HOT PATE: PRIVATE AGGREGATION OF DISTRIBUTIONS FOR DIVERSE TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Private Aggregation of Teacher Ensembles (PATE) framework enables privacy-preserving machine learning by aggregating responses from disjoint subsets of sensitive data. Adaptations of PATE to tasks with inherent output diversity such as text generation, where the desired output is a sample from a distribution, face a core tension: as diversity increases, samples from different teachers are less likely to agree, but lower agreement results in reduced utility for the same privacy requirements. Yet suppressing diversity to artificially increase agreement is undesirable, as it distorts the output of the underlying model, and thus reduces output quality.

We propose Hot PATE, a variant of PATE designed for diverse generative settings. We formalize the notion of a *diversity-preserving ensemble sampler* and introduce an efficient sampler that provably transfers diversity without incurring additional privacy cost. Hot PATE requires only API access to proprietary models and can be used as a drop-in replacement for existing "cold" PATE samplers. Our empirical results corroborate the theoretical guarantees, showing that Hot PATE achieves orders-of-magnitude improvements in utility per privacy budget on in-context learning tasks.

1 INTRODUCTION

Generative models, and in particular large language models (LLMs), can perform a variety of tasks without explicit supervision (Radford et al., 2019; Brown et al., 2020). Unlike conventional machine learning models, generative models support open-ended tasks with inherently *diverse* outputs, where many different outputs may be appropriate. This diversity, which is often essential for functionality, is tunable via a temperature parameter, with higher temperatures yielding greater variation in outputs.

When training or performing analytics on sensitive data such as medical records, incident reports, or emails, privacy of individual data records must be protected. Mathematical frameworks for privacy guarantees include Differential privacy (DP) (Dwork et al., 2006), considered a gold standard, and k-anonymity and its extensions, which require that each released record be indistinguishable from at least $k-1$ others Sweeney (2002). In practice, many large-scale analyses (e.g., Anthropic's Clio and OpenAI's usage reports Tamkin et al. (2024); OpenAI (2025b) adopt lighter privacy notions based on minimum-support thresholds or suppression of low-frequency categories before releasing aggregates. Ultimately, these approaches all rely on *high agreement*, ensuring that reported outputs are supported by many data records.

A popular paradigm for privacy protection is the Private Aggregation of Teacher Ensembles (PATE) paradigm (Papernot et al., 2017; Bassily et al., 2018; Papernot et al., 2018), based on Nissim et al. (2007), and described as Framework 1.1. PATE partitions sensitive data among several teachers (each of which does *not* preserve privacy) and aggregates their predictions to obtain a privacy-preserving output. In the PATE framework, each data record affects at most one teacher and thus affects at most one vote in the histogram. A `NoisyArgMax` DP aggregation mechanism masks these small differences by adding noise to each c_j to obtain $(\tilde{c}_j)_{j\in V}$ and returning the index $\arg\max_j \tilde{c}_j$ (Duan et al., 2023). Different implementation vary in the noise distribution and privacy analyses (see discussion in Section E), but a label j can be returned only when the noise scale σ is small relative to its count c_j. A proxy for the privacy cost needed to get utility from a histogram is therefore the minimum count T of a relevant label.

Framework 1.1: Cold PATE

1. Partition the dataset D into n disjoint parts: $D = D_1 \sqcup \cdots \sqcup D_n$.
   For each $i \in [n]$, train a *teacher* model $\mathcal{A}_i$ on D_i.
2. For each example $x \in X$:
   - For each teacher $i \in [n]$, compute label prediction: $y_i := \mathcal{A}_i(x) \in V$.
   - Construct the histogram $\boldsymbol{c}$ of votes: for $j \in V$, $\quad c_j = \sum_{i \in [n]} \mathbb{1}\{y_i = j\}$.
   - Apply a privacy preserving aggregation mechanism to $\boldsymbol{c}$ to produce a final label $y \in V$. Abort if no confident agreement. Output y.

1.1 PATE IN THE DIVERSE SETTING

In *diverse settings*, such as those involving generative models, the underlying model outputs a probability distribution over V and returns a sample from the distribution. Such distributions are typically *diverse*, supporting open-ended responses with many plausible outcomes. In the corresponding PATE setup, each teacher $i \in [n]$ in the ensemble produces its own probability distribution $\boldsymbol{p}^{(i)}$. We formalize the aggregation step through an *ensemble sampler*: a mechanism that maps the set of teacher distributions $(\boldsymbol{p}^{(i)})_{i \in [n]}$ to an aggregate distribution $\mathcal{M}((\boldsymbol{p}^{(i)})_{i \in [n]})$, from which the final output is sampled.

As with basic PATE—referred to here as Cold PATE—the design goal of an ensemble sampler is to achieve a favorable privacy–utility trade-off. We define *basic utility* as the *yield*: output any relevant response. A more nuanced notion of utility is that the aggregate distribution *preserves diversity*: Informally (we formalize this notion below), when teachers have sufficient agreement on several plausible responses, the aggregate distribution should reflects this. Cold PATE was designed for classification tasks, where there is a single ground-truth label and knowledge transfer to the student is mediated by a set of non-sensitive unlabeled examples. In contrast, in generative settings the response distribution itself is the knowledge to be transferred. An ensemble sampler that preserves diversity enables this knowledge to be transferred to the student, simply by repeated sampling.

When cold PATE is applied in a diverse setting, the ensemble sampler first *samples a histogram*:

$$\mathbf{c} \sim \mathcal{H}_{\text{ind}}\big((\boldsymbol{p}^{(i)})_{i \in [n]}\big) \quad \stackrel{\text{def}}{=} \quad \Big(c_j = \sum_{i \in [n]} \mathbb{1}\{y_i = j\}\Big)_{j \in V} \quad \text{where } y_i \sim \boldsymbol{p}^{(i)} \text{ independently.} \tag{1}$$

This histogram step introduces an *inherent privacy–utility trade-off*: as output diversity increases, basic utility decreases sharply. For instance, if there are r equally good responses, the n teacher votes are split roughly evenly, so this requires threshold $T \approx n/r$, that is inversely proportion to the diversity. Moreover, the subsequent `NoisyArgMax` aggregation is not diversity-preserving: labels with higher probabilities are disproportionately more likely to be selected.

All prior work we are aware of on applying PATE in diverse settings (Tian et al., 2022; Duan et al., 2023; Wu et al., 2023) either relied on the Cold PATE ensemble sampler or implemented custom samplers that explicitly reduced or constrained diversity (see Section A for a detailed discussion). Moreover, these works evaluated only basic utility (yield) and did not consider the importance of transferring diversity from teachers to the aggregate distribution.

In this work, we ask a key question: is the observed diversity–privacy trade-off inherent? If not, can we design an ensemble sampler that achieves high utility under a fixed privacy budget, even in the presence of substantial diversity?

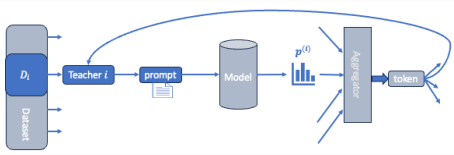
1.2 PATE FRAMEWORK FOR SEQUENTIAL TEXT GENERATION

A motivating application is the task of generating a representative set of synthetic, privacy-preserving data records from a collection of sensitive records. These records may include elements that are identifying and elements that are shared with many other records. A privacy-preserving generator must suppress the identifying components, while still capturing the shared structure and variability of the data. Crucially, it must also *preserve diversity*: without sufficient diversity, the synthetic set will underrepresent rare but valid patterns and fail to reflect the richness of the underlying distribution.

The resulting synthetic records can serve multiple downstream purposes: training a (possibly non-generative) student model, fine-tuning a generative model, or constructing privacy-preserving prompts for tasks.

Framework 1.2 presents a PATE design tailored to sequential text generation and suitable for tasks such as synthetic records generation, summarization, and querying. For our purposes, an *auto-regressive model* is a map $\mathcal{A} : V^* \mapsto \boldsymbol{p}$ that takes a sequence of tokens and outputs a next-token distribution over V. The framework is parametrized by a *model generator* $\mathcal{G}$ and an ensemble sampler $\mathcal{M}$. For each part of the data D_i, we train an autoregressive model $\mathcal{A}_i \leftarrow \mathcal{G}(D_i)$. Response generation then proceeds in lockstep: at each step, each teacher produces its next-token distribution, and the ensemble sampler is invoked to sample the next response token.

Framework 1.2: PATE for sequential text generation



Algorithm 1: PATE for Sequential Text Generation

Parameters: Vocabulary V; Instruction $\mathsf{C} \in V^*$; ensemble sampler $\mathcal{M} : \left(\boldsymbol{p}^{(i)}\right)_{i\in[n]} \mapsto \boldsymbol{p}$ over $V \cup \{\perp\}$; autoregressive model generator $\mathcal{G} : D \mapsto \mathcal{A}$
Input: Dataset D
Output: Response string $R \in V^*$

for $i \in [n]$ **do**
  Randomly partition D into disjoint subsets D_i
  $\mathcal{A}_i \leftarrow \mathcal{G}(D_i)$ // Construct teacher model from D_i
$R \leftarrow \{\}$ // Initialize empty response string
repeat
  for $i \in [n]$ **do** $\boldsymbol{p}^{(i)} \leftarrow \mathcal{A}_i(\mathsf{C} \cdot R)$ // Collect teachers' distributions over V
  $y \sim \mathcal{M}\left((\boldsymbol{p}^{(i)})_{i\in[n]}\right)$ // Aggregate and sample token
  if $y = \perp$ **then** use fallback to obtain y // exception
  $R \leftarrow R \cdot y$ // Append sampled token to response
until *termination condition met*
return R

The model generator abstraction captures two appealing ways to instantiating teachers: in-context learning and fine-tuning. With in-context learning, teacher $\mathcal{A}_i$ is specified by a context C_i constructed from data part D_i for *few shots* learning (Liu et al., 2021; Zhou et al., 2022; Garg et al., 2023). A key advantage of in-context learning is that each teacher is simply a prompt provided to a shared model, requiring no additional training or significant storage. Scaling the number of teachers is inexpensive: prompts are cheap, and the current OpenAI API supports 10^6 context+output tokens for roughly US$1 (OpenAI, 2025a). Thus, the primary bottleneck is the amount of available sensitive data, and larger ensembles are especially attractive since under DP composition, the number of queries allowable for a fixed privacy budget grows quadratically with the number of teachers. With fine-tuning, each teacher $\mathcal{A}_\rangle$ is a model that is fined-tuned on D_i. Parameter-efficient fine-tuning techniques (e.g., LoRA (Hu et al., 2022)) and managed services for fine-tuning proprietary models (OpenAI, 2023; Microsoft Azure, 2024; Anthropic, 2024) make this approach practical. Applying a PATE wrapper on top of such fine-tuned teachers is an appealing way to obtain privacy protection.

1.3 Overview of Contributions and Roadmap

Our primary contribution is the design of *hot*[1] ensemble samplers for PATE that deliver high utility, both in terms of yield and in terms of diversity preservation, in diverse settings. The method is beneficial not just with DP but for any agreement based privacy protection.

[1]The term 'hot' alludes to the temperature parameter that tunes diversity in LLMs.

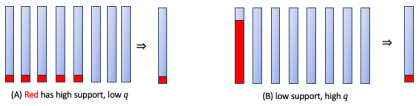


Figure 1: Illustration of two sets of probability distributions (each shown as a rectangle where the red segment indicates the probability of token j). In the left set, many teachers assign low probability q to token j; in the right, few teachers assign high probability q. The average probability of token j is the same in both cases, but the underlying support differs.

