform. Previous bounds [31, 46] either supported single target canaries or uniform priors, but not both simultaneously.

To that end, we focus on three types of data distributions—**uniform random canary**, **numerical password**, and **PII**—and present methods to estimate their prior distributions. Subsequently, use the Secret Sharer [10] setup and DP-finetune GPT-2 on canaries drawn from these distributions, try to extract the canaries, and compare our theoretical bounds to concrete attacks. All our experiments are run on a single A100 GPU. Our experiments on tabular data can be found in Appendix F.

### 4.1 Estimating Priors

**Uniform random canary.** Uniformly chosen canaries are most commonly used when evaluating data extraction attacks, as the explicit prior distribution makes it easy to determine when the attack is non-trivial. Specifically, we derive bounds for the 9-digit uniform random canary from prior work [10] and set the prior probability to $10^{-9}$ for all possible canaries.




(a) Adv bounds for different prior distributions and privacy levels $\varepsilon$

(b) $\varepsilon_{\text{protect}}$ that bounds Adv $\leq 0.05$ for varying prior probabilities

Figure 2: Bounding advantage of data extraction for different prior distributions at fixed $\delta = 10^{-5}$. NB: Dotted line represents Adv $= 0.05$ and markers correspond to canary with highest prior probability for each distribution.

**Numerical password.** Data extraction attacks have also been used to extract sensitive data that appears *naturally* in the dataset such as credit card and social security numbers [10]. Along this line, we consider 9-digit numerical passwords as they follow a similar format to uniformly chosen canaries. Following prior work in similar domains [27, 29, 67], we model the prior distribution with a Zipf's law distribution under the assumption that a few passwords are exceedingly popular amongst users, e.g., the password '123456789' is known to be used roughly 43M times in the wild [34].

**PII.** Lastly, we consider the extraction of Personally Identifiable Information (PII) [11, 45, 53]. However, unlike uniformly chosen canaries and numerical passwords, PIIs do not have a fixed format or a well-known distribution, and can possibly depend on the surrounding text [45].

Therefore, we use a pre-trained LLM to estimate the prior over PIIs. Following recent work [45] we focus on reconstructing names in the Enron emails dataset [40] given the surrounding text. We first sample 500 sequences from the Enron emails dataset and extract a set of candidate PIIs that occur in these sequences. Next, we calculate the perplexity of the pre-trained GPT-2 model [58] for each sequence and candidate PII conditioned on the surrounding text. Lastly, we normalize the perplexities over the set of candidate PIIs to compute the prior distribution of PIIs given the surrounding text.

### 4.2 Reconstructing Data from an LLM

To contextualize our bounds with respect to a concrete attack, we fine-tune a GPT-2 model with DP-SGD [1] on a dataset consisting of canaries drawn from different distributions and task the adversary with reconstructing the canaries. We emphasize that the aim of this section is not to perform a novel and realistic privacy attack, but rather study how well our bounds model concrete attacks against real mechanisms and therefore take a DP auditing approach [55].

Following prior work [55], we construct canaries using new tokens, which has been shown to be essential in achieving meaningful attack success against DP fine-tuned LLMs on uniform canaries. Specifically, each canary consists of two newly created tokens: the prefix, which is fixed to the ID of the canary (<1>, <2>, etc.) and the secret, which is sampled from 10 newly created tokens for each canary. For simplicity, we scale down the prior distributions from their original domains to being over 10 possible secret tokens.

To reconstruct the canary, we instantiate the "Model Attack" on the finetuned model from prior work [31]. We assume that the adversary has access to only the final trained model, the prior distribution over the secret tokens, and the prefix token. The adversary then reconstructs the secret token by choosing the token with the highest probability assigned by the model conditioned on the prefix, combined with the prior.

**Idealized Gaussian Attack.** While prior "one-run" auditing frameworks have evaluated empirical membership inference attacks against real models [46, 64], extending these audits to data extraction with non-uniform priors introduces a new confounding variable: the practical data extraction attack

itself may be highly suboptimal. To isolate the tightness of our theoretical bound from the empirical limitations of the black-box model attack, we also evaluate an idealized Gaussian mechanism. We sample one-hot canary vectors from our priors, add Gaussian noise, and execute a Bayes optimal attack. At a high level, this optimal adversary evaluates every candidate canary by combining the observed noisy vector with the known prior distribution, and guesses the canary with the highest posterior probability (Appendix E). Because the Gaussian mechanism forms the basis of DP-SGD, and its Bayes optimal adversary is mathematically tractable, it provides a ceiling for attack success. Comparing our bound directly against this Bayes optimal adversary tells us exactly how much slack exists in our bound, independent of the empirical limitations of the Model Attack.

### 4.3 Results

In Figure 2a, we plot our bounds on the generalized advantage (Adv) [13] when reconstructing different secrets from the three prior distributions considered. Since the advantage is a measure of an attack's *significance* and not just its success, it allows us to compare the bounds more effectively between canaries with different priors. Using Theorem B.3, we compute posterior bounds for canaries with minimum[8], maximum, and median prior probabilities. We convert these to generalized advantage bounds using the following equation [13]: $\text{Adv} = (\text{Posterior} - \text{Prior})/(1 - \text{Prior})$.

Additionally, in Figure 2b we vary the prior probability and plot the corresponding $\varepsilon_{\text{protect}}$ that would result in an Adv $\leq 0.05$ at fixed $\delta = 10^{-5}$. Informally, Figure 2b plots the privacy level required to protect data with a given prior probability from leaking.[9]

**Non-Uniform Priors Drastically Increase Risk.** While it is intuitive that common secrets are easier to guess, our framework explicitly quantifies how non-uniform priors result in disparate privacy leakage across individuals. Unlike uniform canaries, which share identical advantage bounds, canaries with higher prior probabilities admit attacks with substantially higher adversarial advantage. For example, at $\varepsilon = 1$, the advantage bound for extracting the median numerical password is $\ll 0.05$, but jumps to $0.24$ for the most probable password. Similarly, at $\varepsilon = 10$, the maximum advantage for any uniform canary is negligible, whereas the most probable PII and password canaries reach an advantage bound of $\approx 1.0$. Ultimately, evaluating mechanisms solely on uniform canaries drastically underestimates the theoretical privacy leakage for individuals holding highly probable secrets.

**Calibrating Epsilon to Reconstruction Risks.** A natural research question that arises from our results is what would be an appropriate privacy level that protects against the different types of data extraction. To answer this, for each prior distribution, we compute the $\varepsilon_{\text{protect}}$ value that bounds the advantage for the worst-case canary to $\leq 0.05$ and plot it in Figure 2b. Subsequently, we note that in general as canaries become less probable, they become easier to protect even at low privacy levels e.g., a uniformly drawn 9 digit canary can be protected by $\varepsilon_{\text{protect}} = 17.8$ but numerical passwords and PIIs require $\varepsilon_{\text{protect}} = 0.25$ and $\varepsilon_{\text{protect}} = 2.37$, respectively. This shows that, even though a large $\varepsilon$ has in the past proven effective at preventing the leakage of uniform canaries [10], naturally occurring canaries might require a substantially smaller $\varepsilon$ to be protected. Furthermore, this also illustrates how practitioners can use our bounds in practice to calibrate the privacy parameters to specific privacy risks. While Figure 2b considers the most vulnerable single canary, the same procedure applies to multiple targets: one computes the bound on the number of successfully attacked targets (Theorem B.3) and selects the largest $\varepsilon$ that keeps it below the desired threshold.

**The Gap Between Theoretical and Concrete Attacks.** In Figure 3 we compare our theoretical bounds (Theorem C.4) against concrete reconstruction accuracies of the model attack on GPT-2 fine-tuned on a 100 canaries. We perform hyper-parameter tuning to identify the best clipping norm $[1.0, 10.0, 100.0]$, learning rate $[0.01, 0.1, 1, 10]$, and number of epochs $[1, 3, 10, 32, 100]$ for each privacy level studied. By overlaying the Idealized Gaussian attack, Figure 3 clearly isolates two distinct gaps between our theoretical bound and the concrete model attack.

First, comparing our bound to the Bayes optimal baseline reveals the mathematical slack in our framework. For uniform canaries, our bounds align almost perfectly with the idealized attack.

[8] When estimating the prior distribution for PIIs, we note that the minimum probability can sometimes be prohibitively small ($\approx 0$) and hence we plot the $10^{\text{th}}$ percentile instead.

[9] Here, $0.05$ is chosen arbitrarily, but depending on the context the advantage threshold considered can be made more stringent or loose.

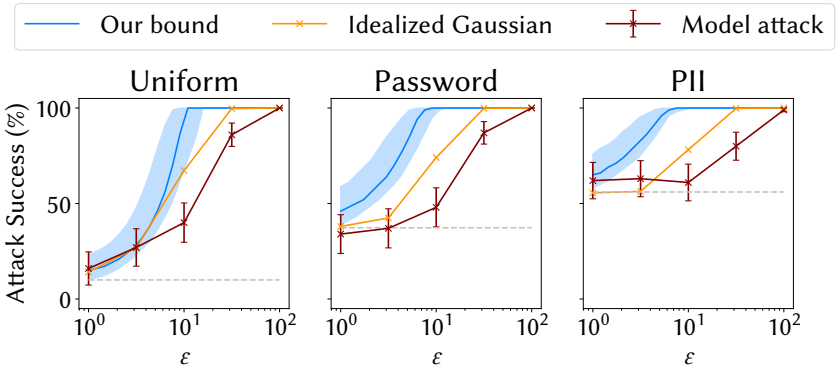



Figure 3: Compare our bounds with success of reconstructing 100 canaries (with 95% CI) from a DP fine-tuned GPT-2 model at different privacy levels and fixed $\delta = 10^{-5}$. NB: Dotted gray line represents baseline attack w/o access to the DP mechanism.

However, for non-uniform priors (passwords and PII), a gap persists between the bound and the optimal attack. This indicates that while our bounds are tight for uniform distributions, there is still room for theoretical refinement to tighten the bounds for mechanisms handling highly skewed priors.[10]

Second, comparing the Idealized Gaussian to the Model Attack reveals that the model attack falls significantly short of the Bayes optimal baseline, particularly in the medium-privacy regime ($\varepsilon = 10$) and for non-uniform priors. On the other hand, it is close to the Idealized Gaussian for high privacy ($\varepsilon = 1.0$) and vacuous regimes ($\varepsilon = 100.0$). This empirical gap likely stems from documented limitations of black box auditing of DP-SGD, like the fact that the bound and idealized attack capture total privacy loss across all iterations, but the black-box model attack only observes the final model weights [3, 12, 54]. The attack would miss data that the model memorized early on and "forgot" during later epochs.

## 5 Conclusion

In this paper, we introduced a novel framework that derives precise privacy leakage bounds for DP mechanisms under previously unexplored, realistic threat models—specifically those involving multiple targets, non-uniform prior distributions, and complex reconstruction metrics. Our framework equips practitioners with the theoretical foundation needed to interpret DP parameters and calibrate them against context-specific privacy risks rather than relying on worst-case heuristics.

