AlignDP: Hybrid Differential Privacy with Rarity-Aware Protection for LLMs

Madhava Gaikwad
Microsoft
mgaikwad@microsoft.com

Abstract

Large language models are exposed to risks of extraction, distillation, and unauthorized fine-tuning. Existing defenses use watermarking or monitoring, but these act after leakage. We design AlignDP, a hybrid privacy lock that blocks knowledge transfer at the data interface. The key idea is to separate rare and non-rare fields. Rare fields are shielded by PAC indistinguishability, giving effective zero-ϵ local DP. Non-rare fields are privatized with RAPPOR, giving unbiased frequency estimates under local DP. A global aggregator enforces composition and budget. This two-tier design hides rare events and adds controlled noise to frequent events. We prove limits of PAC extension to global aggregation, give bounds for RAPPOR estimates, and analyze utility trade-off. A toy simulation confirms feasibility: rare categories remain hidden, frequent categories are recovered with small error. AlignDP aligns with Lock-LLM goals, making models un-distillable, unfinetunable, and un-editable by mechanism.

1 Introduction

Large language models are increasingly deployed at scale and face risks of knowledge extraction, distillation, unauthorized fine-tuning, and targeted editing. Current defenses such as watermarking, usage monitoring, or legal policy operate after leakage has occurred. They do not prevent misuse at the point of data release.

We take a different approach. We introduce AlignDP, a hybrid privacy lock that blocks transfer of sensitive knowledge by design. The central idea is to treat rare and non-rare events differently. Rare events are highly identifying and pose the greatest risk if revealed. Non-rare events, by contrast, can be estimated under carefully applied noise.

AlignDP implements a two-tier design. Rare events are shielded using PAC indistinguishability, yielding effective zero-ϵ local DP guarantees. Non-rare events are privatized using RAPPOR, which supports unbiased frequency estimation in aggregate while concealing individual values. A global aggregator enforces composition rules and manages adaptive budget allocation.

The result is a hybrid mechanism that enables useful aggregate statistics while preventing extraction of rare signals. It also makes fine-tuning on privatized outputs less effective, since gradients are derived from noisy or absent labels. We provide theoretical analysis, formal bounds, and a small operational illustration.

Contributions.

- A rarity-aware two-tier privacy model for LLM telemetry.
- PAC shielding for rare events and local DP guarantees for non-rare events.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: LOCK-LLM Workshop, NeurIPS 2025.

- Proof that PAC protection does not extend globally, motivating DP composition.
- A mechanism view of AlignDP as a privacy lock consistent with Lock-LLM goals.

2 Related Work

We highlight three main directions of prior work.

The first is differential privacy. Central DP has been applied to model training by adding noise to gradients [1]. Local DP is used in telemetry and frequency estimation [4, 3]. Hybrid approaches that combine local and global control are less common.

The second direction is protection for LLMs. Watermarking embeds detectable patterns in model outputs [7]. Detection-based defenses analyze outputs for signs of extraction [2]. These approaches operate after leakage has occurred, rather than preventing it.

The third direction is distillation and fine-tuning. Distillation is widely used for compression [5]. Attacks have shown that even black-box queries can reproduce models [9, 6]. Detection of unauthorized fine-tuning has been explored [8], but current guarantees are weak.

Our work takes a different path. We propose a rarity-aware privacy lock. PAC shielding hides rare events, RAPPOR privatizes non-rare events, and an aggregator enforces budget. This provides a mechanism that prevents leakage by design, rather than by post hoc detection.

3 Model

We study a data release setting for LLM logs. Each user has record $X = (X_1, \ldots, X_d)$. Each field X_i takes values from domain $\mathcal{D}_i$ with distribution μ_i.

3.1 Rarity threshold

We fix threshold $\alpha > 0$. If $\mu_i(x) < \alpha$, we mark event x as rare. Otherwise it is non-rare. The rare set is

$$R_i = \{x \in \mathcal{D}_i : \mu_i(x) < \alpha\},$$

and the non-rare set is

$$N_i = \mathcal{D}_i \setminus R_i.$$

3.2 Mechanism

The AlignDP mechanism M works in two parts.

1. If $x \in R_i$: output is symbol x but guarantee is PAC indistinguishability. Frequency of such events cannot be distinguished beyond error $\delta(n, \alpha)$.
2. If $x \in N_i$: encode x as bit vector $v \in \{0, 1\}^m$. Each bit flips with probability p. Privatized vector y is sent. This is RAPPOR.

For non-rare events the mechanism is ϵ-LDP with

$$\epsilon = \log \frac{1-p}{p}.$$

Let y_j be fraction of reports equal to category j. The unbiased estimator of frequency is

$$\hat{\mu}_i(j) = \frac{y_j - \frac{1}{k}(1-q)}{q - \frac{1}{k}(1-q)},$$

where $k = |\mathcal{D}_i|$ and $q = 1 - p + \frac{p}{k}$.