Cold PATE histogram counts are concentrated around the scaled average probabilities. With threshold T, a sampler can deliver basic utility only when some tokens have average probability over $\approx T/n$. We observe, as depicted in Figure 1, that averaging (and the cold PATE histograms) loses a critical distinction between high teachers' support with low probability q (which we can hope to transfer in a privacy-preserving manner, even when $q \ll T/n$) and low support with high q (which can not be transferred in a privacy-preserving manner). In Section 2 we present a nuanced definition of a robust aggregate distribution that captures this distinction. Informally, for a robustness parameter $\tau \in [n]$, there are two requirements:

- **Transfer requirement:** Any token that has probability at least $q > 0$ (no matter how small) across c teachers where $c \geq \tau$, is "transferred" in that it has probability $\Omega(qc/n)$ in the aggregate distribution.

- **Relevance requirement:** We do not transfer irrelevant tokens, that is, for any token j, its probability in the aggregate distribution is not much higher than its average probability in the teacher distributions.

In Section 3 we propose *coordinated ensembles*. We establish that for agreement threshold T, they facilitate ensemble samplers with $\tau \approx T$ robust aggregate distribution. A coordinated ensemble samples a shared randomness and based on that, each teacher i contributes a token y_i. The marginal distribution of each y_i is $\boldsymbol{p}^{(i)}$, same as with independent ensemble. But the difference is that teachers votes are maximally positively correlated. The frequency c_j of token j has high spread and in particular can (roughly) be $\Omega(\tau)$ with probability $\Omega(q)$. This property is what facilitates the robust transfer. With coordinated ensembles, two teachers with very diverse distributions that have a small total variation distance produce the same token with probability that depends on the distance. In particular, when the distributions are equal (the distance is 0), the same token would be produced. Since the ensemble sampler is histogram based, where each teacher contributes a single vote, any DP aggregation methods for histograms, including those used for PATE (Papernot et al., 2017; 2018), applies to histograms produced by coordinated ensembles. The large gains in utility of Hot PATE stem from the shape of the generated histograms: they are "peaky" with high agreement on few tokens and larger margin (see fig. 12 for a preview).

In Section E, we establish that ensemble samplers that are applied to coordinated histograms deliver the claimed utility privacy tradeoffs. We establish this for two regimes: *homogeneous* and *heterogeneous* ensembles. In the homogeneous setting, data records are randomly partitioned so that each teacher is representative—meaning they possess the core knowledge to be transferred. In this case, we have $\tau = \Omega(n)$ and use the `NoisyArgMax` aggregation. In the heterogeneous case, each teacher may represent one or a few users. In this regime, we need to allow for lower agreement and τ and we use an ensemble sampler that computes a *weighted sample* from the histogram.

In Section 4 we empirically demonstrate the properties and advantages of ensemble coordination for two in-context learning tasks: a natural task of generating synthetic data records from a dataset of sensitive ones and a curated task. We observe orders of magnitude improvement over the baseline of independent ensembles in terms of the privacy cost of achieving certain utility.

In Sections F and G we discuss data-dependent DP privacy analysis methods that can increase the number of queries processed for a given privacy budget by orders of magnitude over naive analysis. We benefit from a high *margin* – separation of the maximizer, which is more likely with coordinated ensembles, and make steps with no yield "free." With heterogeneous ensembles, teachers can be individually charged (instead of the whole ensemble) when they contribute to the final token (Hassidim et al., 2020; Cohen and Lyu, 2023). Related work is surveyed in appendix A.

2 DIVERSITY-PRESERVING AGGREGATION

We introduce a formal definition of when an aggregate distribution preserves diversity:

Definition 1 (Diversity-preserving aggregation of distributions)**.** Let $f(\boldsymbol{p}^{(i)})_{i\in[n]}) \mapsto \boldsymbol{P}$ map from n probability distributions over V to a probability distribution over $V \cup \{\perp\}$. We say that f is *diversity-preserving* with $\tau \in \mathbb{N}$, $\beta \in (0, 1]$, $\gamma \geq 1$ if for any input and $j \in V$

1. For all $q \in [0, 1]$, $(c_{j,q} := \sum_{i\in n} \mathbb{1}\{p_j^{(i)} \geq q\}) \geq \tau \implies P_j \geq \beta \cdot \frac{c_{j,q}}{n} q$.
2. $P_j \leq \gamma \frac{1}{n} \sum_{i\in[n]} p_j^{(i)}$.

The first property is that probability q across enough (τ) teachers, no matter how small is q, is transferred to the aggregate distribution. The second ensures that we do not output irrelevant tokens.

Requirements are stricter (and can be harder to satisfy) when β and γ are closer to 1 and when τ is smaller. A setting of $\tau = 1$ and $\beta = \gamma = 1$ allows only for the average distribution to be the aggregate. A larger τ increases robustness in that more teachers must support the transfer.

Remark 1 (failures)**.** *It is necessary to allow for $\perp$ (failure) in the support of the aggregate distribution when $\tau > 1$. For example, when the prompt instruction ask for a patient ID, and assuming no generalization, the teacher distributions have disjoint supports and no token can be returned. Failures in the generation can be addressed by: (i) Repeating the step with different shared randomness (ii) sample a token from a non-private default prompt or model, or (iii) redesign the prompt instruction.*

Remark 2 (Setting of τ)**.** Homogeneous ensembles *occur when data is randomly partitioned so that most teachers receive a representative part and possess the knowledge we wish to transfer. The goal is to transfer the parts of the distributions that are common to most teachers and $\tau > n/2$ suffices. In* heterogeneous ensembles*, each teacher might have data from one or very few "users." This arises when each teacher has small capacity (prompts currently have limited size of 8k-64k tokens (OpenAI, 2023)) or when by design each teacher is an agent of a single user. The goal here is to transfer parts of the distribution that are common to smaller subgroups of teachers and set $\tau \ll n$.*

3 ENSEMBLE COORDINATION

We propose *ensemble coordination* and establish that it facilitates privacy and diversity preserving aggregation. As with independent ensembles, we define a probability distribution $\mathcal{H}_{coo}\big((\boldsymbol{p}^{(i)})_{i\in[n]}\big)$ over histograms over V with total count $\sum_{j\in V} c_j = n$. The sampling of a histogram $\boldsymbol{c} := (c_j)_{j\in V}$ is described in Algorithm 2. The algorithm samples shared randomness $\rho := (u_j)_{j\in V}$. Each teacher $i \in [n]$ then contributes a single token $y_i \in V$ that is a function of its distribution $\boldsymbol{p}^{(i)}$ and ρ. The frequencies c_j are computed as in (1).

The sampling method in ensemble coordination is a classic technique called *coordinated sampling*. It was first introduced in statistics works in order to obtain samples that are stable under distribution shifts (Kish and Scott, 1971; Brewer et al., 1972; Saavedra, 1995; Rosén, 1997; Ohlsson, 2000) and in computer science works for computational efficiency via sampling-based sketches and a form of Locality Sensitive Hashing (LSH) (Cohen, 1994; 1997; Broder, 2000; Indyk and Motwani, 1998; Haas, 2011). Its recent applications include private learning (Ghazi et al., 2021) and speculative decoding (Leviathan et al., 2023).

Implementation `CoordinatedSamples` is simple to implement with access to the model. With proprietary models, an enhanced API can either (i) provide the shared randomness ρ to the model to facilitate token selection or (ii) give the full distribution to the user. Without API enhancements, the distribution can be approximated by repeated sampling with the same prompt. This impacts computation, as the number of samples needed increases with diversity, but does not impact privacy.

Algorithm 2: `CoordinatedSamples`

Input: Teacher distributions $(\boldsymbol{p}^{(i)})_{i\in[n]}$
foreach *token* $j \in V$ **do** sample i.i.d. $u_j \sim \mathsf{Exp}[1]$ // Sample shared randomness $\rho = (u_j)_{j\in V}$
foreach *teacher* i **do** // Compute coordinated samples $(y_i)_{i\in[n]}$
  $y_i \leftarrow \arg\max_j \frac{p_j^{(i)}}{u_j}$ // bottom-k sampling transform
foreach *token* $j \in V$ **do** // Compute frequencies
  $c_j \leftarrow \sum_{i\in[n]} \mathbb{1}\{y_i = j\}$
return $(c_j)_{j\in V}$, $\rho = (u_j)_j$ // Histogram of frequencies

3.1 Properties of coordinated histograms

Let $(\boldsymbol{p}^{(i)})_{i\in[n]}$ be probability distributions over V and let Y_{coo} and Y_{ind} be the respective distributions of votes $(y_i)_{i\in[n]}$ generated by a coordinated or independent ensemble with teacher distributions $(\boldsymbol{p}^{(i)})_{i\in[n]}$. Let $\mathcal{H}_{\text{coo}}$ and $\mathcal{H}_{\text{ind}}$ be the respective distributions of histograms.

For each token j, its expected frequency, over the sampling of histograms, is the same for coordinated and independent ensembles:

Claim 1 (Expected token frequency)**.**

$$\forall j \in V,\ \mathsf{E}_{\boldsymbol{c}\sim\mathcal{H}_{\text{coo}}}[c_j] = \mathsf{E}_{\boldsymbol{c}\sim\mathcal{H}_{\text{ind}}}[c_j] = \sum_i p_j^{(i)}\ . \tag{2}$$

Proof. The marginal distribution of y_i for teacher i is $\boldsymbol{p}^{(i)}$ with both independent and coordinated ensembles and thus the claim follows from linearity of expectation. □

In a coordinated ensemble, votes of different teachers are much more likely to agree than in an independent ensemble (see Section B for a proof):

Claim 2 (Agreement probability)**.** *For teachers $i, k \in [n]$ and token $j \in V$, the probability $\Pr_{\boldsymbol{y}\sim Y_{\text{coo}}}[y_i = y_k = j]$ that both samples agree on token j is*

$$\frac{\min\{p_j^{(i)}, p_j^{(k)}\}}{\sum_j \max\{p_j^{(i)}, p_j^{(k)}\}} \in \left[\frac{1}{2}, 1\right] \cdot \min\{p_j^{(i)}, p_j^{(k)}\}\ .$$

$$\Pr_{\boldsymbol{y}\sim Y_{\text{coo}}}[y_i = y_k = j] \geq \Pr_{\boldsymbol{y}\sim Y_{\text{ind}}}[y_i = y_k = j] = p_j^{(i)} \cdot p_j^{(k)}\ ,$$

with equality possible only when $\max\{p_j^{(i)}, p_j^{(k)}\} = 1$.

The key benefit of coordinated histograms is that we can generate a sample from a diversity-preserving aggregate distribution as in Definition 1 by exclusively considering tokens that appear with frequency at least $\tau/2$ in the histogram. In particular `NoisyArgMax`, which is not diversity preserving with independent ensembles, preserves diversity with coordinates ensembles (see appendix B for a proof):

Theorem 1 (Utility of Coordinated Ensembles)**.** *We can sample from an aggregate distribution that satisfies Definition 1 with parameters τ, $\beta = 0.34$ and $\gamma = 2$ by sampling a coordinated histogram $\boldsymbol{c} \sim H(Y_{\text{coo}})$ and only considering tokens j with $c_j \geq \tau/2$.*

3.2 Privacy properties

With both independent and coordinated ensembles, we aggregate the histogram in a privacy-preserving way to select a single token. While the distribution of the histograms produced by these ensemble types is very different, the privacy properties in terms of the divergence between neighboring datasets are identical and immediate:

Observation 1. *For every fixture of the shared randomness ρ, changing one of the distributions $\boldsymbol{p}^{(i)}$ given as input to Algorithm 2 changes at most one item of the resulting histogram. That is, letting H and H' denote the resulting histograms before and after the modification, we have that H, H' are at Hamming distance 2 (viewed as vectors in $\mathbb{N}^{|V|}$).*

The following corollary is immediate from Observation 1.