**Limitations and Future Work.** While our bounds currently rely on decomposable success metrics, developing bounds for, non-decomposable success metrics remains an open challenge that would capture an even broader set of adversarial goals. Additionally, we encourage the privacy community to design and study more empirical DP attacks that explicitly incorporate non-uniform priors, as our results show these distributions dramatically alter the true risk landscape and can be harder to execute. Finally, Corollary 3.4 shows that our bounds remain valid under upper confidence bounds on the prior success probabilities. However, obtaining *tight* such bounds for complex, high-dimensional distributions such as natural language remains challenging, and loose estimates propagate directly to loose bounds.

**Broader Impacts.** Our framework equips practitioners to calibrate privacy parameters to context-specific threat models. The primary risk of this framework is the potential for false confidence or policy misuse. Because our bounds can demonstrate that empirical risks are lower than worst-case MI guarantees, practitioners might be tempted to justify weaker privacy protections. These bounds rely on strict structural assumptions (which may not hold for all use cases) and should not be used as a blanket reason to degrade privacy safeguards.

[10] It's not clear we could achieve a tighter bound without losing the generality of our framework, but perhaps one could derive a tighter bespoke bound for the Gaussian mechanism.

## Acknowledgments and Disclosure of Funding

M.S.M.S.A. was supported by a National Science Scholarship (PhD) from the Agency for Science Technology and Research, Singapore (A*STAR). A.S.'s work at BU was partially supported by NSF award CNS-2232694. The authors would also like to thank Bogdan Kulynych for helpful discussions.

## References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep Learning with Differential Privacy. In *CCS*, 2016.

[2] J. M. Abowd. The U.S. Census Bureau Adopts Differential Privacy. In *KDD*, 2018.

[3] M. S. M. S. Annamalai and E. De Cristofaro. Nearly Tight Black-Box Auditing of Differentially Private Machine Learning. In *NeurIPS*, 2024.

[4] M. S. M. S. Annamalai, A. Gadotti, and L. Rocher. A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data. In *USENIX Security*, 2024.

[5] M. S. M. S. Annamalai, B. Balle, J. Hayes, G. Kaissis, and E. De Cristofaro. The Hitchhiker's Guide to Efficient, End-to-End, and Tight DP Auditing. *arXiv:2506.16666*, 2025.

[6] Apple Differential Privacy Team. Learning with privacy at scale. https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf, 2017.

[7] S. Aydore, W. Brown, M. Kearns, K. Kenthapadi, L. Melis, A. Roth, and A. A. Siva. Differentially Private Query Release Through Adaptive Projection. In *ICML*, 2021.

[8] B. Balle, G. Cherubin, and J. Hayes. Reconstructing Training Data with Informed Adversaries. In *IEEE S&P*, 2022.

[9] K. Cai, X. Lei, J. Wei, and X. Xiao. Data Synthesis via Differentially Private Markov Random Fields. *VLDB*, 2021.

[10] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *USENIX Security*, 2019.

[11] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting Training Data from Large Language Models. In *USENIX Security*, 2021.

[12] T. Cebere, A. Bellet, and N. Papernot. Tighter Privacy Auditing of DP-SGD in the Hidden State Threat Model. In *ICLR*, 2025.

[13] G. Cherubin. Bayes, not Naïve: Security Bounds on Website Fingerprinting Defenses. *PETS*, 2017.

[14] G. Cherubin, B. Köpf, A. Paverd, S. Tople, L. Wutschitz, and S. Zanella-Béguelin. Closed-Form Bounds for DP-SGD against Record-level Inference Attacks. In *USENIX Security*, 2024.

[15] A. Cohen. Attacks on Deidentification's Defenses. In *USENIX*, 2022.

[16] A. Cohen and K. Nissim. Linear Program Reconstruction in Practice. *Journal of Privacy and Confidentiality*, 2020.

[17] A. Cohen and K. Nissim. Towards formalizing the GDPR's notion of singling out. *PNAS*, 2020.

[18] E. Cohen, H. Kaplan, Y. Mansour, S. Moran, K. Nissim, U. Stemmer, and E. Tsfadia. Data Reconstruction: When You See It and When You Don't. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2025.

[19] R. Cummings, S. Hod, J. Sarathy, and M. Swanberg. ATTAXONOMY: Unpacking Differential Privacy Guarantees Against Practical Adversaries. *arXiv:2405.01716*, 2024.

[20] DataSF. Fire Department Calls for Service. https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3, 2020.

[21] T. Dick, C. Dwork, M. Kearns, T. Liu, A. Roth, G. Vietri, and Z. S. Wu. Confidence-ranked reconstruction of census microdata from published statistics. *PNAS*, 2023.

[22] W. Diffie and M. E. Hellman. New Directions in Cryptography. In *Democratizing Cryptography: The Work of Whitfield Diffie and Martin Hellman*. 2022.

[23] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer. Detecting Violations of Differential Privacy. In *CCS*, 2018.

[24] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[25] J. Dong, A. Roth, and W. J. Su. Gaussian Differential Privacy. *arXiv:1905.02383*, 2019.

[26] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, 2006.

[27] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *CCS*, 2014.

[28] A. Gadotti, F. Houssiau, L. Rocher, B. Livshits, and Y.-A. de Montjoye. When the signal is in the noise: Exploiting Diffix's Sticky Noise. In *USENIX*, 2019.

[29] A. Gadotti, F. Houssiau, M. S. M. S. Annamalai, and Y.-A. de Montjoye. Pool Inference Attacks on Local Differential Privacy: Quantifying the Privacy Guarantees of Apple's Count Mean Sketch in Practice. In *USENIX Security*, 2022.

[30] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In *CCS*, 2018.

[31] J. Hayes, B. Balle, and S. Mahloujifar. Bounding Training Data Reconstruction in DP-SGD. *NeurIPS*, 2023.

[32] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genetics*, 2008.

[33] T. Humphries, S. Oya, L. Tulloch, M. Rafuse, I. Goldberg, U. Hengartner, and F. Kerschbaum. Investigating membership inference attacks under data dependencies, 2023. URL https://arxiv.org/abs/2010.12112.

[34] T. Hunt, C. Hunt, and S. J. Sigurðarson. Have I been pwned. https://haveibeenpwned.com/, 2025.

[35] M. Jagielski, J. Ullman, and A. Oprea. Auditing Differentially Private Machine Learning: How Private is Private SGD? *NeurIPS*, 2020.

[36] B. Jayaraman and D. Evans. Are Attribute Inference Attacks Just Imputation? In *CCS*, 2022.

[37] P. Kairouz, S. Oh, and P. Viswanath. The Composition Theorem for Differential Privacy. In *ICML*, 2015.

[38] S. P. Kasiviswanathan, M. Rudelson, A. Smith, and J. Ullman. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In *STOC*, 2010.

[39] A. Keinan, M. Shenfeld, and K. Ligett. How Well Can Differential Privacy Be Audited in One Run? *arXiv:2503.07199*, 2025.

[40] B. Klimt and Y. Yang. Introducing the Enron Corpus. In *CEAS*, 2004.

[41] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

[42] B. Kulynych, J. F. Gomez, G. Kaissis, J. Hayes, B. Balle, F. Calmon, and J. L. Raisaro. Unifying Re-Identification, Attribute Inference, and Data Reconstruction Risks in Differential Privacy. In *NeurIPS*, 2025.

[43] X. Li, F. Tramer, P. Liang, and T. Hashimoto. Large Language Models Can Be Strong Differentially Private Learners. In *ICLR*, 2022.

[44] K. Z. Liu, C. A. Choquette-Choo, M. Jagielski, P. Kairouz, S. Koyejo, P. Liang, and N. Papernot. Language Models May Verbatim Complete TextThey Were Not Explicitly Trained On. *arXiv:2503.17514*, 2025.

[45] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin. Analyzing Leakage of Personally Identifiable Information in Language Models. In *IEEE S&P*, 2023.

[46] S. Mahloujifar, L. Melis, and K. Chaudhuri. Auditing $f$-Differential Privacy in One Run. In *ICML*, 2025.

[47] R. McKenna, G. Miklau, and D. Sheldon. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *JPC*, 2021.

[48] R. McKenna, B. Mullins, D. Sheldon, and G. Miklau. AIM: An Adaptive and Iterative Mechanism for Differentially Private Synthetic Data. *VLDB Endowment*, 2022.

[49] B. McMahan and A. Thakurta. Federated Learning with Formal Differential Privacy Guarantees. https://ai.googleblog.com/2022/02/federated-learning-with-formal.html, 2022.

[50] I. Mironov. Rényi Differential Privacy. In *CSF*, 2017.

[51] Multiple Authors. Differential Privacy. https://github.com/google/differential-privacy, 2020.

[52] M. Nasr, J. Hayes, T. Steinke, B. Balle, F. Tramèr, M. Jagielski, N. Carlini, and A. Terzis. Tight Auditing of Differentially Private Machine Learning. In *USENIX Security*, 2023.

[53] M. Nasr, J. Rando, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, F. Tramèr, and K. Lee. Scalable Extraction of Training Data from Aligned, Production Language Models. In *ICLR*, 2025.

[54] M. Nasr, T. Steinke, A. Ganesh, B. Balle, C. A. Choquette-Choo, M. Jagielski, J. Hayes, A. G. Thakurta, A. Smith, and A. Terzis. The Last Iterate Advantage: Empirical Auditing and Principled Heuristic Analysis of Differentially Private SGD. In *ICLR*, 2025.

[55] A. Panda, X. Tang, C. A. Choquette-Choo, M. Nasr, and P. Mittal. Privacy Auditing of Large Language Models. In *ICLR*, 2025.

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

[57] N. Ponomareva, H. Hazimeh, A. Kurakin, Z. Xu, C. Denison, H. B. McMahan, S. Vassilvitskii, S. Chien, and A. G. Thakurta. How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy. *Journal of Artificial Intelligence Research*, 2023.

[58] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[59] M. Rigaki and S. Garcia. A Survey of Privacy Attacks in Machine Learning. *ACM Computing Surveys*, 2023.

[60] A. Salem, G. Cherubin, D. Evans, B. Köpf, A. Paverd, A. Suri, S. Tople, and S. Zanella-Béguelin. SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning. In *IEEE S&P*, 2023.

[61] Shlomi Hod and Ran Canetti. Differentially Private Release of Israel's National Registry of Live Births. In *IEEE S&P*, 2025.

[62] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership Inference Attacks against Machine Learning Models. In *IEEE S&P*, 2017.

[63] A. Sinha, T. Mesnard, R. McKenna, D. Liu, C. A. Choquette-Choo, Y. Huang, D. Yu, G. Kaissis, Z. Charles, R. Liu, et al. VaultGemma: A Differentially Private Gemma Model. *arXiv:2510.15001*, 2025.

[64] T. Steinke, M. Nasr, and M. Jagielski. Privacy Auditing with One (1) Training Run. *NeurIPS*, 2024.