Corollary 1. *Let $\mathcal{A}$ be an algorithm whose input is a histogram $H \in \mathbb{N}^{|V|}$, such that for any two neighboring histograms H, H' (differing by at most one item) it holds that $\mathcal{A}(H) \approx_{(\varepsilon,\delta)} \mathcal{A}(H')$. Then the composed algorithm $\mathcal{A}$ (`CoordinatedSamples`$(\cdot)$) is (ε, δ)-differentially private.*[2]

Therefore, the same DP aggregation schemes and analyses used with independent ensembles apply off-the-shelf to coordinated ensembles. The benefit of coordinated ensembles, per Theorem 1, is in the *shape* of the histogram that results in better utility for same privacy.

4 Empirical demonstration for sequential text generation

We compare coordinated ensembles (Hot PATE) to a baseline of independent ensembles (Cold PATE) for sequential text generation as described in Framework 1.2. We evaluate both on a natural and a curated task. We use default temperature settings and took a few minutes on a single A100 GPU.

Evaluation metrics: In our evaluation, at a given generation step, corresponding to a set of contexts $\mathsf{C}_i \cdot R$ for $i \in [n]$, we sample $r = 10^3$ vote histograms $(\mathbf{c}^{(h)})_{h=1}^r$ from each of the coordinated and independent ensembles. Each histogram aggregates votes from n teachers, with each teacher contributing a single token. We denote by $c_j^{(h)}$ the count of token j in the hth histogram (for $h \in [r]$).

We use a threshold value $T \in [n]$ on token counts as a *streamlined proxy* for the *inverse privacy cost*: higher T implies proportionally lower cost. This is because a token can be reliably reported only when its count exceeds the scale of the noise introduced by the privacy-preserving mechanism (e.g., Gaussian or Laplace noise). We evaluate the utility of an ensemble type at a threshold value T using the following measures: (i) *transferred probability mass (coverage)*: $\frac{1}{r}\sum_{h=1}^{r}\sum_{j\in V} c_j^{(h)}\mathbf{1}\{c_j^{(h)} \geq T\}$, the fraction of total votes assigned to tokens with frequency at least T; (ii) *transferred support size*: $|\{j \in V : \max_{h\in[r]} c_j^{(h)} \geq T\}|$, the number of distinct tokens that appear above threshold in at least one histogram; and (iii) average *yield* per sample: $\frac{1}{r}\sum_{h=1}^{r}|\{j \in V : c_j^{(h)} \geq T\}|$, the average number of above-threshold tokens per histogram.

4.1 Natural task: synthetic instruction generation from a sensitive dataset of instructions

Dataset: We used **Dolly 15K** (Conover et al., 2023), a dataset of instructions and corresponding responses intended for training "chat" models like ChatGPT (in this work, we only use the instructions). We filter the dataset to include only instructions without a context that are shorter than 256 characters, resulting in a pool of about 10K examples of the original 15K.

Model and setup: To generate synthetic instructions, we use the pre-trained `Llama-3.1-8B` (lla, 2024) base model which is capable of in-context learning. Specifically, when we present this model with a few instructions as context, it consistently generates another instruction. The data was randomly partitioned to $n = 512$ teachers with initial contexts $(\mathsf{C}_i)_{i\in[n]}$ of 10 instructions. At each step of the generation, for a fixed partial response R, we sampled $r = 1000$ histograms. We discuss the results, additional results are reported in appendix C.

Gains in utility: Figure 2 reports the coverage and support-size of the transfer for two prefixes R. Coordinated ensembles attain high coverage and support even with $T = 0.5n$ whereas independent ensembles transfer no diversity, only one token, for the first prefix and fail to even have yield (return a relevant token) for the second prefix with $T > 0.17n$. This because independent ensembles can only transfer tokens when their average probability is $\gtrsim T/n$. Figure 3 (left) shows the distribution of the maximum count in the histogram: for prefixes with diverse next-token, independent ensembles require much lower T (high privacy cost) even for the basic utility of a yield.

[2] This corollary holds for all variants of differential privacy, and is written here with (ε, δ)-DP for concreteness.

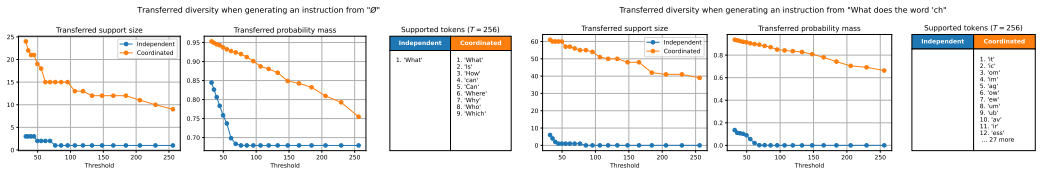


Figure 2: The transferred support-size and coverage per threshold T with coordinated and independent ensembles. Generating with prefixes $R = \emptyset$ (left) and $R =$"`What does the word 'ch`" (right).

Additional privacy analysis benefits: The privacy noise scale (proxied by the threshold T) is a "first order" indicator for the privacy cost with basic privacy analysis. The variety of data-dependent privacy analysis techniques (Dwork et al., 2006; Papernot et al., 2018; Cohen and Lyu, 2023) benefit by "not charging" for failed aggregations and "charging less" when there is a larger *margin* between the highest and second highest count. We demonstrate that coordinated ensembles reap more of these benefits as well. Figure 3 (middle) demonstrates that retries (with the same noise scale) are beneficial with coordinated ensembles, as the maximum count over several tries can be much larger than in a single try. With independent ensembles, counts concentrate around their expectations, and there is little benefits in retries. Additionally, fig. 3 (right) demonstrates large margins with coordinated ensembles. In independent ensembles, margins are smaller when diversity is higher as they simply reflect the difference in expected counts between the highest and second highest frequency in the average distribution. A large margin means that the output is much more *stable* which is a significant benefit with a refined privacy analysis Thakurta and Smith (2013); Bassily et al. (2018); Cohen and Lyu (2023) (see appendix C).

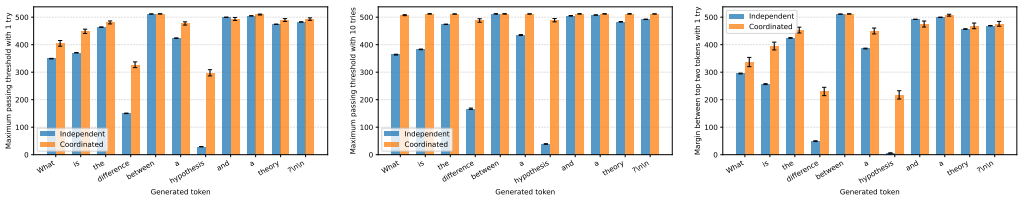


Figure 3: Maximum token count per histogram for different prefixes R (left: single attempt, middle: max in 10 attempts). Margin between highest and second highest counts (right).

4.2 Curated task

We designed a task for which (i) the pre-trained model has no prior exposure so that the "sensitive" context *must* be used for generating a good response, (ii) some mechanism is necessary for protecting the "private" information, and (iii) diversity is tunable. For simplicity, the task is designed to return a single token. We use the instruction-tuned `Llama-3-8B` (lla, 2024) (lla, 2024; AI@Meta, 2024) model.

Prompts: For each experiment we use $n = 10^4$ text prompts (teachers) of the form:

```
On planet Z, some numbers are edible.  <name> from planet Z eats
the following numbers for breakfast:  <random permutation of C      ∪
{<priv num>} > Give me an example breakfast number in planet Z.
Respond with just the number.
```

The fixed set C is a uniform sample of size $|C| = k$ from the set $\mathbb{N}_{100}^{999} = \{100, \ldots, 999\}$ of the 900 3-digit numbers. The strings <name> and <priv num> $\sim U[\mathbb{N}_{100}^{999} \setminus C]$ were generated separately for each prompt $i \in [n]$. For our purposes, the set C is the information we want transferred whereas the <name>, <priv num>, and the ordering of C in the prompt are prompt-specific and sensitive. Each prompt is designed to have $k + 1$ correct answers. We report results with $k \in \{20, 100\}$. `Llama-3-8B` uses a vocabulary V of 128k tokens and 3-digit numbers are encoded as single tokens. The distributions $\boldsymbol{p}^{(i)}$ exhibited biases towards certain numbers and high variation. The probability of returning a 3-digit number was 0.995 but the model generalized and

returned with 25% probability numbers outside the input set. Note that our goal is simply to reflect what the model does, including biases and generalizing. See appendix D.1 for further details.

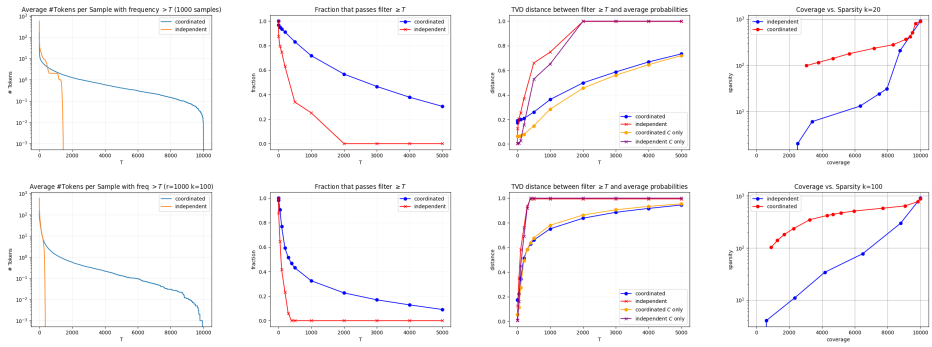


Figure 4: Left: Average yield per sample. Middle left: Coverage. Middle right: Total Variation Distance between transferred and average distribution; all as a function of T. Right: Coverage versus support-size with coordinated and independent ensembles, when sweeping the parameter T (not shown). Top: $k = 20$. Bottom: $k = 100$.

Utility Evaluation Figure 4 (left) shows the average yield per sample for varying T. Observe that with independent ensembles, the maximum frequency $\max_{h,j\in V} c_j^h$ (over histograms and tokens) corresponds to the maximum token average probability: for $k = 20$ it is $0.14n$ and for $k = 100$ it is $0.03n$. With coordinated ensembles, the majority of samples contained a token with frequency above $0.25n$ (that is much higher than the maximum token average probability). Figure 4 (middle right) reports the total variation distance from the average distribution and fig. 4 (middle left) reports coverage for varying T. We observe much higher coverage with coordinated ensembles compared with independent ensembles. Additionally, we observe that the coverage corresponds to the T/n-robust part of the distribution shown in fig. 9, that is, it corresponds to what we can hope to transfer (see Theorem 1 and Section D.2). For $k = 100$, we see 20% coverage with $T = 2000$ with coordinated sampling but we need $T \leq 250$ with independent sampling ($8\times$ in privacy budget). For $k = 20$, we see 40% coverage with $T = 4000$ with coordinated sampling but we need $T \leq 1000$ with independent sampling ($4\times$ in privacy budget). Moreover, independent samples have 0% coverage with $T \geq 1500$ for $k = 20$ and with $T \geq 400$ for $k = 100$ (when T/n exceeds the maximum average frequency) whereas coordinated ensembles are effective with high T. Figure 4 (right) shows a parametric plot (by threshold T, not shown) relating coverage and support size for coordinated and independent ensembles. Coordinated ensembles exhibit substantially greater diversity, achieving significantly larger support sizes at the same coverage levels, often with an order-of-magnitude gap compared to independent ensembles.