[65] P. Stock, I. Shilov, I. Mironov, and A. Sablayrolles. Defending against Reconstruction Attacks with Rényi Differential Privacy. *arXiv:2202.07623*, 2022.

[66] US Census Bureau. American Community Survey. https://www.census.gov/programs-surveys/acs.html, 2025.

[67] T. Wang, J. Blocki, N. Li, and S. Jha. Locally Differentially Private Protocols for Frequency Estimation. In *USENIX Security*, 2017.

[68] A. C. Yao. Theory and AppDcations of Trapdoor Functions. In *Symposium on Foundations of Computer Science (SFCS)*, 1982.

[69] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, A. Manoel, L. Wutschitz, S. Yekhanin, and H. Zhang. Differentially Private Fine-tuning of Language Models. In *ICLR*, 2022.

# A Pure DP Proofs

Recall the definition of stochastic dominance.

**Definition A.1** (Stochastic Dominance)**.** *Let $X, Y \in \mathbb{R}$ be random variables. We say $X$ is* stochastically dominated *by $Y$ if for all $t \in \mathbb{R}$, $\Pr(X \geq t) \leq \Pr(Y \geq t)$.*

**Theorem 3.2.** *Let $\mathcal{M}$ satisfy $(\varepsilon, 0)$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an adversary and let $\mathcal{L}$ be decomposable into $\ell_1 \cdots \ell_n$. Then, for all $v \in \mathbb{R}$ and for all mechanism outputs $a \in Supp(\mathcal{M})$:*

$$\Pr_{X \sim \mathcal{D}}\left[\mathcal{L}(X, \mathcal{A}(a)) \geq v | \mathcal{M}(X) = a\right] \leq \Pr_{S_i \sim \mathsf{Bernoulli}(\beta_i(\mathbf{z}, \varepsilon))}\left[\sum_{i \in [n]} S_i \geq v\right],$$

*where $\mathbf{z} = \mathcal{A}(a)$ and $\beta_i(\mathbf{z}, \varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X \sim \mathcal{D}_i}[\ell_i(X, \mathbf{z}) = 1]}}$*

*Proof.* Fix a product distribution distribution $\mathcal{D} = \mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_n$, a $(\varepsilon, 0)$-DP mechanism $\mathcal{M}$, a metric $\mathcal{L}$ that decomposes into $\ell_1 \cdots \ell_n$, and an attacker $\mathcal{A}$, and fix any mechanism output $a$. We can assume WLOG that the attacker $\mathcal{A}$ is deterministic. Let $\mathbf{z} = \mathcal{A}(a)$ be the attack attempts for our fixed mechanism output $a$.

Following the techniques of Steinke, Nasr, Jagielski, we will prove the upper bound by induction over data records. Fix some $i \in [n]$ and fix $x_{<i} = (x_1, \ldots, x_{i-1}) \in Supp(\mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_{i-1})$. Then, applying Bayes' theorem and the $\varepsilon$-DP guarantee:

$$\begin{aligned}
&\Pr[\ell_i(X_i) = 1 | \mathcal{M}(X) = a, X_{<i} = x_{<i}] \\
&= \frac{\Pr[\mathcal{M}(X) = a | \ell_i(X_i) = 1, X_{<i} = x_{<i}] \cdot \Pr[\ell_i(X_i) = 1 | X_{<i} = x_{<i}]}{\Pr[\mathcal{M}(X) = a | X_{<i} = x_{<i}]} \\
&= \frac{\Pr[\mathcal{M}(X) = a | \ell_i(X_i) = 1, X_{<i} = x_{<i}] \cdot \Pr[\ell_i(X_i) = 1]}{\sum_{b \in \{0,1\}} \Pr[\mathcal{M}(X) = a | \ell(X_i, \mathbf{z}) = b, X_{<i} = x_{<i}] \cdot \Pr[\ell_i(X_i) = b]} \\
&= \frac{1}{1 + \frac{\Pr[\mathcal{M}(X) = a | \ell_i(X_i) = 0, X_{<i} = x_{<i}] \cdot \Pr[\ell_i(X_i) = 0]}{\Pr[\mathcal{M}(X) = a | \ell_i(X_i) = 1, X_{<i} = x_{<i}] \cdot \Pr[\ell(X_i, \mathbf{z}) = 1]}} \\
&\leq \frac{1}{1 + e^{-\varepsilon} \cdot \frac{\Pr[\ell_i(X_i) = 0]}{\Pr[\ell_i(X_i) = 1]}} \\
&= \frac{1}{1 + e^{-\varepsilon} \cdot \frac{1 - \Pr[\ell_i(X_i) = 1]}{\Pr[\ell_i(X_i) = 1]}} \\
&= \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr[\ell_i(X_i) = 1]}}.
\end{aligned}$$

By a similar argument,

$$\begin{aligned}
&\Pr[\ell_i(X_i) = 1 | \mathcal{M}(X) = a, X_{<i} = x_{<i}] \\
&\in \left[\frac{e^{-\varepsilon}}{e^{-\varepsilon} - 1 + \frac{1}{\Pr[\ell_i(X_i) = 1]}}, \ \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr[\ell_i(X_i) = 1]}}\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Pr[\ell_i(X_i, \mathbf{z}) = 1 | \mathcal{M}(X) = a, X_{<i} = x_{<i}] \\
&\leq \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr[\ell_i(X_i) = 1]}} := \beta_i(\mathbf{z}, \varepsilon),
\end{aligned}$$

Note that $\beta_i(\mathbf{z}, \varepsilon) \in [0, 1]$ for all $\varepsilon$ and all $\mathbf{z} \in Supp(\mathcal{D})$. Now we will proceed with induction. Let $W_{i-1} = \sum_{m \in [i-1]} \ell_m(X_m, \mathbf{z})$ be the loss metric the first $i - 1$ data records. Assume inductively that $W_{i-1}$ is stochastically dominated by $\hat{W}_{i-1} = \sum_{m \in [i-1]} S_m$ where

$S_m \sim \mathsf{Bernoulli}(\beta_i(\mathbf{z}, \varepsilon))$. We have that, conditioned on $W_{i-1}$ the variable $\ell_i(X_i, \mathbf{z})$ is stochastically dominated by $\mathsf{Bernoulli}(\beta_i(\mathbf{z}, \varepsilon))$. Thus, we may apply Lemma 4.9 from Steinke, Nasr, Jagielski to conclude that $W_i = W_{i-1} + \ell_i(X_i, \mathbf{z})$ is stochastically dominated by $\hat{W}_i := \sum_{i \in [n]} S_i$ where $S_i \sim \mathsf{Bernoulli}(\beta_i(\mathbf{z}, \varepsilon))$. □

### A.1 Our bound is tight for RR with priors that are sufficiently close to uniform

Next, we show that the bound in Theorem 3.2 is tight for randomized response on distributions that are not too far from uniform.

**Proposition A.2.** *For some domain $[m]$ let $RR_\varepsilon : \mathcal{X}^n \to [m]$ be the mechanism that applies randomized response independently to each data record, and let $\mathcal{L}$ be the zero-one reconstruction loss. Then, for every data distribution $\mathcal{D}$ such that $\mathcal{D}(a) \leq e^\varepsilon \cdot \mathcal{D}(b)$ for all $a, b \in [m]$, the bound in Theorem 3.2 is achieved by the Bayes' optimal reconstruction attacker. That is, the bound is tight for this mechanism.*

*Proof.* Fix $n = k > 0$, and fix data domain $[m]$. Let $\mathcal{L}(\mathbf{x}, \mathbf{z}) = \sum_{i \in [n]} \mathbb{I}\{\mathbf{x}_i = \mathbf{z}_i\}$. The optimal prior-aware adversary $R$ against randomized response will operate on each term $a \in [m]$ independently and for each one output

$$R(a, \mathcal{D}, \varepsilon) = argmax_{b \in [m]} \Pr_{X \sim \mathcal{D}}(X = b | RR_\varepsilon(X) = a).$$

Thus, we have

$$\begin{aligned}&\Pr_{X \sim \mathcal{D}}(X = b | RR_\varepsilon(X) = a)\\ &= \frac{\Pr(RR_\varepsilon(X) = a | X = b) \cdot \Pr(X = b)}{\Pr(RR_\varepsilon(X) = a)}.\end{aligned}$$

Note that the conditional probability in the numerator is

$$\begin{aligned}&\Pr(RR_\varepsilon(X) = a | X = b)\\ &= \mathbb{I}\{a = b\} \cdot \frac{e^\varepsilon}{e^\varepsilon - 1 + m} + \mathbb{I}\{a \neq b\} \cdot \frac{1}{e^\varepsilon - 1 + m}\end{aligned}$$

and the denominator is

$$\begin{aligned}&\Pr(RR_\varepsilon(X) = a)\\ &= \mathcal{D}(a) \cdot \frac{e^\varepsilon}{e^\varepsilon - 1 + m} + (1 - \mathcal{D}(a)) \cdot \frac{1}{e^\varepsilon - 1 + m}.\end{aligned}$$

Thus, when we simplify the terms we get

$$\begin{aligned}&\Pr_{X \sim \mathcal{D}}(X = b | RR_\varepsilon(X) = a)\\ &= \begin{cases} \frac{\mathcal{D}(a)e^\varepsilon}{\mathcal{D}(a)(e^\varepsilon - 1) + 1} & \text{if } a = b \\ \frac{\mathcal{D}(b)}{\mathcal{D}(a)(e^\varepsilon - 1) + 1} & \text{if } a \neq b. \end{cases}\end{aligned}$$

Thus, the optimal adversary will use the following strategy:

$$R(a, \mathcal{D}, \varepsilon) = \begin{cases} a & \text{if } \mathcal{D}(a) \cdot e^\varepsilon \geq \mathcal{D}(b) \text{ for all } b \in [m] \setminus \{a\} \\ b^* = argmax_{b \in [m]} \mathcal{D}(b) & \text{otherwise} \end{cases},$$

that is, $R$ *ignores* the prior if the mechanism output is consistent with the prior and otherwise *defaults* to the prior.

Now that we have defined the optimal adversary, for every fixed $a \in [m]$ we can show that the bound is tight for distributions $\mathcal{D}$ such that $\mathcal{D}(a) \cdot e^\varepsilon \geq \mathcal{D}(b)$ for all $b \in [m]$.

In this case, optimal attacker will output $a$ as its reconstruction guess. So, the attacker's success rate is simply

$$\begin{aligned}&\Pr(X = a | RR_\varepsilon(X) = a)\\ &= \frac{\mathcal{D}(a)e^\varepsilon}{\mathcal{D}(a)(e^\varepsilon - 1) + 1} = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\mathcal{D}(a)}},\end{aligned}$$

which exactly matches the upper bound in Theorem 3.2.

□

## B Approximate DP Theorems and Proofs

**Theorem B.1** (Extension of Prop 5.7 in [64]). *Let $\mathcal{M}$ satisfy $(\varepsilon, \delta)$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an attacker and let $\mathcal{L}$ be decomposable into $\ell_1 \cdots \ell_n$. Then, for all $v \in \mathbb{R}$ and for all mechanism outputs $a \in Supp(\mathcal{M})$:*

$$\Pr_{X\sim\mathcal{D}}[\mathcal{L}(X, \mathcal{A}(a)) \geq v | \mathcal{M}(X) = a] \leq \Pr_{S_i \sim \mathsf{Bernoulli}(\beta_i(\mathbf{z},\varepsilon))} \left[ G(a) + \sum_{i\in[n]} S_i \geq v \right],$$

*where $\mathbf{z} = \mathcal{A}(a)$ and $\beta_i(\mathbf{z}, \varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X_i\sim\mathcal{D}_i}[\ell_i(X_i,\mathbf{z})=1]}}$ and where $G : Range(\mathcal{M})^n \to \{0, \ldots, n\}$ is independent from $S_1, \ldots, S_n$ with $\mathbb{E}_{A\sim\mathcal{M}(X),G}[G(A)] = n \cdot \delta$.*

*Proof.* The proof follows same general proof structure as that of Proposition 5.7 in Steinke et al. [64].

Fix a product distribution $\mathcal{D} = \mathcal{D}_1 \otimes \ldots \mathcal{D}_n$, an $(\varepsilon, \delta)$-DP mechanism $\mathcal{M}$, a mechanism output $a$, and an attack $\mathcal{A}$. For $i \in [n] \cup \{0\}$, $x_{<i} \in \mathcal{X}^i$, and $b \in \{0, 1\}$, let $\mathcal{M}(x_{<i})$ denote the distribution of $\mathcal{M}(X)$ where $X \sim \mathcal{D}$ conditioned on $X_{<i} = x_{<i}$. In an abuse of notation, we will use the shorthand $\mathcal{M}(x_{<i}, b)$ to denote the distribution of $\mathcal{M}(X)$ for $X \sim \mathcal{D}$ conditioned on $X_{<i} = x_{<i}$ and $\ell_i(X_i, \mathcal{A}(a)) = b$ for $b \in \{0, 1\}$.

Now, for distributions $P$ and $Q$ on $Range(\mathcal{M})$, let $E_{P,Q}$ be the randomized function promised by Lemma 5.6 in [64]. As in their analysis, the internal randomness of $E_{P,Q}$ is independent of everything else. Given a fixed sequence $x_1, ..., x_n$, we will apply the lemma once for each index $i = 1, ..., n$, with $P = \mathcal{M}(x_{<i}, 0)$ and $Q = \mathcal{M}(x_{<i}, 1)$.

To simplify notation, in the remainder of the proof, we assume that the adversary's prediction algorithm $\mathcal{A}$ is simply the identity; we can think of $\mathcal{A}$ as being folded into the loss functions $\ell_i$. Thus we set $\mathbf{z} = \mathcal{A}(a) = a$ and aim to bound $\mathcal{L}(X, a) = \sum_i \ell_i(X_i, a)$.

Thus, for all $i \in [n]$ and for all $x_{<i}$, and all mechanism outputs $a$, we have

$$\Pr_{\substack{X\sim\mathcal{D} \\ a\sim\mathcal{M}(X),E}} \left[ \ell_i(X_i, a) = 1 \wedge E_{\mathcal{M}(x_{<i},0),\mathcal{M}(x_{<i},1)}(a) = 1 \middle| X_{<i} = x_{<i}, \mathcal{M}(X) = a \right] \leq \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr(\ell_i(X_i,a)=1)}} = \beta_i(a, \varepsilon), \tag{1}$$

and for $b \in \{0, 1\}$,

$$\Pr_{\substack{X\sim\mathcal{D} \\ a\sim\mathcal{M}(X),E}} \left[ E_{\mathcal{M}(x_{<i},0),\mathcal{M}(x_{<i},1)}(a) = 1 \middle| X_{<i} = x_{<i}, \ell_i(X_i, a) = b \right] \geq 1 - \delta. \tag{2}$$

In the remainder of the proof we will define three quantities: $W$, which is what we want to upper bound but which isn't always well-behaved, $\widetilde{W}$ which is well-behaved but hard to quantify, and $\widehat{W}$ which is well-behaved and easy to quantify and upper-bounds the other two quantities (with an extra term $F$ to handle the poorly-behaved situations).

Now, for fixed $x \in \mathcal{X}^n$, $k \in [n]$, $a \in Range(\mathcal{M})$, let

$$\widetilde{W}_{i-1}(x, a) = \sum_{m\in[i-1]} \ell_m(x_m, a) \cdot E_{\mathcal{M}(x_{<i},0),\mathcal{M}(x_{<i},1)}(a)$$

and let

$$\widehat{W}_{i-1}(a) = \sum_{m\in[i-1]} S_m \,,$$

where $S_m \sim \mathsf{Bernoulli}(\beta_i(a,\varepsilon))$ independently for each $m \in [i-1]$

Equation (1) shows that each term $S_m$ (in the $\hat{W}$'s) stochastically dominates the term $\ell_m(x_m, a) \cdot E_{\mathcal{M}(x_{<i},0),\mathcal{M}(x_{<i},1)}(a)$ (in the $\tilde{W}$'s), even conditioned on the corresponding terms with smaller indices. By induction (e.g., see Lemma 4.9 in Steinke et al. [64]), for every $i \in [n]$ and every $a$, the conditionally-distributed random variable $(\widetilde{W}_i(X,a)|\mathcal{M}(X)=a)$, where $X \sim \mathcal{D}$, is stochastically dominated by $\widehat{W}_i(a)$.

Now, for fixed $\mathbf{x} \in \mathcal{X}^n$ and $a \in Range(\mathcal{M})$, define

$$G(\mathbf{x}, a) = \sum_{i\in[n]} \mathbb{I}\{E_{\mathcal{M}(x_{<i},0),\mathcal{M}(x_{<i},1)}(a) = 0\},$$

so that

$$\begin{aligned} W_n(\mathbf{x}, a) &:= \mathcal{L}(x, a) := \sum_{i\in[n]} \ell_i(\mathbf{x}_i, a) \\ &\leq \widetilde{W}_n(\mathbf{x}, a) + G(\mathbf{x}, a). \end{aligned}$$

Since that the conditional distribution $(\widetilde{W}_n(X,a)|\mathcal{M}(X)=a)$ where $X \sim \mathcal{D}$ is stochastically dominated by $\widehat{W}_n(a)$, we know that $W_n$ is stochastically dominated by the convolution $\widehat{W}_n(A) + G(X,A)$ for $A \sim \mathcal{M}(X)$.

To summarize:

$$\begin{aligned} &(\widetilde{W}_n(X,a)|\mathcal{M}(X)=a) \preceq_{s.d.} \widehat{W}_n(a) \quad \forall a \\ \implies &W_n(X,A) \preceq_{s.d.} \widehat{W}_n(A) + G(X,A) \quad \text{for } A \sim \mathcal{M}(X) \end{aligned}$$

Furthermore, $\mathbb{E}[G(\mathbf{x},a)] = \sum_{i\in[n]} \Pr[E_{\mathcal{M}(x_{<i},0),\mathcal{M}(x_{<i},1)}(a) = 0] \leq n \cdot \delta$. Lastly, since $\widehat{W}_{i-1}(\mathcal{M})$ does not depend on $X$, the input dataset $X$ does not contribute to the dependence between $G(X,A)$ and $\widehat{W}_{i-1}(A)$ so we can ignore this input to $F$ and assume $G(A) = G(X,A)$ for $X \sim \mathcal{D}$. □

**Theorem B.2.** *Let $\mathcal{M}$ satisfy $(\varepsilon,\delta)$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an attacker and let $\mathcal{L}$ be decomposable into $\ell_1 \cdots \ell_n$. Then, for all $v \in \mathbb{R}$:*

$$\Pr_{\substack{X\sim\mathcal{D}\\ a\sim\mathcal{M}(X)}}[\mathcal{L}(X,\mathcal{A}(a)) \geq v] \leq \Pr_{\substack{X\sim\mathcal{D}\\ a\sim\mathcal{M}(X)\\ S_i\sim\mathsf{Bernoulli}(\beta_i(\mathcal{A}(a),\varepsilon))}} \left[\sum_{i\in[n]} S_i \geq v\right] + n\cdot\delta,$$

*where* $\beta_i(\mathcal{A}(a),\varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X\sim\mathcal{D}_i}[\ell_i(X,\mathcal{A}(a))=1]}}$.

Note that since $\delta$ is typically much smaller than $1/n$, the additive term brought on by approximate DP is fairly small, however the flipping probability $\beta$ is now a random variable itself, which must be estimated from multiple runs of the mechanism on fresh data samples.

Next we present an approximate DP bound that gives a tighter dependence on $\delta$ and which can be computed in only one run of the mechanism, under the restriction that the attack success metric is *target-separable*, that is that $\ell_i(X_i, z) = \ell_i(X_i, z_i)$. The flipping probabilities $\beta$ are computed using the Bayes optimal *a priori* attack $\mathbf{z}^*$ on each target. When the prior distributions are close to uniform, then using $\mathbf{z}^*$ to compute the prior should be fairly tight, but for highly skewed data distributions, Theorem B.2 will likely be much tighter.

**Theorem B.3** (Generalization of Theorem 5.2 [64])**.** *Let $\mathcal{M}$ satisfy $(\varepsilon,\delta)$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an attacker outputting $k = n$ guesses $z = (z_1, \ldots, z_n)$, and let $\mathcal{L}$ be decomposable into $\ell_1 \ldots \ell_n$ satisfying target separability: $\ell_i(X_i, z) = \ell_i(X_i, z_i)$ Then, for all $v \in \mathbb{R}$,*

$$\Pr_{\substack{X\sim\mathcal{D}\\ a\sim\mathcal{M}(X)}}[\mathcal{L}(X,\mathcal{A}(a)) \geq v] \leq \Pr_{S_i^*\sim\mathsf{Bernoulli}(\beta_i(\varepsilon))} \left[\sum_{i\in[n]} S_i^* \geq v\right] + \alpha\cdot n\cdot\delta,$$

*where* $\beta_i(\varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X\sim\mathcal{D}_i}[\ell_i(X, z_i^*)=1]}}$ *where* $z_i^*$ *is the a priori Bayes optimal attack attempt on* $\mathcal{D}_i$ *and*

$$\alpha = \max\left\{\frac{1}{j}\left(\Pr_{S_i^*\sim\mathsf{Bernoulli}(\beta_i(\varepsilon))}\left[\sum_{i\in[n]} S_i^* \geq v - j\right] - \Pr_{S_i^*\sim\mathsf{Bernoulli}(\beta_i(\varepsilon))}\left[\sum_{i\in[n]} S_i^* \geq v\right]\right) : j \in [n]\right\}.$$

*Proof.* Fix an $(\varepsilon, \delta)$-DP mechanism $\mathcal{M}$, fix a metric $\mathcal{L}$ that is decomposable into $\ell_1 \cdots \ell_n$, and fix an attacker $\mathcal{A}$. Let $\mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be the data distribution.