CONCLUSION

We introduced *Hot PATE*, an enhancement of the PATE framework that achieves significantly higher utility and effective diversity transfer for tasks with diverse outputs. We demonstrated orders-of-magnitude improvements over the baseline "cold" PATE in in-context learning scenarios, such as generating privacy-preserving synthetic data records from sensitive inputs.

Our core technical contribution is a formal notion of a robust, diversity-preserving aggregation of distributions, along with the proposal of *coordinated ensembles*, a method that enables both high utility and diversity transfer under the same privacy budget. Compared to cold PATE, which uses independent ensembles, coordinated ensembles produce voting histograms with properties more favorable to privacy analysis, including higher maximum counts and greater margins.

Finally, our design supports not only differential privacy but also lighter forms of protection that offer higher utility for tasks such as synthetic record generation. These relaxed goals include robustness to a small number of outliers and suppression of idiosyncratic subsequences—patterns that depend on one or a few examples and do not arise from generalization—while preserving diversity. Such protections can be achieved with greater utility by using fewer teachers, a lower robustness threshold, and omitting DP noise from the vote counts.

REFERENCES

The llama 3 model: A deep learning approach to language understanding. https://ai.meta.com/blog/meta-llama-3/, 2024.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Anthropic. Fine-tuning for claude 3 haiku in amazon bedrock, 2024. URL https://www.anthropic.com/news/fine-tune-claude-3-haiku-ga.

Raef Bassily, Om Thakkar, and Abhradeep Guha Thakurta. Model-agnostic private learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/aa97d584861474f4097cf13ccb5325da-Paper.pdf.

K. R. W. Brewer, L. J. Early, and S. F. Joyce. Selecting several samples from a single population. *Australian Journal of Statistics*, 14(3):231–239, 1972.

A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proc.of the 11th Annual Symposium on Combinatorial Pattern Matching*, volume 1848 of *LNCS*, pages 1–10. Springer, 2000.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Mark Bun, Thomas Steinke, and Jonathan Ullman. *Make Up Your Mind: The Price of Online Queries in Differential Privacy*, pages 1306–1325. 2017. doi: 10.1137/1.9781611974782.85. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974782.85.

Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *J. Mach. Learn. Res.*, 20:94:1–94:34, 2019. URL http://jmlr.org/papers/v20/18-549.html.

E. Cohen. Estimating the size of the transitive closure in linear time. In *Proc. 35th IEEE Annual Symposium on Foundations of Computer Science*, pages 190–200. IEEE, 1994.

E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.*, 55:441–453, 1997.

Edith Cohen and Xin Lyu. The target-charging technique for privacy accounting across interactive computations. *CoRR*, abs/2302.11044, 2023. doi: 10.48550/arXiv.2302.11044. URL https://doi.org/10.48550/arXiv.2302.11044.

Edith Cohen, Ofir Geri, Tamas Sarlos, and Uri Stemmer. Differentially private weighted sampling. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*. PMLR, 2021. URL https://proceedings.mlr.press/v130/cohen21b.html.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models, 2023.

David Durfee and Ryan M. Rogers. Practical differentially private top-k selection with pay-what-you-get composition. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3527–3537, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/b139e104214a08ae3f2ebcce149cdf6e-Abstract.html.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. On the complexity of differentially private data release: Efficient algorithms and hardness results. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 381–390, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585062. doi: 10.1145/1536414.1536467. URL https://doi.org/10.1145/1536414.1536467.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023.

Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20172–20184, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a89cf525e1d9f04d16ce31165e139a4b-Abstract.html.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM J. Comput.*, 41(6):1673–1693, 2012. URL https://doi.org/10.1137/09076828X.

Peter J. Haas. Sketches get sketchier. *Commun. ACM*, 54(8):100, 2011. doi: 10.1145/1978542.1978565. URL https://doi.org/10.1145/1978542.1978565.

Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Yossi Matias, and Uri Stemmer. Adversarially robust streaming algorithms via differential privacy. In *Annual Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annual ACM Symposium on Theory of Computing*, pages 604–613. ACM, 1998.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018. URL https://api.semanticscholar.org/CorpusID:53342261.

Haim Kaplan, Yishay Mansour, and Uri Stemmer. The sparse vector technique, revisited. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 2747–2776. PMLR, 2021. URL http://proceedings.mlr.press/v134/kaplan21a.html.

L. Kish and A. Scott. Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66(335):pp. 461–470, 1971. URL http://www.jstor.org/stable/2283509.

Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 171–180. ACM, 2009. doi: 10.1145/1526709.1526733. URL https://doi.org/10.1145/1526709.1526733.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 1: Images. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YEhQs8POIo.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *CoRR*, abs/2101.06804, 2021. URL https://arxiv.org/abs/2101.06804.

Yunhui Long, Boxin Wang, Zhuolin Yang, Bhavya Kailkhura, Aston Zhang, Carl Gunter, and Bo Li. G-pate: Scalable differentially private data generator via private aggregation of teacher discriminators. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2965–2977. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/171ae1bbb81475eb96287dd78565b38b-Paper.pdf.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pages 94–103. IEEE Computer Society, 2007. doi: 10.1109/FOCS.2007.41. URL https://doi.org/10.1109/FOCS.2007.41.

Microsoft Azure. Fine-tuning models with azure openai service, 2024. URL https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/fine-tuning-now-available-with-azure-openai-service/3954693.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

E. Ohlsson. Coordination of pps samples over time. In *The 2nd International Conference on Establishment Surveys*, pages 255–264. American Statistical Association, 2000.

OpenAI. OpenAI text completion API documentation, 2023. URL https://platform.openai.com/docs/api-reference/completions/create#logprobs.

OpenAI. Fine-tuning guide, 2023. URL https://platform.openai.com/docs/guides/fine-tuning.

OpenAI. Openai api pricing. https://openai.com/api/pricing, 2025a. Accessed: 2025.

OpenAI. How people use chatgpt: Usage analysis with aggregation thresholds, 2025b. URL https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf. Technical Report.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HkwoSDPgg.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rkZB1XbRZ.

Gang Qiao, Weijie J. Su, and Li Zhang. Oneshot differentially private top-k selection. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8672–8681. PMLR, 2021. URL http://proceedings.mlr.press/v139/qiao21b.html.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

B. Rosén. Asymptotic theory for order sampling. *J. Statistical Planning and Inference*, 62(2):135–158, 1997.

P. J. Saavedra. Fixed sample size pps approximations with a permanent random number. In *Proc. of the Section on Survey Research Methods*, pages 697–700, Alexandria, VA, 1995. American Statistical Association.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. doi: 10.1142/S0218488502001648.

Alex Tamkin, Miles McCain, Esin Durmus, Kunal Handa, et al. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*, 2024. URL https://arxiv.org/abs/2412.13678.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZtt0pRnOl.

Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 819–850, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Guha13.html.

Zhiliang Tian, Yingxiu Zhao, Ziyue Huang, Yu-Xiang Wang, Nevin L. Zhang, and He He. Seqpate: Differentially private text generation via knowledge distillation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11117–11130. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/480045ad846b44bf31441c1f1d9dd768-Paper-Conference.pdf.

Salil Vadhan. *The Complexity of Differential Privacy*. 04 2017. ISBN 978-3-319-57047-1. doi: 10.1007/978-3-319-57048-8_7.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models, 2023.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 2: Text. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=jnF53uXmBS.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. Teaching algorithmic reasoning via in-context learning, 2022.

A Related work

We place our contribution in the context of prior and independent concurrent works on PATE adaptations for text generation. These works either (i) did not consider diversity or (ii) recognized it and the importance of transferring it but proposed aggregation schemes where utility decreases with diversity together with methods to limit diversity as to mitigate this perceived privacy-diversity trade-off. Our technique of ensemble coordination can improve utility by replacing the respective component in some of these designs.

Tian et al. (2022) proposed a PATE extension for sequential text generation tasks in diverse settings. Their approach limited diversity: Average the teachers distributions and then truncate the tail by keeping only the top-k frequencies. The work of Tang et al. (2024) (independent concurrent) took a similar approach. The distribution of each teacher is reduced to a uniform distribution over its top-k token probabilities. An independent ensemble is then applied to this set of reduced distributions. This design limits diversity to k and modifies the distributions and still incurs the privacy-utility trade-off of independent ensembles.

Duan et al. (2023) explored adaptations of PATE for in-context learning via prompting, where each part D_i of the data is used to create a text prompt C_i. The ensemble is then used to label curated queries. But while some design elements were tailored to LLMs, the workflow and privacy analysis were identical to cold PATE (Papernot et al., 2018), and in particular, did not consider diverse responses.

Wu et al. (2023) (independent concurrent work) proposed approaches to private aggregation for in-context learning with diversity. They proposed to reduce the perceived diversity in sequentially-generated text outputs by different teachers by clustering together outputs that are semantically equivalent and aggregating each cluster in a semantic space. This essentially reduces the dimensionality of the output space. The aim then is to extract and transfer this common semantics in a privacy preserving way: Map responses into a common low dimensional embedding space and privately aggregate embedding vectors or identify frequent keywords in diverse teachers' responses. The limitations are that the approach only addresses same-semantics diversity and offers no solution for semantically-distinct diverse responses and are subjected to a privacy diversity trade-off. Additionally and importantly, they require hand crafted tools to map and curate responses back and forth from a semantic space. The added value of such a mapping approach, if combined with coordinated ensembles, depends on whether the reduction of diversity that is achieved is within or across teachers. The *across* variety (see Figure **??** (B)), where the knowledge of each teacher only contains one or limited variations of the same semantic, is not eliminated by ensemble coordination and thus there is added value by addressing it via other means. The *within* variety (see Figure **??** (A)) is handled effectively by ensemble coordination and can be transferred fluidly with no privacy loss and without the need for mitigation of diversity via additional engineering. We suspect that for the in-context learning use case, and for semantic similarity that can be captured by tools external to the model (such as an embedding), the diversity eliminated is anyhow encapsulated in the base model and thus present in most teacher distributions. That is, we expect the diversity to overwhelmingly be the "within" variety.

Lin et al. (2024); Xie et al. (2024) (independent concurrent work) proposed an approach called *private evolution* for generating synthetic examples from private examples. The design used heterogeneous teachers, where each

is a single private example. Initially, the base model is sampled to generate a collection of candidate (full) responses. The teachers then vote on candidates by nearest neighbor to their sensitive example in an embedding space. The next iteration then consist of a weighted sample from a privacy-preserving vote histogram. The resulting candidates are then used to generate a new set of candidates by the base model that are closer to the private distribution. This is repeated for multiple iterations. The inherent drawbacks of this approach, compared with sequential text generation, are that it is not suitable for transferring specific patterns (such as extension numbers for specific departments within an org) that are common in the private data but do not exist in the pre-training data and are not memorized by the model and can not be generalized by it. Additionally, it requires a number of candidates that is exponential in the intrinsic dimensionality of the candidate space. Therefore the realm of applications is different than Hot Pate and they are not directly comparable.

Papernot et al. (2017) (Appendix B.1) discussed using additional outputs (beyond just the noisy the maximizer) in the teachers' votes histogram for distillation tasks. They concluded that it is beneficial for utility but does not justify the privacy loss. Despite the superficial resemblance, this is very different from what we do as we capture diversity in the generation of the histogram where we "force" the teachers to agree but there is a distribution on the agreement token.

Finally, there are multiple innovative adaptations of PATE to non-categorical settings (aggregate vectors rather than labels) applied with generative models. The works we are aware of address different problems and use different techniques than Hot PATE. For example, image generation using generative adversarial networks (GAN): Jordon et al. (2018) proposed to train student discriminator using a cold-PATE like labeling approach. Long et al. (2021) proposed to train a student generator by aggregating the gradients produced by teachers discriminators. Notably, as with Hot PATE, this design does not require external generation of examples in order to facilitate transfer. Instead, it uses the built-in property of generators to produce examples from random strings.