We begin by using Theorem B.1 and taking the expectation over the potential mechanism outputs to get

$$\begin{aligned}
&\Pr_{X\sim\mathcal{D}, a\sim\mathcal{M}(X)}[\mathcal{L}(X, a) \geq v] \\
&\leq \Pr_{\substack{a\sim\mathcal{M}(X) \\ S_i\sim\mathsf{Bernoulli}(\beta_i(a,\varepsilon))}}\left[F(a) + \sum_{i\in[n]} S_i \geq v\right],
\end{aligned}$$

where $\beta_i(a, \varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X\sim\mathcal{D}_i}[\ell_i(X,a)=1]}}$ and where $F : Range(\mathcal{M}) \to \{0, \ldots, n\}$ is independent from $S_1, \ldots, S_n$ with $\mathbb{E}_{A\sim\mathcal{M}(X),F}[F(A)] = n \cdot \delta$.

Since the constraints on $F$ are fairly simple, we will set up a linear program to find the optimal distribution for random variable $F(a)$ for a given mechanism output $a$ and threshold $v \in \mathbb{R}$. In this linear program, the variables are the values on $\Pr(F(a) = j)$ for each $j \in \{0, \ldots, n\}$ and the two constraints are: (1) it must be a probability distribution, and (2) the expectation of $F$ is at most $n \cdot \delta$. So, for a fixed mechanism output $a$ and threshold $v$:

$$\begin{aligned}
\text{maximize} \quad & \Pr_{S_i\sim\mathsf{Bernoulli}(\beta_i(a,\varepsilon))}\left[F(a) + \sum_{i\in[n]} S_i \geq v\right] \\
& = \sum_{j=0}^{n} \Pr_F[F(a) = j]) \cdot \Pr_{S_1\ldots S_n}\left[\sum_{i\in[n]} S_i \geq v - j\right] \\
\text{subject to} \quad & \mathbb{E}_F[F(a)] = \sum_{j=0}^{n} \Pr[F(a) = j] \cdot j \leq n \cdot \delta, \\
& \sum_{j=0}^{n} \Pr[F(a) = j] = 1, \text{and} \\
& \Pr[F(a) = j] \geq 0 \quad \forall j \in \{0, \ldots, n\}.
\end{aligned}$$

By strong duality, the linear program above has the same value as its dual (again, for fixed $a, v$):

$$\begin{aligned}
\text{minimize} \quad & n\delta\alpha + \gamma \\
\text{subject to} \quad & \alpha \cdot j + \gamma \geq \Pr_{S_1\ldots S_n}\left[\sum_{i\in[n]} S_i \geq v - j\right] \\
& \forall j \in \{0, \ldots, n\}, \\
& \alpha \geq 0
\end{aligned}$$

Any feasible solution to the dual gives an upper bound on the primal. So, we can use the solution given by

$$\gamma = \Pr[S^* \geq v] \quad \text{and}$$
$$\alpha = \max\left(\{0\} \cup \frac{1}{j}\Pr[S^* \geq v - j] - \gamma : j \in [n]\right),$$

where $S^*$ is any distribution that satisfies

$$\Pr_{S^*}[S^* \geq v - j] \geq \Pr_{S_i \sim \mathsf{Bernoulli}(\beta_i(a,\varepsilon))}\left[\sum_{i\in[n]} S_i \geq v - j\right]$$

for all $j \in \{0, \ldots, n\}$ and for all $a \in Supp(\mathcal{M})$.

To construct such a distribution $S^*$, we utilize the target-separability assumption, $\ell_i(X_i, z) = \ell_i(X_i, z_i)$. Because the attacker's guesses are completely decoupled, we can construct the joint a priori optimal attack $z^* = (z_1^*, \ldots, z_n^*)$ component-wise, by choosing each local guess to maximize its individual success probability:

$$z_i^* \in \arg\max_{z_i} \Pr_{X_i \sim \mathcal{D}_i}[\ell_i(X_i, z_i) = 1]$$

For any mechanism output $a$, the adversary produces an attack attempt $z(a) = \mathcal{A}(a) = (z_1(a), \ldots, z_n(a))$. By the definition of $z_i^*$, it must hold for every target $i \in [n]$ that:

$$\Pr_{X_i \sim \mathcal{D}_i}[\ell_i(X_i, \mathcal{A}(a)) = 1] \leq \Pr_{X_i \sim \mathcal{D}_i}[\ell_i(X_i, z_i^*) = 1]$$

Because the Bernoulli parameter function $\beta(p) = \frac{e^\varepsilon}{e^\varepsilon - 1 + 1/p}$ is strictly monotonically increasing with respect to $p \in (0, 1]$, this local probability bound guarantees pointwise parameter domination for all targets simultaneously:

$$\beta_i(a, \varepsilon) \leq \beta_i(z^*, \varepsilon) \quad \forall i \in [n], \quad \forall a \in Supp(\mathcal{M}).$$

Let $S_i^* \sim \text{Bernoulli}(\beta_i(\varepsilon))$ be independent random variables. Since $\beta_i(a, \varepsilon) \leq \beta_i(\varepsilon)$ for all $i$, a standard Poisson-binomial coupling guarantees that $\sum_{i\in[n]} S_i \leq \sum_{i\in[n]} S_i^*$ almost surely. Consequently, we obtain the required upper-tail stochastic domination:

$$\Pr\left[\sum_{i\in[n]} S_i^* \geq v - j\right] \geq \Pr\left[\sum_{i\in[n]} S_i \geq v - j\right].$$

Thus, we can choose $S^* = \sum_{i\in[n]} S_i^*$, which is independent of the mechanism output $a$ and satisfies the dual LP constraint. □

# C Omitted $f$-DP Theorems and Proofs

First, recall the definition of $f$-DP.

**Definition C.1** (Trade-off function [25])**.** *For any two probability distributions, $P$ and $Q$ on the same space, the trade-off function $T(P, Q) : [0, 1] \to [0, 1]$ is defined as:*

$$T(P, Q)(\alpha) \triangleq \inf_\phi\{\beta_\phi : \alpha_\phi \leq \alpha\}$$

*where the infimum is taken over all rejection rules $\phi$ for which $\alpha_\phi$ and $\beta_\phi$ are the type I and type II errors, respectively.*

**Definition C.2** ($f$-DP [25])**.** *A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies $f$-DP if, for any two adjacent datasets $\mathbf{x}, \mathbf{x}' \in \mathcal{D}$, and $\alpha \in [0, 1]$ it holds:*

$$T(\mathcal{M}(\mathbf{x}), \mathcal{M}(\mathbf{x}'))(\alpha) \geq f(\alpha)$$

We present two $f$-DP theorems that were omitted from the main body. Then, we will present the proofs of all three $f$-DP bounds.

**Theorem C.3** (Generalization of Proposition 5.7 [64]). *Let $\mathcal{M}$ satisfy $f$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an attacker and let $\mathcal{L}$ be decomposable into $\ell_1 \cdots \ell_n$. Then, for all $v \in \mathbb{R}$, $\varepsilon > 0$, and mechanism outputs $a \in Supp(\mathcal{M})$:*

$$\Pr_{X\sim\mathcal{D}}[\mathcal{L}(X, \mathcal{A}(a)) \geq v | \mathcal{M}(X) = a] \leq \Pr_{S_i \sim \text{Bernoulli}(\beta_i(\mathbf{z},\varepsilon))} \left[ G(a) + \sum_{i\in[n]} S_i \geq v \right],$$

*where $\mathbf{z} = \mathcal{A}(a)$ and $\beta_i(\mathbf{z}, \varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X_i\sim\mathcal{D}_i}[\ell_i(X_i,\mathbf{z})=1]}}$ and where $G : Range(\mathcal{M})^n \to \{0, \ldots, n\}$ is independent from $S_1, \ldots, S_n$ with $\mathbb{E}_{a\sim\mathcal{M}(X),G}[G(a)] = n \cdot \delta_f(\varepsilon)$.*

The main challenge with this bound is that $G$ is unspecified and generally depends on the mechanism.

In order to avoid flipping probabilities that depend on the attack as in Theorem 3.3 (and thus require multiple runs of the mechanism to estimate), we present a version of this bound that uses the worst-case prior instead. We compute the $\beta_1, \ldots, \beta_n$ with respect to the guesses $\mathbf{z}^*$ that maximize the attack success on $\mathcal{D}$. This approach provides a single distribution that dominates the true posterior distribution *for all mechanism outputs*. Because we're essentially paying for the heaviest elements in $\mathcal{D}$, the bound in Theorem B.3 is tightest for data distributions that are close to uniform.

**Theorem C.4** (Generalization of Theorem 5.2 [64]). *Let $\mathcal{M}$ satisfy $f$-DP. Let dataset $X \sim \mathcal{D} = \mathcal{D}_1 \otimes \ldots \otimes \mathcal{D}_n$ be drawn from a product distribution. Let $\mathcal{A}$ be an attacker outputting $k = n$ guesses $z = (z_1, \ldots, z_n)$ and let $\mathcal{L}$ be decomposable into $\ell_1 \cdots \ell_n$ satisfying target separability: $\ell_i(X_i, z) = \ell_i(X_i, z_i)$. Then, for all $\varepsilon > 0$, $v \in \mathbb{R}$,*

$$\Pr_{\substack{X\sim\mathcal{D} \\ a\sim\mathcal{M}(X)}}[\mathcal{L}(X, \mathcal{A}(a)) \geq v] \leq \Pr_{S_i^* \sim \text{Bernoulli}(\beta_i(\varepsilon))} \left[ \sum_{i\in[n]} S_i^* \geq v \right] + \alpha \cdot n \cdot \delta_f(\varepsilon),$$

*where $\beta_i(\varepsilon) = \frac{e^\varepsilon}{e^\varepsilon - 1 + \frac{1}{\Pr_{X\sim\mathcal{D}_i}[\ell_i(X,z_i^*)=1]}}$ where $z_i^*$ is the a priori Bayes optimal attack attempt for target $i$ on $D_i$ and*

$$\alpha = \max\left\{ \frac{1}{j} \left( \Pr_{S_i^* \sim \text{Bernoulli}(\beta_i(\varepsilon))} \left[ \sum_{i\in[n]} S_i^* \geq v - j \right] - \Pr_{S_i^* \sim \text{Bernoulli}(\beta_i(\varepsilon))} \left[ \sum_{i\in[n]} S_i^* \geq v \right] \right) : j \in [n] \right\}.$$

The approx-DP analog of this bound is Theorem B.3.

To prove Theorem C.3, Theorem 3.3, and Theorem C.4, will use the fact that $f$-DP is known to be equivalent to an infinite collection of $(\varepsilon, \delta_f(\varepsilon))$-DP guarantees (Proposition 2.12 [25]) where $\delta_f(\varepsilon) = 1 + f^*(-e^\varepsilon)$ and $f^*$ is the convex conjugate of $f$. The upper bounds for the entire collection of $(\varepsilon, \delta(\varepsilon))$-DP guarantees will simply be satisfied simultaneously by an $f$-DP mechanism. Thus, the proofs for Theorem C.3, Theorem 3.3, and Theorem C.4 follow immedately from their $(\varepsilon, \delta)$-DP analogues, Theorem B.1, Theorem B.2, and Theorem B.3, respectively.

# D Proof of Robustness to Prior Misspecification

**Corollary 3.4.** *Let $\mathcal{M}$, $\mathcal{D}$, $\mathcal{A}$, and $\mathcal{L}$ be as in Theorem 3.2. Let $\hat{p}_1, \ldots, \hat{p}_n$ be estimates computed independently of $X$ and of the randomness of $\mathcal{M}$, such that for some $\eta \in (0, 1)$,*

$$\Pr_{\hat{p}}\left[ \hat{p}_i(\mathbf{z}) \geq p_i(\mathbf{z}) \;\; \forall i \in [n], \; \forall \mathbf{z} \right] \geq 1 - \eta.$$

*Then, with probability at least $1 - \eta$ over $\hat{p}$, for all $v \in \mathbb{R}$ and all $a \in Supp(\mathcal{M})$,*

$$\Pr_{X\sim\mathcal{D}}\left[\mathcal{L}(X, \mathcal{A}(a)) \geq v \mid \mathcal{M}(X) = a\right] \;\leq\; \Pr_{\hat{S}_i \sim \text{Bernoulli}(\beta(\hat{p}_i(\mathbf{z}),\varepsilon))} \left[ \textstyle\sum_{i\in[n]} \hat{S}_i \geq v \right], \qquad \mathbf{z} = \mathcal{A}(a).$$

*The same substitution $p_i \mapsto \hat{p}_i$ preserves Theorem 3.3 and Theorems B.1 to B.3, C.3 and C.4. For Theorems B.3 and C.4, it suffices to upper bound the $n$ scalars $p_i^* = \max_{z_i} p_i(z_i)$, and $\hat{p}_i^*$ replaces $p_i^*$ in both $\beta_i$ and $\alpha$.*

To save space, we give a proof sketch for the affected theorems.

*Proof.* Let $G$ be the event that $\hat{p}_i(\mathbf{z}) \geq p_i(\mathbf{z})$ for all $i \in [n]$ and all $\mathbf{z}$. By assumption, $\Pr[G] \geq 1 - \eta$. Since $\hat{p}$ is independent of $X$ and of the coins of $\mathcal{M}$, conditioning on any realization of $\hat{p}$ leaves the joint distribution of $(X, \mathcal{M}(X))$ unchanged. Fix a realization of $\hat{p}$ in $G$.

*Pointwise domination.* The map $p \mapsto \beta(p, \varepsilon)$ has derivative $\frac{e^{\varepsilon}}{(p(e^{\varepsilon}-1)+1)^2} > 0$, so it is increasing on $[0, 1]$. Hence, for every $\mathbf{z}$ and $i$, $\hat{\beta}_i(\mathbf{z}, \varepsilon) := \beta(\hat{p}_i(\mathbf{z}), \varepsilon) \geq \beta(p_i(\mathbf{z}), \varepsilon) = \beta_i(\mathbf{z}, \varepsilon)$.

*Coupling.* Let $U_1, \ldots, U_n \sim \mathrm{Unif}[0, 1]$ be independent, and set $S_i = \mathbb{I}\{U_i \leq \beta_i(\mathbf{z}, \varepsilon)\}$ and $\hat{S}_i = \mathbb{I}\{U_i \leq \hat{\beta}_i(\mathbf{z}, \varepsilon)\}$. Then $S_i \sim \mathsf{Bernoulli}(\beta_i)$, $\hat{S}_i \sim \mathsf{Bernoulli}(\hat{\beta}_i)$, and $\sum_i S_i \leq \sum_i \hat{S}_i$ almost surely. So for every $t \in \mathbb{R}$, $\Pr[\sum_i S_i \geq t] \leq \Pr[\sum_i \hat{S}_i \geq t]$. The same holds after adding any random variable independent of $U_1, \ldots, U_n$ to both sums.

*Pure DP.* Combining the coupling with Theorem 3.2 (taking $t = v$) gives the displayed bound for every $a$ simultaneously, because $G$ holds uniformly over $\mathbf{z}$.

*Theorems B.1 and C.3.* Apply the coupling with the independent variable $G(a)$ added to both sums.

*Theorems B.2 and 3.3.* The domination holds for every $a$, so it is preserved by taking the expectation over $X \sim \mathcal{D}$ and $a \sim \mathcal{M}(X)$.

*Theorems B.3 and C.4.* Here we only need $\hat{p}_i^* \geq p_i^* = \max_{z_i} p_i(z_i)$ for all $i$. In the proof of Theorem B.3, for every $a$ and $i$ we then have $\beta_i(a, \varepsilon) \leq \beta(p_i^*, \varepsilon) \leq \beta(\hat{p}_i^*, \varepsilon)$. Let $\hat{S}_i^* \sim \mathsf{Bernoulli}(\beta(\hat{p}_i^*, \varepsilon))$ be independent. The coupling shows that $\hat{S}^* = \sum_i \hat{S}_i^*$ satisfies the dual constraint $\Pr[\hat{S}^* \geq v - j] \geq \Pr[\sum_i S_i \geq v - j]$ for all $j$ and $a$. Therefore $\hat{\gamma} = \Pr[\hat{S}^* \geq v]$ and $\hat{\alpha} = \max\left(\{0\} \cup \{\frac{1}{j}(\Pr[\hat{S}^* \geq v - j] - \hat{\gamma}) : j \in [n]\}\right)$ form a feasible dual solution, which yields the upper bound $\hat{\gamma} + \hat{\alpha} \cdot n \cdot \delta$. The $f$-DP case (Theorem C.4) follows by replacing $\delta$ with $\delta_f(\varepsilon)$. □

# E Idealized Gaussian Attack

The idealized setting for DP-SGD is a simple gaussian mechanism that calculates the noisy sum of orthogonal unit vectors, i.e., in Algorithm 1, each $\mathcal{D}_i$ is a distribution over $k$ $d$-dimensional orthogonal unit vectors $\{V_1^i, \ldots, V_k^i\}$ and $a = \mathcal{M}(X_1, \ldots, X_n) = \sum_i X_i + \mathcal{N}(0, \sigma^2 I)$. In this case, the bayes optimal adversary $\mathcal{A}$ reconstructs each sample by choosing the vector with highest probability combined with the prior, i.e., $\mathbf{z}_i = \arg\max_{V_j^i} \Pr[X_i = V_j^i | a] \propto \arg\max_{V_j^i} \Pr[a | X_i = V_j^i] \cdot \Pr[V_j^i \sim \mathcal{D}]$. Lastly, the success of the adversary is exact match, i.e., $\mathcal{L}(X, \mathbf{z}) = \sum_i \mathbb{I}\{X_i = \mathbf{z}_i\}$.

In our experiments, since we are reconstructing a single digit, we set $k = 10$ and set $d = n * k$ where $n$ is the number of canaries we are reconstructing. Each $V_j^i$ is then a one-hot vector where the 1 is at index $i * k + j$. This ensures that the canaries both *within* and *between* each distribution $\mathcal{D}_i$ are orthogonal to each other. Note that this setting is functionally equivalent to an adversary with access to all intermediate models in DP-SGD in the case that the gradients of all possible canaries are orthogonal to each other (e.g., DP auditing [46, 52]).

# F Bounding Tabular Data Reconstruction from Noisy Marginals

In this section, we investigate the tightness of our theoretical framework with respect to another commonly used DP mechanism. Specifically, we compare our theoretical bounds against the success rate of a multi-attribute inference (M-AI) attack run against noisy $k$-way marginals, a commonly used DP mechanism in synthetic data. In the process we also show how our framework can model novel privacy risks and attacks targeting multiple users.

## F.1 Experimental Setup

**$k$-way marginals.** $k$-way marginals are useful summary statistics that effectively describe tabular datasets. Specifically, a $k$-way marginal counts the number of records in the dataset whose values agree on $k$ attributes, e.g., "How many people are **aged 40**, **unemployed** and **unmarried**" is a 3-way marginal.

Although $k$-way marginals are useful summary statistics, they have been known to be vulnerable to powerful attacks [21, 32, 38]. Therefore, calibrated Gaussian noise is typically added to these marginals so that they satisfy DP, which forms the basis for many popular DP synthetic data algorithms such as MST [47], AIM [48], and RAP [7].

**Noisy $k$-way marginals.** In our experiments, we assume the "replace-one" adjacency and additionally assume that for a given set of attributes, full marginal vectors across all possible values are released. Therefore, we calibrate the level of Gaussian noise added to a $L_2$ sensitivity of $\sqrt{2}$ and $\delta = 10^{-5}$ for the entire marginal vector using Gaussian DP [25] accounting. Furthermore, as we are mainly focused on M-AI, we only calibrate the noise for the attributes that are *unknown*, i.e., when $d'$ columns are unknown out of $d$ columns, the noise is calibrated to $\binom{d}{k} - \binom{d-d'}{k}$ compositions of the noisy $k$-way marginal mechanism. Lastly, we fix $k = 3$ as these were the most powerful attacks in prior work [21]. We provide a formal definition of the noisy $k$-way marginal mechanism below:

**Definition F.1** (**$k$-way marginal query**). *Given a dataset $X$ over a data domain with $d$ attributes, $\mathcal{X} = \mathcal{X}_1 \times ...\mathcal{X}_d$, a $k$-way marginal query is defined by a subset of attributes $V \subseteq [d]$, $|V| = k$ and corresponding values for each of the attributes $v \in \prod_{i \in V} \mathcal{X}_i$. Given the pair $(V, v)$, define $\mathcal{X}(V, v) = \{x \in \mathcal{X} : x_i = v_i \; \forall i \in V\}$. The corresponding $k$-way marginal query is defined as follows where $\mathbb{1}$ is the indicator function that maps elements of the set to 1 and the rest to 0:*

$$Q_{V,v}(X) = \frac{1}{|X|} \sum_{x \in X} \mathbb{1}(x \in \mathcal{X}(V, v))$$

**Definition F.2** (**Noisy $k$-way marginal**). *Given a dataset $X$ over a data domain with $d$ attributes, $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. The noisy $k$-way marginal mechanism is defined by a subset of attributes $V \subseteq [d]$ and noise level $\sigma$ and outputs:*

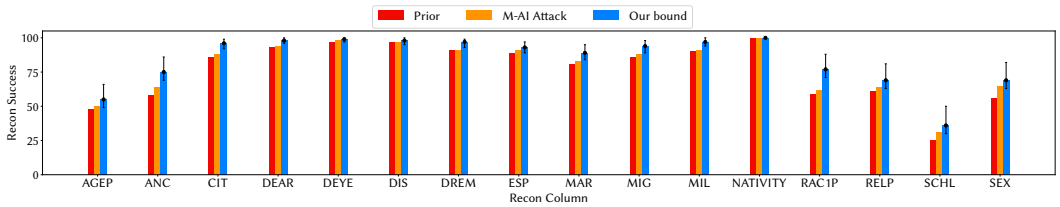
$$\mathcal{M}_{V,\sigma}(X) = \left( Q_{V,v}(X) : \forall v \in \prod_{i \in V} \mathcal{X}_i \right) + \mathcal{N}(0, 2\sigma^2 I)$$

**Adapting RAP attack.** In prior work, Dick et al. [21] present a reconstruction attack against *exact* $k$-way marginals by using the optimization objective introduced by Aydore et al. [7], called the Relaxed Adaptive Projection (RAP) attack. Specifically, they relax discrete datasets and the corresponding $k$-way marginals to the continuous domain and use standard optimization techniques such as Stochastic Gradient Descent to reconstruct the dataset using the following optimization objective: $\arg\min_{\tilde{X}'} ||\tilde{Q}(\tilde{X}) - \tilde{Q}(\tilde{X}')||_2$ where $\tilde{Q}$, $\tilde{X}$, and $\tilde{X}'$ are the relaxed marginals, original dataset, and reconstructed dataset, respectively. Finally, the relaxed reconstructed dataset is projected back to the discrete domain using rounding techniques.