B Properties of Coordinated Ensembles

Proof of Claim 2. The first statement in the claim follows from the denominator satisfying

$$1 \leq \sum_j \max\{p_j^{(i)}, p_j^{(k)}\} \leq 2 - \max\{p_j^{(i)}, p_j^{(k)}\} \leq 2\,. \tag{3}$$

The inequality follows using the more refined upper bound (3) on the denominator. □

The overall agreement probability of the two teachers (over all tokens) is the (weighted) Jaccard index (Jaccard, 1901) of the distributions:

$$\Pr_{\boldsymbol{y}\sim Y_{\text{coo}}}[y_i = y_k] = \frac{\sum_j \min\{p_j^{(i)}, p_j^{(k)}\}}{\sum_j \max\{p_j^{(i)}, p_j^{(k)}\}}\,.$$

In particular, when two teacher distributions are identical, the samples are the same

$$\boldsymbol{p}^{(i)} = \boldsymbol{p}^{(k)} \implies \Pr_{\boldsymbol{y}\sim Y_{\text{coo}}}[y_i = y_k] = 1.$$

We establish the claim in Theorem 1. We show that a token j for which m teachers i have $p_j^{(i)} > q$ has frequency at least $m/2$ with probability at least $0.34q$. This follows by substituting $p = 1/2$ in the following more general claim:[3]

Lemma 1 (diversity transfer)**.** *For any token j and $p, q \in [0, 1]$,*

$$\Pr_{\boldsymbol{c}\sim H(Y_{\text{coo}})}\left[c_j \geq \left\lfloor p \cdot \sum_{i\in n} \mathbf{1}\{p_j^{(i)} \geq q\}\right\rfloor\right] \geq \frac{1}{2}\ln(1/p)q\,.$$

Proof. Let i be such that $p_j^{(i)} \geq q$. Fix the sampled min value $x \sim \mathsf{Exp}[q]$ for q part of the probability of j. The distribution of the remaining part is $y \sim \mathsf{Exp}[1 - p_j^{(i)}]$ which is stochastically smaller than $\mathsf{Exp}[1 - q]$. We get that

$$\Pr[y_i = j] \geq \Pr_{y\sim\mathsf{Exp}[1-q]}[y > x] = e^{-x(1-q)}\,.$$

Fix $p \in [0, 1)$. It follows that the probability that $\Pr[y_i = j]$, conditioned on $x < \frac{-\ln p}{1-q}$ is at least $e^{-x(1-q)} \geq p$. The respective random variables y_i on different teachers that may share part of the distribution can only be

[3]The general statement allows for different tradeoffs between β and the threshold in Theorem 1

nonnegatively correlated. Therefore, if there are $c_{j,q}$ teachers with $p_j^{(i)} \geq q$ then the distribution of the number of teachers with $y_i = j$ is stochastically larger than $\mathsf{Bin}[e^{-x(1-q)}, c_{j,q}]$, which for any $x \leq \frac{-\ln p}{1-q}$ is stochastically larger than $\mathsf{Bin}[p, c_{j,q}]$. The median of the Binomial distribution $\mathsf{Bin}[p, c_{j,q}]$ with probability at least $1/2$ is larger than $\lfloor pc_{j,q} \rfloor$. Therefore, with this conditioning on x, there are at least $\lfloor pc_{j,q} \rfloor$ teachers with $y_i = j$.

$$\Pr_{(y_i)_{i\in[n]} | x < \frac{-\ln p}{1-q}} [c_j \geq \lfloor pc_{j,q} \rfloor] \geq 1/2 \,. \tag{4}$$

The event $x < \frac{-\ln p}{1-q}$ occurs with probability at least

$$\Pr_{x\sim\mathsf{Exp}[q]} [x < \frac{-\ln p}{1-q}] = 1 - e^{(\ln p)q/(1-q)} \geq -(\ln p)q \,.$$

Combining with (4), we obtain the claim in the statement of the Lemma. □

To establish relevance we show that high frequency must have a "backing." The following is immediate from (2) and Markov's inequality (and is tight in the sense that for any T there are distributions where equality holds):

Lemma 2 (relevance)**.** *For any token j and T,*

$$\Pr_{\mathbf{c}\sim H(Y_{\mathrm{coo}})} [c_j \geq T] \leq \frac{1}{T} \sum_{i\in[n]} p_j^{(i)} \,.$$

C Further details for the instruction generation demonstration

Diversity transfer: Diversity transfer with coordinated and independent ensembles for additional prefixes R are reported in fig. 5. We observe that with coordinated ensembles, more of the probability mass is transferred and it is much more diverse.

Maximum count: Figure 6 (left) shows the distribution of the maximum count for additional prefixes; (right) shows the maximum count over 10 tries (histograms generated with different samplings of shared randomness). We observe that with coordinated ensembles, the maximum token count is consistently at or above $0.6n$ with one try and above $0.9n$ for the maximum over 10 tries. In particular, there is significant benefit to repetitions. As for independent ensembles, we observe that when there is high diversity (many appropriate choices for the next-token), the maximum count is frequently below $0.2n$ and there is nearly no benefits for retries. As explained, the noise scale of the DP aggregation depends linearly in this maximum count. This means that even with basic privacy analysis (which does not benefit from margin), coordinated ensembles require over 4 times the number of teachers (and data) for the *basic utility* of producing an instruction. As demonstrated, the produced instruction by independent ensembles would also be much less diverse. Furthermore, by using privacy accounting with `BetweenThresholds` (Cohen and Lyu, 2023; Bun et al., 2017) we can generate a number of tokens that is exponential in the number of teachers when histograms are such that the maximum count is either very high (say above $0.6n$) or very low (say below $0.4n$).

Margin: The vote histograms generated by coordinated ensembles benefit not only a higher *maximum* count but also from a high *margin* between the highest count and second highest count tokens. Additional results that show the size of the margin between the highest and second highest counts in the histogram are reported in fig. 6. We observed a margin that is consistently above $0.4n$, where n is the number of teachers, with coordinated ensembles whereas a very small margin occurs frequently with independent ensembles.

Benefits of high margin: We explain how high margins are leveraged in data dependent data analysis using the techniques of Bassily et al. (2018).[4] Similar benefits are reaped via other methods such as (Cohen and Lyu, 2023). Informally, their technique is based on a coupling argument between the *distance to instability* framework of Thakurta and Smith (2013) and the *sparse vector* technique of Dwork et al. (2009). More specifically, the algorithm of Bassily et al. (2018) uses the sparse vector technique in order to continuously verify that the number of "unstable queries" seen so far does not cross some predefine threshold k; and uses the distance to instability framework to answer queries as long as the number of unstable queries is indeed below k. If we assume, as is supposed by our experiments, that the margin in our algorithm is consistently above $\eta\, n$ (in our experiments we observed $\eta = 0.4$), then it suffices to assert that $\eta\, n \geq \frac{32\sqrt{2}}{\varepsilon} \log\left(\frac{4m}{\delta}\right) \sqrt{\log\left(\frac{2}{\delta}\right)}$ in order to generate m tokens while satisfying (ε, δ)-DP. This means that (with high margin histograms) the number of tokens generated for given privacy parameters increases *exponentially* with the number of teachers. This can be contrasted with only a quadratic increase with the number of teachers obtained using standard analysis with advanced composition.

[4] See Algorithm 3 in Bassily et al. (2018).

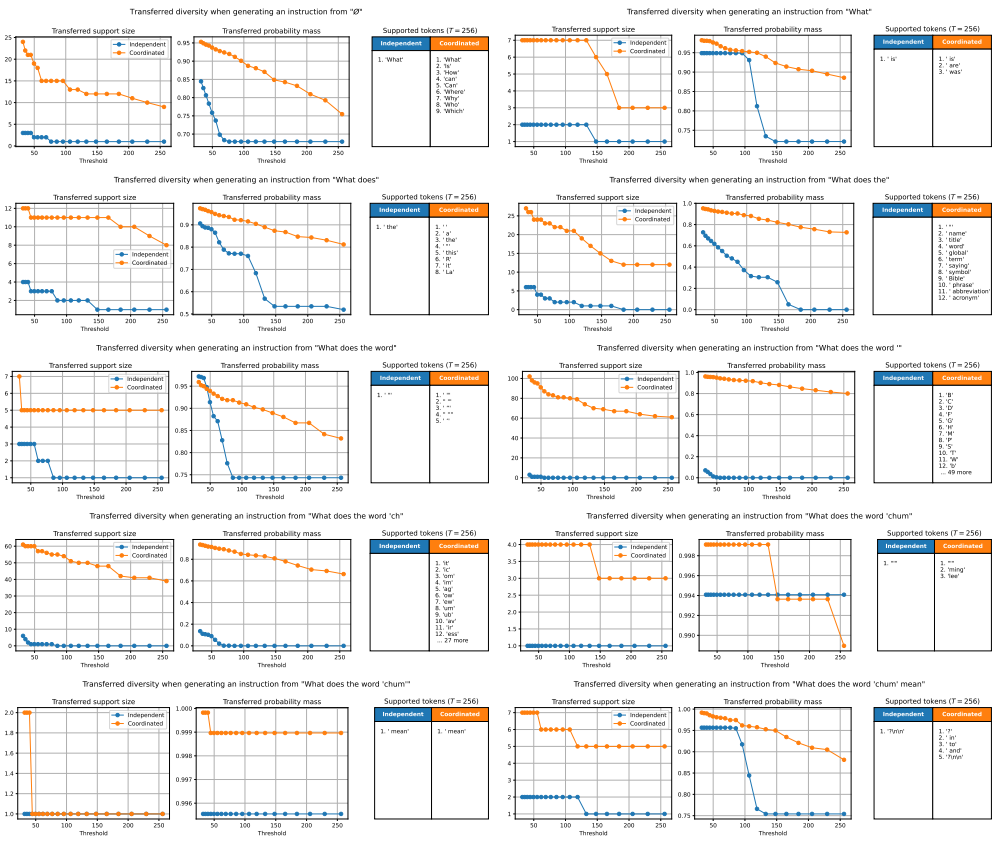


Figure 5: The transferred support-size and coverage per threshold T with coordinated and independent ensembles, when generating a synthetic instruction. For multiple prefixes R.

D Further details on Planet Z Demonstration

D.1 Properties of the Generated Distributions

The distributions deviated from the "intended" one of a uniform distribution over the numbers in the prompt: The model exhibited bias towards certain numbers, had spurious dependencies on private components, and generalized. Note that our evaluation focuses on the effectiveness of transferring the *knowledge of the model*, as reflected in its generated response distributions, including its biases and generalizations. We observed the following:

- The probability assigned by the model to tokens that are not 3-digit numbers is negligible: The average probability (over teachers) of a response token in $\mathbb{N}_{100}^{999}$ was $\mathsf{E}_{i\in[n]} \sum_{j\in\mathbb{N}_{100}^{999}} p_j^i \approx 0.997$ for $k = 20$ and ≈ 0.994 for $k = 100$.

- Tokens in C dominate but other 3-digit numbers are likely: The average probability of a token in C was $\mathsf{E}_{i\in[n]} \sum_{j\in C} p_j^i \approx 0.716$ ($k = 20$ tokens) and ≈ 0.75 ($k = 100$). Recall that only one in k numbers in the prompt was in $\mathbb{N}_{100}^{999} \setminus C$, therefore the probability of 25%+ assigned to these tokens is explained by the model generalizing that additional 3-digit numbers are edible on Planet Z.

- Despite symmetric prompt construction, there is significant variability in the average probability of different tokens in C and in the probability across teachers of the same token. This is an artifact of the model. Figure 8 reports the average (over prompts) of the probability of each token and demonstrates variability between tokens. The error bars indicate variability in the token probability across teachers.

Figure 6: Maximum token count per next-token vote histogram for different prefixes R in a single attempt (left) and in 10 attempts (right)

D.2 Quantifying how much is Transferable

Remark 3 (Robust Average)**.** *We use the τ-robust part of the average of the teachers distributions as an indicative upper bound on the part that is privately transferrable:*

$$P_j(\tau) := \frac{1}{n} \sum_{i \in [n]} \min \left\{ p_j^{(i)}, (\{p_j^{(h)}\}_{h \in [n]})_{(\tau)} \right\} \text{ for } j \in V \tag{5}$$

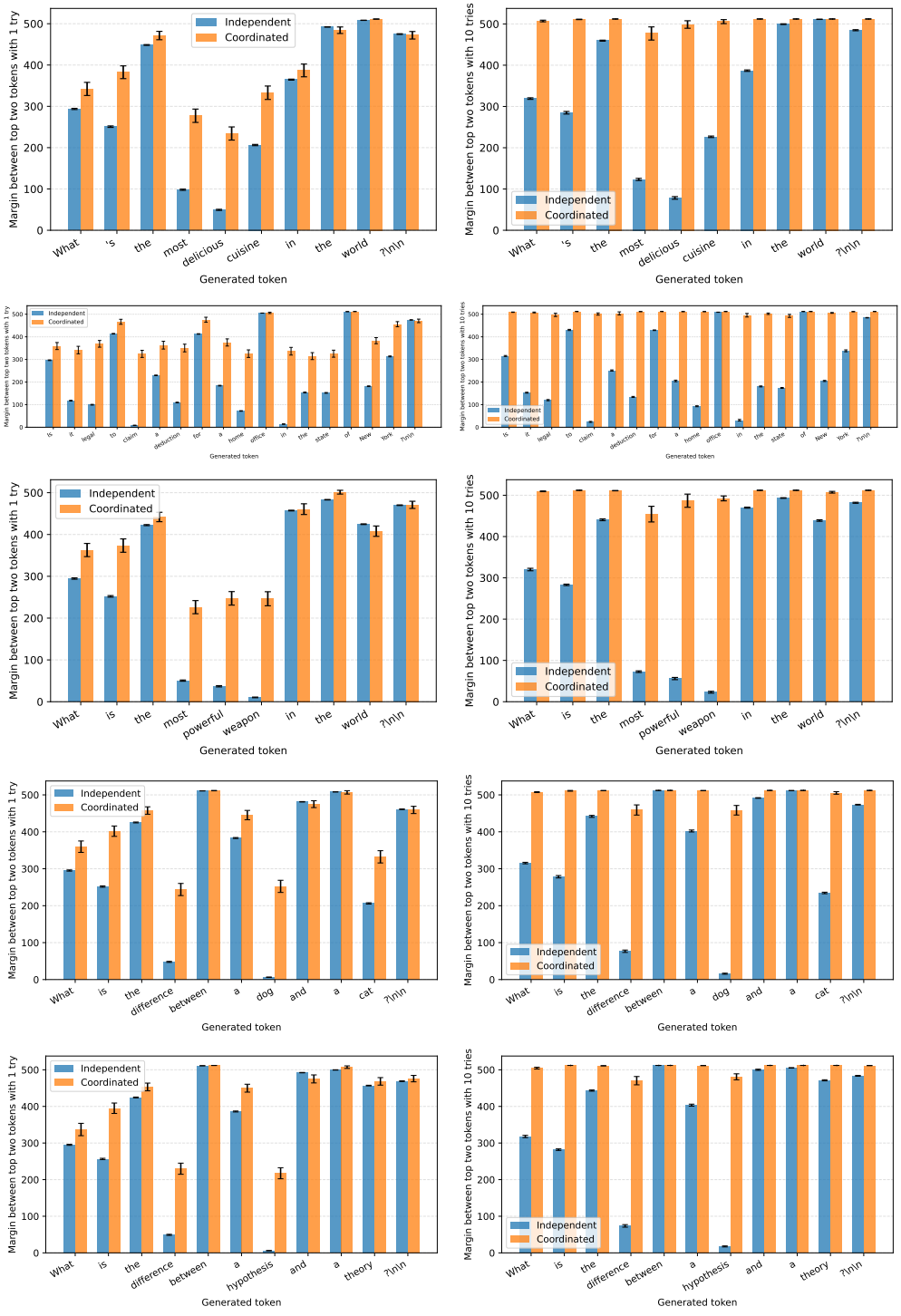


Figure 7: Margin between highest and second highest counts per histogram. A single try (left) and largest of 10 tries (right).

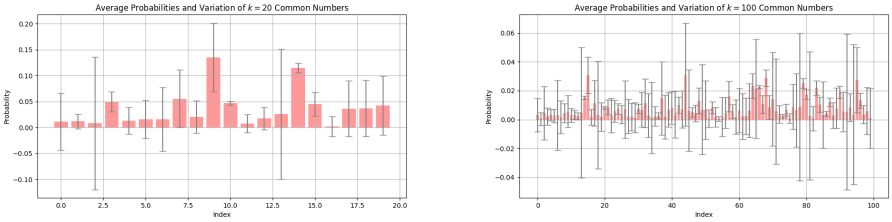


Figure 8: Average probability, over teachers, of the k tokens in C (left is $k = 20$, right is $k = 100$). The error bars indicate the contribution of the token to the average total variation distance over pairs of teacher distributions.

where $(\{p_j^{(h)}\}_{h\in[n]})_{(\tau)}$ *is the* τ*th largest probability of token* j *in a teacher distribution. Note that* $(P_j(1))_{j\in V}$ *is the average distribution and the values are non-increasing with* τ*. The* τ-robust probability mass*, defined as* $P(\tau) := \sum_{j\in V} P_j(\tau) \leq 1$*, upper bounds the transferrable probability mass. The complement* $1 - P(\tau)$ *is indicative lower bound on the probability of* $\perp$ *in the robust aggregate.*

Figure 9 reports the τ-robust fraction of the average distribution for varying τ (see Remark 3). This is the part of the average distribution that we can hope to transfer via coordinated ensembles with support τ. Recall that variability in the same token among teachers decreases transferability whereas variability among tokens does not.

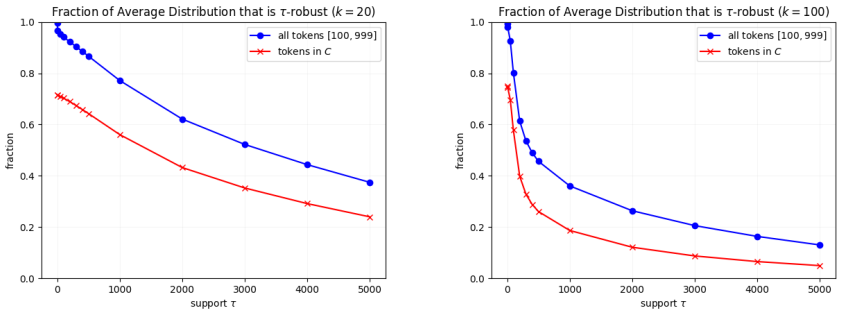


Figure 9: The τ-robust part of the distribution for varying τ (see Remark 3). Left is $k = 20$ right is $k = 100$.

D.3 Independent versus Coordinated Histograms

Figures 10 and 11 visualize the average probability $\frac{1}{n}\sum_{i\in[n]} p_j^{(i)}$ of each token $j \in \mathbb{N}_{100}^{999}$ across teacher distributions and the average frequency $\frac{1}{r}\sum_{h=1}^{r} c_j^h$ over the $r = 10^3$ samples from each of independent and coordinated ensembles. This demonstrates the property in claim 1 that the expected number of votes for each token is the same for the two ensemble types and corresponds to the average distribution. The qualitative difference between coordinated and independent ensembles (see claim 2) is visualized in Figure 12 which zooms on individual sampled histograms, showing one for independent sampling and two for coordinated sampling. With independent sampling, frequency counts of each token j are concentrated close to the expectation $\sum_i p_j^{(i)}$ and are similar across different samples and to the averages shown in Figures 10 and 11. With coordinated ensembles there is high variability in the shape of different samples and it is possible for the frequency of a token to far exceed the average value $\sum_i p_j^i$.

D.4 Visualized histograms of transferred mass

Figures 13 and 14 visualize the histograms of the covered votes (averaged over the r samples) per token, for varying thresholds T. For each T we list coverage and support size. We can see that independent ensembles become ineffective with very low T, when T/n exceeds the maximum average frequency of a token (0.14 with $k = 20$ and 0.03 with $k = 100$), and transfer support-size is effectively limited to tokens with frequency at least T/n. In particular, no generalization (shown in blue) is transferred. In contrast, coordinated ensembles are effective also when $T > 0.2n$ and transfer larger support size.

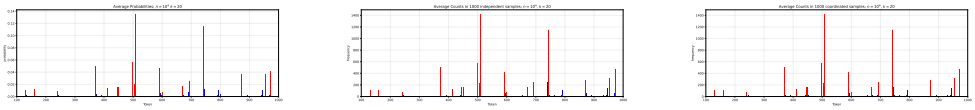


Figure 10: $k = 20$: For all tokens (tokens in C shown in read): Average probability over teachers (left). Average frequency of $r = 1000$ samples using independent (middle) and coordinated (right) ensembles.

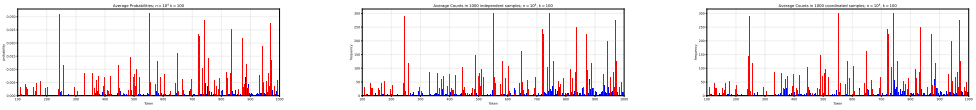


Figure 11: $k = 100$: For all tokens (tokens in C shown in read): For all tokens (tokens in C shown in read): Average probability over teachers (left). Average frequency of $r = 1000$ samples using independent (middle) and coordinated (right) ensembles.

E AGGREGATION METHODS OF FREQUENCY HISTOGRAMS

Our aggregation methods are applied to frequency histograms generated by a coordinated ensemble and return a token or $\perp$. We propose two meta schemes that preserves diversity in the sense of Definition 1: One for homogeneous ensembles, where we use $\tau > n/2$, in Section E.1 and one for heterogeneous ensembles, where $\tau \ll n/2$ (but large enough to allow for DP aggregation), in Section E.2. To establish diversity preservation, we consider the end-to-end process from the teacher distributions to the aggregate distribution. To establish privacy, it suffices to consider the histogram in isolation, as it has the same sensitivity as vote histograms with cold PATE: When one teacher distribution changes, one token can gain a vote and one token can lose a vote. Noting that the shared randomness ρ is considered "public" data. We then explore (Sections F and G) DP implementations that admit data-dependent privacy analysis so effectively many more queries can be performed for the same privacy budget. We can avoid privacy loss on responses that agree with the prior distribution of the public model with a public prompt. We can benefit from the particular structure of histograms generated by coordinated ensembles. The privacy loss does not depend on queries with no yield, with high agreement, or with agreement with a public prior. With heterogeneous ensembles we can also gain from individualized per-teacher privacy charging.