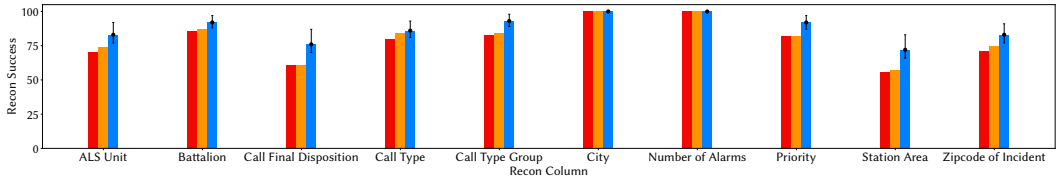
In our experiments, we make three main modifications to what has been done in prior work and refer to this modified attack as our "M-AI attack". First, we evaluate the effectiveness of the RAP attack against *noisy* $k$-way marginals instead of exact. By doing so, we measure the gap between our theoretical success bounds provided by DP and concrete attacks that can be run against DP mechanisms. On the other hand, Dick et al. [21] focus on non-private mechanisms only.

Second, we adapt the RAP attack to perform *multi-attribute inference* (M-AI) instead of reconstruction. Unlike reconstruction, in M-AI the adversary is given access to a subset of attributes of the original dataset and tasked with reconstructing the remaining subset of attributes. Naively, the reconstructed dataset $\tilde{X}'$ could be instantiated with these known attributes as the prior [21]. However, we note that doing so does not guarantee that the final reconstructed dataset will be consistent with these known attributes, which might weaken the M-AI attack. Therefore, along with instantiating $\tilde{X}'$ with the known attributes, during the optimization process, we additionally freeze these parameters so that the optimization process focuses purely on reconstructing the unknown attributes.

Lastly, instead of deriving the prior implicitly from other similar datasets, in our experiments we *explicitly calculate the prior*. Specifically, we fit a Bayesian Network (BayNet) model [41] to raw datasets and sample the "original" datasets from this model that the adversary then attempts to reconstruct. We do so as this enables us to attack realistic looking datasets, while at the same time explicitly defining the prior distribution necessary for our bounds calculations. Therefore, in our modified RAP attack, we initialize the unknown attributes of $\tilde{X}'$ to the explicit prior distribution conditioned on the known attributes calculated from the BayNet model, as opposed to initializing them to an adjacent dataset as done previously [21].




(a) ACS




(b) FIRE

Figure 4: Success of reconstructing a single column at $\varepsilon = 1$ and $\delta = 10^{-5}$ for different datasets and select columns.

**Datasets.** We experiment with two raw datasets in this work that have been commonly used for DP research [9, 47, 48]: American Community Survey (ACS) [66] and San Francisco Fire Department Calls for Service (FIRE) [20]. Following the data preparation in prior work [4], we discretize and post-process the two datasets to have 16 and 10 attributes, respectively. These datasets are then treated as the "raw" datasets used to fit a BayNet model, from which 100 fresh records are sampled as the "original" dataset.

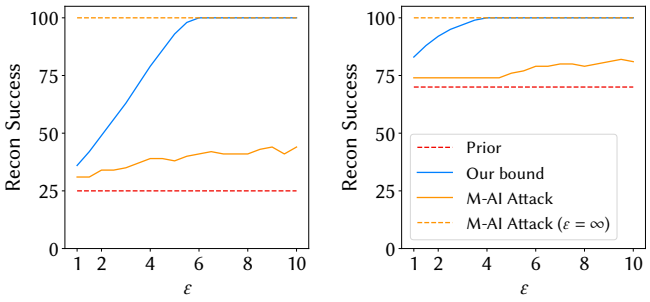
### F.2 1 Column Reconstruction

We first begin with reconstructing a single column. Note that this is very similar to traditional AI attacks, but we reconstruct the entire column for all records in the dataset, as opposed to AI attacks that typically only reconstruct a single attribute of a single record. In Figure 4 we report the total number of correctly reconstructed records when using our M-AI attack to reconstruct different columns from noisy 3-way marginals of different datasets at a privacy level of $\varepsilon = 1$. Additionally, we compare the success of our attack with our theoretical upper bounds (derived from Theorem C.4) at varying levels of confidence (5%, 50%, 95%) as error bars and the success of attempting to reconstruct the column from the prior distribution (derived from the BayNet model) alone.

First, we find that in such real-world settings the bounds can sometimes be vacuous on simple tasks such as reconstructing a single attribute. Specifically, even for a high privacy level ($\varepsilon = 1$), our theoretical bounds predict that most of the columns can be reconstructed almost fully ($\geq 80/100$) with a high level of confidence (95%), with the exception of 'AGEP' and 'SCHL' columns of the ACS dataset. This is mainly because real-world priors can be very large when reconstructing a single column. Hence, we echo the concerns of Jayaraman and Evans [36], as care has to be taken when evaluating privacy attacks with respect to large priors in real-world datasets, which might seemingly lead to successful attacks even against DP mechanisms. However, these attacks may not necessarily be *significant* as they may not perform much better than the prior or anywhere close to the theoretical limit posed by DP.

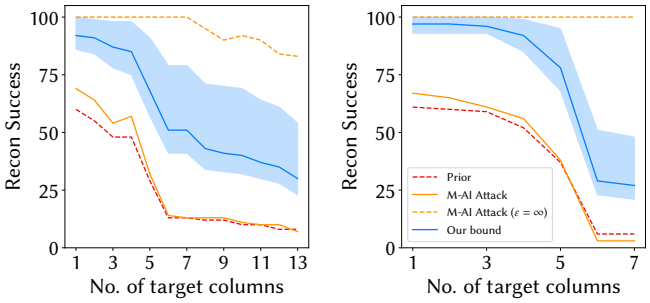
Second, similar to our previous findings on data extraction, we note that real-world state-of-the-art attacks are still far from reaching our bound. Although across both datasets and all columns, the success of our M-AI attack either matches or exceeds the success of using the prior alone, our M-AI attack is still far from realizing the limit of our theoretical bounds in many settings. For instance, our M-AI attack achieves $65\%$ reconstruction success for the 'SEX' attribute in the ACS dataset compared to the prior, which only achieves $56\%$. However, our 95% confidence level upper bound on the reconstruction success is $84\%$.

**Impact of privacy level $\varepsilon$.** Next, in Figure 5 we look at the impact of the privacy level $\varepsilon$ on the reconstruction success of our bound and our M-AI attack. To that end, we vary the $\varepsilon$ and plot the corresponding success of reconstructing the 'SCHL' and 'ALS Unit' columns of the ACS and FIRE




(a) ACS, 'SCHL' (b) FIRE, 'ALS Unit'

Figure 5: Success of reconstructing a single column at varying privacy levels $\varepsilon$ and fixed $\delta = 10^{-5}$.




(a) ACS (b) FIRE

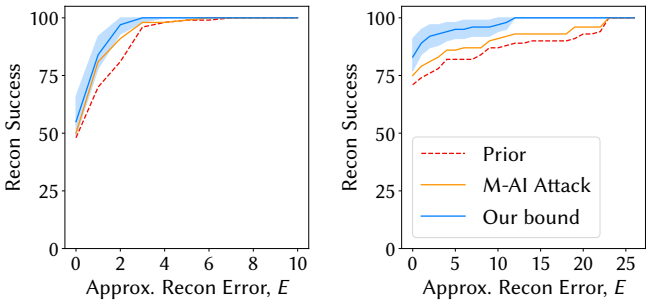
Figure 6: Success of reconstructing multiple columns at $\varepsilon = 3$ and $\delta = 10^{-5}$.

datasets, respectively. Overall, as the $\varepsilon$ increases, our attack becomes more significant, outperforming the prior more and becoming closer to our theoretical bounds. In fact, for the vacuous privacy guarantee of $\varepsilon = \infty$ our attack matches the theoretical bound and completely reconstructs the target columns of all records, whereas the prior success remains very low. This indicates that, our M-AI attack is not fundamentally limited in its ability to leverage privacy leakage. However, the success of state-of-the-art attacks that might work well against non-private mechanisms may not necessarily transfer to significant attacks for private mechanisms, especially at high levels of privacy. Therefore, we believe that more research should focus on designing attacks specifically for DP mechanisms that can better leverage the privacy leakage and prior probabilities from these mechanisms specifically.

### F.3 Multi-Column Reconstruction

We evaluate the success of our M-AI attack when reconstructing varying number of attributes and compare it with our theoretical bounds at fixed $\varepsilon = 3$ in Figure 6. To that end, for each raw dataset, we fit a single BayNet model and reconstruct an increasing number of columns in the visit order of the network. Note that in order to perform M-AI, a fixed number of columns must be known to match the reconstructed records with the original records, which we set to 3 in this case. Therefore the maximum number of columns reconstructed for the ACS and FIRE datasets are 13 and 7, respectively.

Here, we observe that our M-AI attack remains most significant when the number of target columns reconstructed is low and quickly converges to the success of the prior. For instance, for both datasets, we observe that when reconstructing more than 5 columns, our M-AI attack performs equivalently to just using the prior alone. Furthermore, for the FIRE dataset, even though the theoretical bound stays roughly the same when reconstructing 3 columns as 1 column, the success of the M-AI attack reduces in tandem with the success of the prior. This indicates that the M-AI attack is not optimal and largely depends on the prior as it struggles to reconstruct records as effectively under more difficult settings, even when the privacy leakage is the same.