E.1 HOMOGENEOUS ENSEMBLES

Algorithm 3: `DistAgg` homogeneous

$\boldsymbol{c}, \rho \leftarrow$ `CoordinatedSamples`$((\boldsymbol{p}^{(i)})_{i\in[n]})$ `// Algorithm 2`
$(j, \hat{c}_j) \leftarrow$ `NoisyArgMax`$_L(\boldsymbol{c})$ `// DP noisy maximizer with error` L
if $\hat{c}_j > (n/2 + L)$ **then return** j **else return** $\perp$

When $\tau > n/2$, there can be at most one token j with frequency $c_j \geq \tau$. If there is such a token, we aim to report it. Otherwise, we return $\perp$. Our scheme is described in Algorithm 3 in terms of a noisy maximizer (`NoisyArgMax`$_L$) procedure. The latter is a well studied construct in differential privacy (McSherry and Talwar, 2007; Durfee and Rogers, 2019; Qiao et al., 2021). Generally, methods vary with the choice of noise distribution and there is a (high probability) additive error bound L that depends on the privacy parameters and in some cases also on the support size and confidence. For our purposes, we abstract this as `NoisyArgMax`$_L$ that is applied to a frequency histogram $\boldsymbol{c}$ and returns $(j, \hat{c}_j)$ such that $|c_j - \hat{c}_j| < L$ and $\max_{h\in V} c_h - c_j \leq 2L$. We show that the method is diversity preserving:

Lemma 3 (Diversity-preservation of Algorithm 3)**.** *For $\mu > 1$, Algorithm 3, instantiated with `NoisyArgMax`$_L$ as described, is diversity preserving in the sense of Definition 1 with $\tau = \mu(n/2 + 2L)$, $\beta = \ln(\mu)/2$ and $\gamma = 2$.*

Proof. We apply Lemma 1 with $p = 1/\mu$. We obtain that the token j has frequency at least $c_j \geq n/2 + 2L$ with probability at least $0.5\ln(\mu)q$. Therefore we have $\hat{c}_j \geq n/2 + L$ with probability at least $0.5\ln(\mu)q$. Note that a token can only be reported if its frequency is $c_j > n/2$. Using $T = n/2$ in Lemma 2 we obtain that the relevance requirement is satisfied with $\gamma = 2$. □

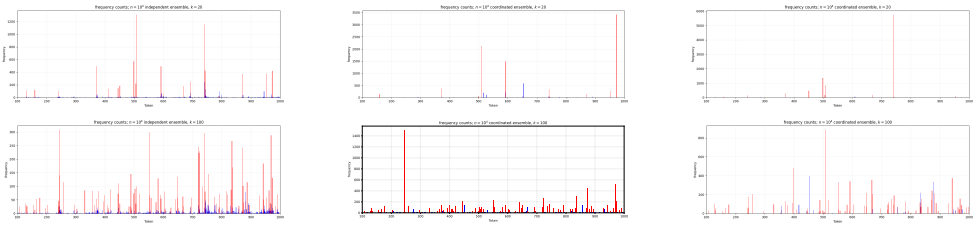


Figure 12: Frequency counts per token in individual sampled histograms. Left: Independent ensemble. Middle and Right: Coordinated ensemble. Top $k = 20$ bottom $k = 100$.

$T = 100$
coverage: coo: 94% ind: 75%
support: coo: 417 ind: 24

$T = 200$
coverage: coo: 91% ind: 63%
support: coo: 365 ind: 13

$T = 500$
coverage: coo: 83% ind: 34%
support: coo: 277 ind: 6

$T = 1000$
coverage: coo: 72% ind: 25%
support: coo: 233 ind: 2

$T = 2000$
coverage: coo: 57% ind: 0%
support: coo: 178 ind: 0

$T = 5000$
coverage: coo: 30% ind: 0%
support: coo: 98 ind: 0

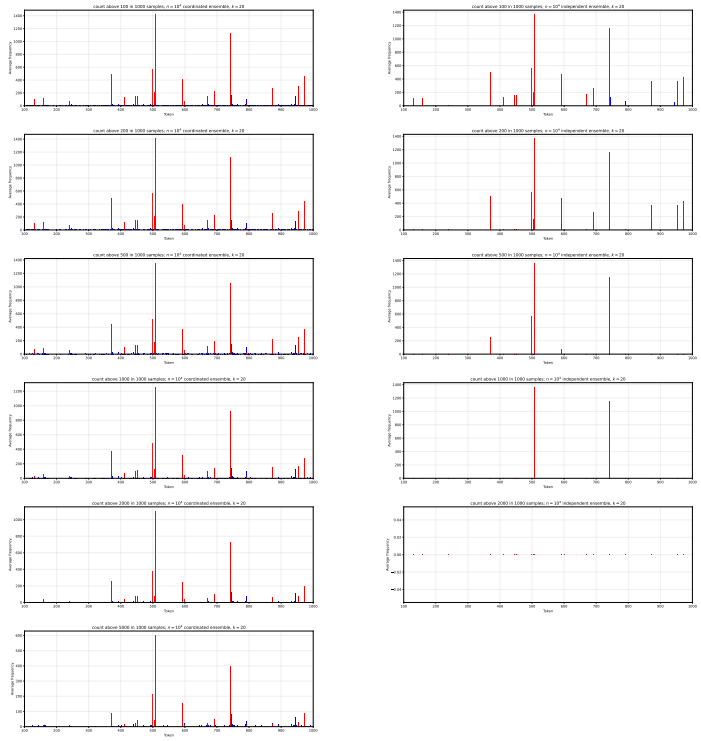


Figure 13: Coverage histograms averaged over $r = 10^3$ samples. Filter $T \in [100, 200, 500, 1000, 2000, 5000]$. $k = 20$. Left: Coordinated. Right: Independent.

The two most common noise distributions for DP are Gaussian and Laplace noise. (Cold) PATE was studied with both. The Gaussian-noise based Confident-GNMax aggregator (Papernot et al., 2018; Duan et al., 2023) empirically outperformed the Laplace-based LNMAX (Papernot et al., 2017) on cold PATE. The advantages of Gaussian noise are concentration (less noise to separate a maximizer from low frequency tokens) and efficient composition. and more effective data dependent privacy analysis. Laplace-based noise on the other hand can benefit from sparsity of the histogram (with approximate DP), a consideration as the key space of tokens or strings of token can be quite large, there is an optimized mechanism with weighted sampling. Both benefit from data dependent privacy analysis that benefits from consistently large maximum counts or large margins using tools such as (Cohen and Lyu, 2023). Our privacy analysis in Section F uses a data-dependent Laplace-based approach.

E.2 Heterogeneous Ensembles

For lower values of τ, we propose the meta-scheme described in Algorithm 4: We perform weighted sampling of a token from $\boldsymbol{c}$ and return it if its count exceeds $2L$. If it is below $2L$ we may return either j or $\perp$. We propose DP implementations in Section G. We establish that Algorithm 4 is diversity-preserving:

Lemma 4 (Diversity-preservation of Algorithm 4)**.** *For $\mu > 1$, Algorithm 4 is diversity preserving in the sense of Definition 1 with $\tau = \mu 2L$, $\beta = \frac{1}{2\mu}\ln(\mu)$ and $\gamma = 1$.*

$T = 50$
coverage: coo: 90% ind: 65%
support: coo: 638 ind: 77

$T = 100$
coverage: coo: 77% ind: 42%
support: coo: 576 ind: 34

$T = 200$
coverage: coo: 59% ind: 23%
support: coo: 509 ind: 11

$T = 300$
coverage: coo: 52% ind: 6%
support: coo: 468 ind: 4

$T = 400$
coverage: coo: 47% ind: 0%
support: coo: 441 ind: 0

$T = 1000$
coverage: coo: 33% ind: 0%
support: coo: 347 ind: 0

$T = 2000$
coverage: coo: 23% ind: 0%
support: coo: 234 ind: 0

$T = 5000$
coverage: coo: 9% ind: 0%
support: coo: 102 ind: 0

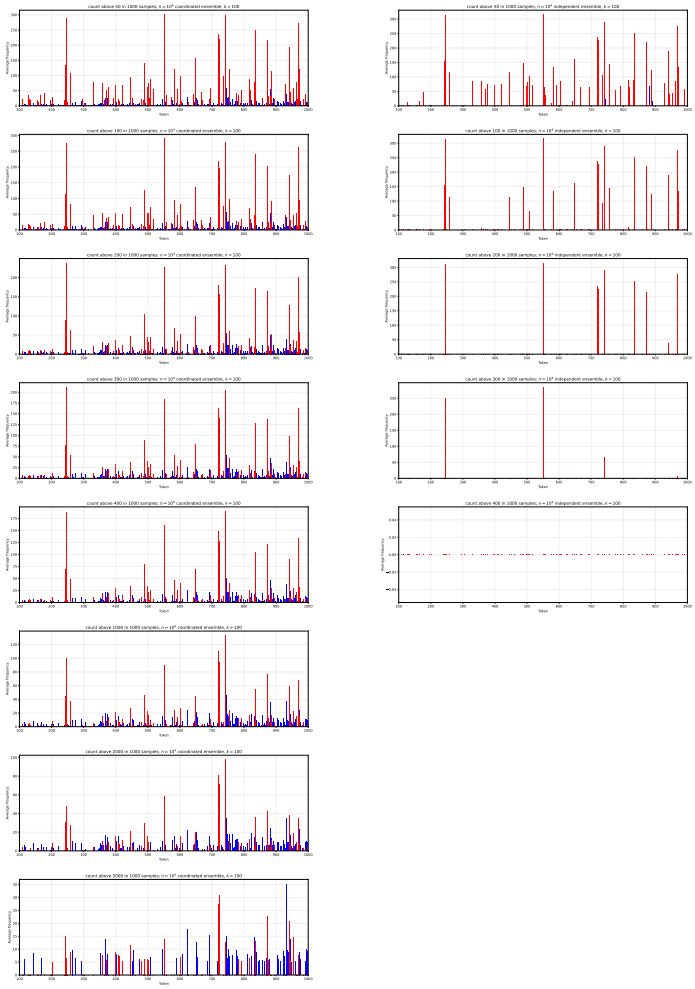


Figure 14: Average of $r = 10^3$ sampled histograms for filter $T \in [50, 100, 200, 300, 400, 1000, 2000, 5000]$. $k = 100$ Left: coordinated Right: Independent

Algorithm 4: `DistAgg` Heterogeneous

$\boldsymbol{c}, \rho \leftarrow$ `CoordinatedSamples`$((\boldsymbol{p}^{(i)})_{i\in[n]})$ `// Algorithm 2`
Sample $j \in V$ with probability $\frac{c_j}{n}$ `// Weighted sampling of a token from` $\boldsymbol{c}$
if $c_j \geq 2L$ **then return** j **else return** j *or* $\perp$

Proof. Consider the first requirement of Definition 1. Consider a token j with $c_{j,q} \geq \tau$. From Lemma 1 using $p = 1/\mu$ we obtain that the token j has frequency at least $c_j \geq c_{j,q}/\mu \geq 2L$ with probability at least $0.5\ln(\mu)q$. The token is sampled with probability $\min\{1, kc_j/n\}$ and if so appears also in $\boldsymbol{c}^*$ (since $c_j \geq 2L$). The expected size (number of entries) of $\boldsymbol{c}^*$ is at most k and thus it is returned if sampled with probability at least $1/k$. Overall it is sampled and reported with probability at least $\min\{1/k, c_j/n\}$. In total, the probability is $P_j \geq \min\{1/k, c_{j,q}/(\mu n)\}0.5\ln(\mu)q \geq \frac{1}{2k\mu}\ln(\mu)\frac{c_{j,q}}{n}q$.

The second requirement of Definition 1 is immediate. The expected frequency of token j is $\sum_{i\in[n]} p_j^{(i)}$ and it is sampled with probability at most $\frac{k}{n}\sum_{i\in[n]} p_j^{(i)}$. It can only be the output if sampled. □

F PRIVACY ANALYSIS CONSIDERATIONS

The effectiveness of Hot PATE depends on the number of queries with yield (token returned) that can be returned for a given privacy budget. In this section we explore the benefits of data-dependent privacy analysis when the aggregation follows Algorithm 3 (homogeneous ensembles). We use synthetically generated teacher distributions with varying size common component (that can be arbitrarily diverse) and distinct (private) components.