(a) ACS, 'AGEP' (b) FIRE, 'Zipcode of Incident'

Figure 7: Success of approximately reconstructing a single column (up to $L_1$ error) at $\varepsilon = 1$ and $\delta = 10^{-5}$.

### F.4 Approximate Reconstruction

Lastly, one of the key features of our framework is its ability to support both *complex success metrics* and *multiple targets* simultaneously, which none of the prior work has achieved before. To that end, in this section, we focus on approximately reconstructing multiple records at the same time.

In Figure 7 we report our reconstruction success when reconstructing the 'AGEP' and 'Zipcode of Incident' columns of the ACS and FIRE datasets, respectively, and compare it against our bounds when the success metric is defined as being within some $L_1$ error $E$, i.e., $\ell_i(X_i, \mathbf{z}) = \mathbb{I}\{|X_i - \mathbf{z}_i| \leq E\}$. Note that the 'AGEP' and 'Zipcode of Incident' columns have a domain of 10 and 26, respectively, which corresponds to the maximum reconstruction error possible in the Figure. When $E = 0$, this corresponds to the "exact match" metric we have been using previously.

Here, we observe that even for approximate definitions of reconstruction, our attack consistently performs better than the prior, at times even matching our bound, albeit at very low levels of confidence. This shows that not only can our framework successfully handle complex success metrics such as approximate reconstruction, but also that real world attacks might potentially perform better with respect to complex metrics as compared to simple metrics, which can be overly restrictive.

### F.5 Key Takeaways

We show that while our bounds hold empirically, they remain loose when compared to the reconstruction success observed from actual attacks. We believe that the main reason for this is that current state of the art reconstruction attacks may not be optimal for DP mechanisms, even though they work well under the non-private setting. Nevertheless, in this section we show how our framework is capable of modelling *novel privacy risks*, providing meaningful *population-level* attack success bounds, and accommodating *complex success metrics*, which have not been simultaneously captured by any of the prior work discussed.

## G Bridging Existing Attack Frameworks

In this section, we show how our framework bridges and extends the current state of the art frameworks bounding the privacy leakage for a handful of influential works.

### G.1 Privacy Auditing in One Run

The influential work of Steinke, Nasr, and Jagielski [64] is perhaps the most closely related work to ours. At a high level, their work also gives a high probability bound on the success of an adversary with multiple targets, but only for MI. In Section 3.3 we describe how to encode their specific MI setting in our bounds. We generalize their bounds beyond MI attacks in two ways. First, we allow the data to be generated according to general product distributions rather than just sampling membership bits. Second, we generalize beyond measuring success via exact match by incorporating complex success metrics into the prior success probability.

**Auditing $f$-Differential Privacy in One Run (MMC).** Mahloujifar, Melis, and Chaudhuri [46] also derive MI bounds for $f$-DP mechanisms, which better captures common mechanisms like the Gaussian mechanism. They also generalize their membership inference bounds to what we call a "multi-membership inference" attack where each canary is drawn uniformly from a set of possible canaries, and the attacker is successful if they correctly guess the *index* of the true canary out of the possible ones.

Multi-membership has a number of limitations in real-world applicability. In multi-membership inference, canaries have to have some ordering within their domain and attacks typically search over the potentially large domain of canaries to find the correct index. Furthermore, multi-membership inference cannot handle approximate reconstruction, since the attack success is measured by exactly matching the *index* of the canary output by the adversary. On the other hand, in a full reconstruction attack, which is handled by our framework, the adversary outputs *the canary itself* (or an approximation of it) that does not need any ordering and can be compared against original records using potentially approximate distance functions. Lastly, the MMC bound is not flexible enough to handle cases where there are more guesses than canaries, i.e., $k > n$.

### G.2 Narcissus Resiliency

Perhaps the most interesting connection is that to the Narcissus Resiliency framework proposed by Cohen et al. [18]. They define a general framework for evaluating a wide range of data reconstruction attacks and protections, including: Predicate singling out [17], membership inference, differential privacy, one-way functions [22, 68], and various guarantees of encryption schemes. The core insight of their framework is to measure the attacker's success probability relative to its baseline probability of success on a fresh sample of data rather than using a fixed baseline for all attackers. A mechanism is considered "narcissus resilient" if the two success probabilities are close for all attackers. Specifically,

**Definition G.1** (Narcissus resiliency, [18])**.** *Let $\mathcal{C}$ be a class of data distributions and let $\mathcal{L}$ be a predicate over dataset, attack attempt pairs. Algorithm $\mathcal{M}$ is $(\varepsilon, \delta, \mathcal{C})$-$\mathcal{L}$-Narcissus-resilient if for all $\mathcal{D} \in \mathcal{C}$ and for all attackers $\mathcal{A}$ it holds that*

$$\Pr_{\substack{X\sim\mathcal{D}\\ a\leftarrow\mathcal{M}(X)\\ \mathbf{z}\leftarrow\mathcal{A}(a)}}[\mathcal{L}(X,\mathbf{z})=1] \le e^{\varepsilon}\cdot \Pr_{\substack{X\sim\mathcal{D}\\ Y\sim\mathcal{D}\\ a\leftarrow\mathcal{M}(X)\\ \mathbf{z}\leftarrow\mathcal{A}(a)}}[\mathcal{L}(Y,\mathbf{z})=1]+\delta.$$

We can directly compare our Theorem B.2 to the Narcissus Resiliency bounds for $(\varepsilon, \delta)$-DP mechanisms. In Theorem F.1 of [18], they bound the RHS by $e^{\varepsilon}\cdot p+\delta$ where $p$ is the prior probability that predicate $\mathcal{L}$ is true. This is always strictly larger than our bound of $\frac{e^{\varepsilon}}{e^{\varepsilon}-1+1/p}+\delta$, especially when $p$ or $\varepsilon$ are relatively large. In addition, we give high probability bounds that hold even when conditioned on the mechanism output, which describes the dependence on the number of successful attacks and the privacy parameters. Therefore, while our bounds enjoy the same philosophical benefits that inspired the narcissus resiliency framework, our bounds on DP mechanisms are substantially tighter.

### G.3 Reconstruction Robustness (ReRo)

The reconstruction robustness (ReRo) game in [8, 31] can be viewed as a specific case of our setting when there is only a single target, $n = 1$, and a single reconstruction guess, $k = 1$, and the probability is taken over the coins of the mechanism as well as the sampling of the target record. That is, translating the pure DP version of their bound (Corollary 3 in [8]) into our notation, we get for $X' \sim \mathcal{D}$

$$\mathbb{E}_{a\sim\mathcal{M}(X')}\left[\Pr_{X\sim\mathcal{D}}[\mathcal{L}(X,R(a))=1] \mid \mathcal{M}(X)=a\right] \le e^{\varepsilon}\cdot \sup_{z\in Supp(\mathcal{D})}\Pr_{X\sim\mathcal{D}}[\mathcal{L}(X,z)=1], \quad (3)$$

where $\mathcal{L}(X,z)=\mathbb{I}\{\ell(X,z)\le\eta\}$ is some approximate reconstruction measure. In the ReRo game only one data record, $X$, is unknown while the rest are fixed and known to the adversary (one can imagine that $\mathcal{M}$ has the known records hardcoded).

Notice with these specific choices, the LHS is equal to $\Pr_{X\sim\mathcal{D},A\leftarrow\mathcal{M}(X)}(\ell(X,R(A))<\eta)$ and the supremum on the RHS is exactly what their prior, denoted by $\kappa_{\ell,\mathcal{D}}(\eta)$.

The ReRo bounds are a function of the prior success probability of the *heaviest* item in the target's data distribution, which may be very far from the prior success probability of a real attack. As Cohen et al. [18] point out, this can lead to cases where a sophisticated reconstruction of a relatively low-probability data record doesn't contradict the ReRo bound. Our bounds avoid this problem by computing the prior with respect to a specific attack attempt. In later work, Stock et al. [65] derived a similar RDP reconstruction bound which has the same limitations as ReRo.

**Distributional Reconstruction Robustness.** Cummings et al. [19] extend the pure-DP ReRo bounds to an attack setting with $n$ targets and one attack attempt ($k = 1$). While they capture settings where the adversary has substantially less information about the dataset, their bounds suffer from the same problems as ReRo. Namely, the prior is still dependent on the heaviest item (the optimal a priori attack) and the bounds only give an expectation over mechanism outputs.

### G.4 Additional Related Work

Our work fits into a substantial body of previous work [8, 14, 19, 31, 42, 46, 64, 65] that bounds the success probability of particular attacks based on a DP guarantee. These works consider several broad categories of attacks. *Membership inference (MI)* [62]—in which an attacker given (partial) knowledge of a target individual and of the population from which the data are drawn aims to determine if the target individual's records was in the input data set—is the attack category that most closely tracks the negation of DP's guarantees. *Attribute inference (AI)* [30] is similar but one aims to infer a specific sensitive attribute of the target, again given partial knowledge. *Reconstruction* [24] is the broadest class, encompassing multiple settings where the attacker infers one or more whole records. In Table 1 we compare our bounds to those of prior work.

While we focus here on works that aim to help interpret DP guarantees, many of the works we discuss also aim to check the quantitative tightness of the DP guarantee for a specific algorithm [46, 64].[11]

[11] Such assessment is often referred to as *auditing* [5, 23, 35, 46, 52, 64]. Mathematically, results on auditing overlap significantly with the type of bounds considered in this paper, but the goals and interpretation differ. Our comparison to previous work includes bounds motivated by both considerations.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in our abstract is that we provide a unified framework that captures DP guarantees against a broad family of attack settings and apply the bound to two empirical settings. These claims is substantiated by Section 3 and Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly address the mathematical and empirical limitations of our framework—including the requirement for independent priors (Section 3.1), decomposable metrics, sensitivity to prior estimation (Section 3.5), and the gap between theoretical bounds and current attacks—in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The text directs the reader to Appendices A,B,C, and D for the complete proofs of the theorems in this paper. All of the assumptions are stated in the theorem statements.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes all experimental details, models, datasets, and hyper-parameters have been disclosed under Appendix F.

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Since the exact models, datasets, and hyper-parameters have been disclosed in the main paper, we do not find a need to provide open access to our code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All models, datasets, and hyper-parameters have been disclosed in Section 4 and Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide analytical 95 percent confidence intervals on our bound and empirically calculated ones for our attacks.

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we disclose that all experiments were run on a single A100 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. Our research focuses on deriving theoretical privacy bounds and conducting empirical evaluations on publicly available datasets, and it fully conforms to the ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 5, we discuss how our risk framework enables safer context-specific DP calibration, alongside the negative risk that practitioners might over-rely on empirical upper bounds to justify dangerously high privacy parameters ($\epsilon$).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work is focused on theoretically quantifying privacy leakage of DP algorithms and does not pose a risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use three datasets: Enron emails (which is in the public domain), American Community Survey (which is in the public domain), and San Francisco Fire Department Calls for Service (which is licensed under the Open Data Commons Public Domain Dedication and License). Citations to the particular datasets are included where they are first referenced (Section 4 and Appendix F).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.