Broadly speaking, with data-dependent analysis, we incur privacy loss on "borderline" queries where the output of the DP aggregation has two or more likely outputs. Queries that return a particular token with high probability or return $\perp$ with high probability incur little privacy loss.

We demonstrate that with Algorithm 3, we can expect that only a small fraction of frequency histograms generated by coordinated ensembles are "borderline." (i) For queries with high *yield* (high probability of returning a token over the sampling of the shared randomness), the generated histograms tend to have a dominant token (and thus lower privacy loss). This because coordinated ensembles tend to "break ties" between tokens. (ii) For queries with low yield (high probability of $\perp$ response and low probability of returning a token), the total privacy loss only depends on yield responses. This means that high $\perp$ probability does not cause performance to deteriorate.

This is important because both these regimes are likely in sequential text generation and with coordinated ensembles. We expect many of the tokens to follow the base model distribution and therefore have high agreement and not incur privacy loss. Or alternatively, instructions that require private data have no agreement and return $\perp$. The dependent privacy analysis means that generally we can process many more queries for the privacy budget than if we had just used a DP composition bound.

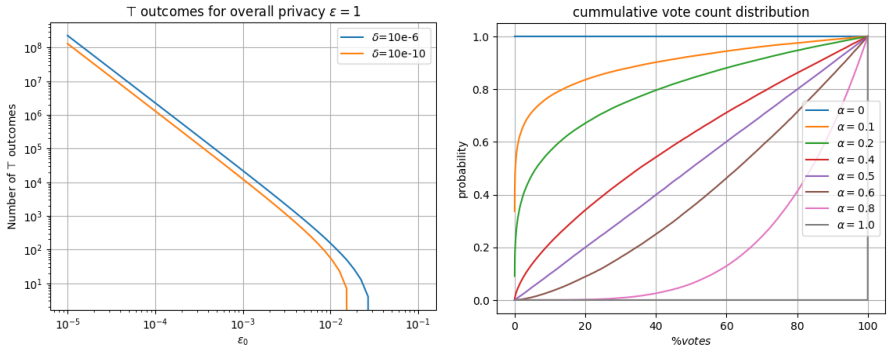


Figure 15: Left: Number of $\top$ responses for ε_0-DP queries for total $\varepsilon = 1$ loss. Right: Cummulative maximum frequency for varying common part α.

Our evaluation here uses (ε, δ) differential privacy (Dwork et al., 2006):

Definition 2 ((ε, δ)-Differential Privacy)**.** A randomized mechanism $\mathcal{M}$ provides (ε, δ)-differential privacy if, for any two datasets D and D' differing in at most one element, and for any subset of outputs $S \subseteq \text{Range}(\mathcal{M})$, the following holds:

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

Concretely we consider `NoisyArgMax` using (Cohen et al., 2021) [5] with the maximum sanitized frequency, with privacy parameters $(\varepsilon_0, \delta_0)$. For privacy analysis across queries we applied the Target Charging Technique (TCT) of Cohen and Lyu (2023) with the *boundary-wrapper* method. The wrapper modifies slightly the output distribution of the query algorithm (after conditioning on ρ!) to include an additional outcome $\top$ (*target*). The wrapper returns $\top$ with this probability (that depends on the response distribution) and otherwise returns a sample from the output distribution of the wrapped algorithm. The probability of $\top$ is at most $1/3$ and decreases with agreement (vanishes when there is response with probability closer to 1). The technique allows us to analyse the privacy loss by only counting target hits, that is, queries with $\top$ response. Since the probability of $\top$ is at most $1/3$, we get in expectation at least two useful responses per target hit. But in case of agreements, we can get many more. Figure 15 (left) reports the number of $\top$ (target) responses we can have with the boundary wrapper method as a function of ε_0 with overall privacy budget is $\varepsilon = 1$. When $\varepsilon_0 \leq 0.01$, it is about $(10\varepsilon_0)^{-2}$.

With Hot PATE, we are interested in *yield* responses, those that return a token (not $\perp$, and when we apply the boundary wrapper, also not $\top$). We study how the yield probability behaves for histograms generated by coordinated ensembles.

[5] We mention the related (non optimized) sparsity-preserving methods (Bun et al., 2019; Korolova et al., 2009; Vadhan, 2017) and optimized but not sparsity-preserving (Ghosh et al., 2012).

Synthetic Teacher distributions: We parametrize the set of teacher distributions by $\alpha \in (0, 1]$, which is the probability of a common part to all distribution. This component is what we aim to transfer to the student. The teacher distributions have probability vectors of the form

$$\boldsymbol{p}^{(i)} = \alpha \cdot \boldsymbol{s} + (1 - \alpha) \cdot \boldsymbol{r}^{(i)} ,$$

where $\boldsymbol{s}$ and $\boldsymbol{r}^{(i)}$ are probability vectors. That is, with probability α there is a sample from the common distribution $\boldsymbol{s}$, and with probability $(1 - \alpha)$, there is a sample from an arbitrary distribution that is specific to each teacher. Note that the common component $\boldsymbol{s}$ can be arbitrarily diverse, that is, $\|\boldsymbol{s}\|_1$ is permitted to be arbitrarily small.

When the histogram is generated by a coordinated ensemble, then the distribution of the maximum frequency c of a token is dominated by sampling $y \sim \mathsf{Exp}[\alpha]$ and then $c \sim \mathsf{Bin}[e^{-y\cdot(1-\alpha)}, n]$. It is visualized in Figure 15 (right) for varying values of α. Note that across all weights $\alpha > 0$ of the shared component, no matter how small α is, there is probability $\approx \alpha$ of being above a high threshold (and returning a token). The probability of $\perp$ (no agreement) in this case can be $\approx 1 - \alpha$. Therefore α parametrizes the probability of yield over the sampling of the shared randomness.

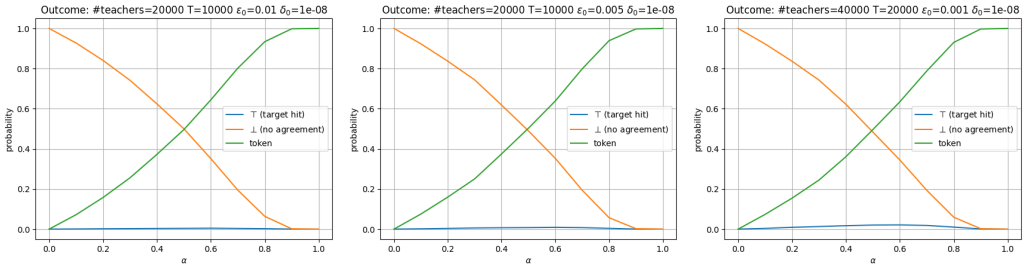


Figure 16: Sweep of α, showing probabilities of outcomes: token, $\perp$, $\top$ (target hit).

Figure 16 shows the distribution of responses as we sweep α, broken down by $\top$ (target hit), $\perp$ (abort), and token (yield). The number of queries we process per target hit, which is the inverse of the probability of $\top$, is $\gtrapprox \varepsilon_0 n$. It is lowest at $\alpha \approx T/n$ and is very high for small and large α, meaning that the privacy cost per query is very small.

The yield (probability of returning a token) per query is $\approx \alpha$. Note that as α decreases, both yield and target probabilities decrease but their ratio remains the same: In the regime $\alpha \leq T/n$, the yield per target hit is $\approx \varepsilon_0 n/2$. Queries with $\alpha \gg T/n$ are essentially free in that the yield (token) probability is very high and the $\top$ (target hit) probability is very low.

When using $n = C_\delta/\varepsilon_0$ ($C_\delta \approx 2\log(1/\delta_0)$ teachers and plugging this in, we obtain that we get $\gtrapprox 0.005\frac{1}{C_\delta}n^2$ yields for overall privacy budget $\varepsilon = 1$. This means that we pay only for yield and not for queries. Note that this holds in the "worst case" across all α values, but the number of yields can be much higher when queries have large α (and "yields" do not incur privacy loss).

G DP METHODS FOR HETEROGENEOUS ENSEMBLES

We propose two DP methods to implement Algorithm 4 (Section E.2) with different trade offs. In both cases we can apply data-dependent privacy analysis so that queries that do not yield a token (that is, return $\perp$) are essentially "free" in terms of the privacy loss. The parameter L depends on the privacy parameters (and logarithmically on $|V|$).

Importantly, with the second method we can apply privacy analysis with individual charging, where instead of charging the whole ensemble as a unit we only charge teachers that contributed to a response. With heterogeneous ensembles we expect the diversity to arise both from individual distributions and from differences between teachers and therefore with individual charging allows for much more efficient privacy analysis when different groups of teachers support each prediction.

Private Weighted Sampling This method gains from sparsity (histogram support being much smaller than $|V|$) but the calculation of privacy loss is for the whole ensemble. We can do the analysis in the TCT framework (Cohen and Lyu, 2023) so that privacy loss only depends on yield queries (those that return a token). We perform weighted sampling by frequency of each token to obtain the sampled histogram $\boldsymbol{c}'$ and then sanitize the frequencies of sampled tokens using the end-to-end sparsity-preserving method of Cohen et al. (2021) to

obtain $\boldsymbol{c}^*$. The sanitizing prunes out some tokens from $\boldsymbol{c}'$ with probability that depends on the frequency c_j, privacy parameters, and sampling rate. All tokens in $\boldsymbol{c}'$ with frequency above $2L$, where L only depends on the privacy parameters, remain in $\boldsymbol{c}^*$.[6] The final step is to return a token from $\boldsymbol{c}^*$ selected uniformly at random or to return $\perp$ if $\boldsymbol{c}^*$ is empty.

Individual Privacy Charging This method does not exploit sparsity, but benefits from individual privacy charging (Kaplan et al., 2021; Cohen and Lyu, 2023). It is appropriate when $2L \ll n$. The queries are formulated as counting queries over the set of teachers. The algorithm maintain a per-teacher count of the number of counting queries it "impacted." A teacher is removed from the ensemble when this limit is reached. Our queries are formed such that at most $O(2L)$ teachers (instead of the whole ensemble) can get "charged" for each query that yields a token.

To express Algorithm 4 via counting queries we do as follows: We sample a sampling rate $\nu \sim U[1/n, 1]$ of teachers and sample a token $v \in V$ uniformly. We sample the teachers so that each one is included with probability ν and count the number c'_v of sampled teachers with $y_i = v$. We then do a `BetweenThresholds` test on c'_j (using (Cohen and Lyu, 2023) which improves over Bun et al. (2017)) to check if $c'_v \geq 2L$. For "above" or "between" outcomes we report v. If it is a "between" outcome we increment the loss counter of all sampled teachers with $y_i = v$ (about $2L$ of them). We note that this process can be implemented efficiently and does not require explicitly performing this "blind" search.

Teachers that reach their charge limit get removed from the ensemble. The uniform sampling of the sampling rate and token emulates weighted sampling, where the probability that a token gets selected is proportional to its frequency. The sub-sampling of teachers ensures that we only charge the sampled teachers. Teachers are charged only when the query is at the "between" regime so (with high probability) at most $\approx 2L$ teachers are charged. Because we don't benefit from sparsity, there is overhead factor of $\log(|V|(n/L))$ in the privacy parameter (to bound the error of this number of queries) but we gain a factor of n/L by not charging the full ensemble for each query in the heterogeneous case where most teachers have different "solutions" to contribute.

[6] We note that the method also produces sanitized (noised) frequency values c^*_j for tokens in $\boldsymbol{c}^*$ such that $|c^*_j - c_j| \leq L$. And hence can also be used for `NoisyArgMax`