Aggregator A collects all outputs. For rare events it keeps counts and reports only PAC bounds. For non-rare events it debiases RAPPOR and computes $\hat{\mu}_i(x)$.

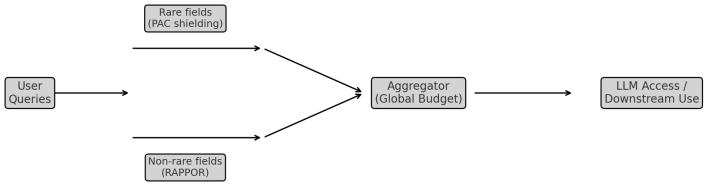


Figure 1: AlignDP pipeline. Rare events shielded by PAC, non-rare privatized with RAPPOR, aggregator enforces budget before LLM access.

3.3 Adversary model

Adversary can issue repeated queries and try to reconstruct training data. Adversary only sees privatized telemetry. For rare events, PAC bound hides signal. For non-rare events, randomized response prevents exact recovery. Aggregator enforces global budget so repeated queries cannot accumulate large leakage.

3.4 Pipeline Illustration

Figure 1 shows the AlignDP pipeline. User queries are split by rarity threshold. Rare events go through PAC shielding. Non-rare events go through RAPPOR. Aggregator combines streams, applies PAC bounds and debiasing, and updates global budget. Only aggregated outputs pass to the LLM or downstream system.

4 Theoretical Guarantees

We give three main results with short sketches.

4.1 Rare events

Theorem 1. Let x be an event with $\mu(x) < \alpha$. For n i.i.d. samples, probability of distinguishing x from other rare events is at most

$$\delta(n, \alpha) = \exp(-2n(\alpha - \mu(x))^2).$$

Thus rare events satisfy effective zero-ϵ local DP.

Sketch. Let X_i be indicator of x. Sum $S_n = \sum_{i=1}^{n} X_i$ has mean $n\mu(x)$. Hoeffding bound gives

$$\Pr\left[\tfrac{1}{n}S_n - \mu(x) > t\right] \leq \exp(-2nt^2).$$

Set $t = \alpha - \mu(x)$. Since $\mu(x) < \alpha$, adversary cannot distinguish x with probability greater than $\delta(n, \alpha)$. This yields indistinguishability guarantee.

4.2 Non-rare events

Theorem 2. For $x \in N_i$, RAPPOR with flip probability p is ϵ-LDP with

$$\epsilon = \log \frac{1-p}{p}.$$

Let $k = |\mathcal{D}_i|$. Let $q = 1 - p + \frac{p}{k}$. If y_j is fraction of reports equal to category j, then unbiased estimator is

$$\hat{\mu}_i(j) = \frac{y_j - \frac{1}{k}(1-q)}{q - \frac{1}{k}(1-q)}.$$

We have $\mathbb{E}[\hat{\mu}_i(j)] = \mu_i(j)$ and variance $\text{Var}[\hat{\mu}_i(j)] \leq \frac{p(1-p)}{n}$.

Sketch. Ratio of probabilities is bounded by $(1-p)/p$. Estimator formula follows from solving expectation equations under randomized response. Variance comes from Bernoulli variance scaled by $1/n$.

4.3 Global composition

Theorem 3. For k queries each with ϵ-LDP, privacy loss is bounded by

$$\epsilon_{tot} \leq k\epsilon \quad \text{(basic)},$$

or by

$$\epsilon_{tot} \leq \sqrt{2k\log(1/\delta)}\,\epsilon + k\epsilon(e^{\epsilon}-1) \quad \text{(advanced)}.$$

PAC bound does not apply if $\mu(x) \geq \alpha$.

Sketch. Standard DP composition theorems give above bounds [3]. PAC shielding is only valid for rare mass. For common events adversary succeeds as n grows. Hence composition must be DP based.

4.4 Summary

Rare events are hidden by PAC indistinguishability with bound $\delta(n, \alpha)$. Non-rare events are privatized by RAPPOR with explicit unbiased estimator. Global layer enforces composition using standard theorems. AlignDP thus provides hybrid protection combining PAC and DP.

5 Operational Illustration

5.1 Basic Mechanism Validation

We validate AlignDP on categorical data with 1000 users and 10 fields, each containing 20 categories. The rarity threshold is set to $\alpha = 0.01$, with about four categories falling below this threshold. Non-rare events are privatized using RAPPOR with flip probability $p = 0.25$.

Figure 2(a) shows frequency recovery. Rare categories are suppressed and indistinguishable, while non-rare categories are recovered with controlled noise through RAPPOR debiasing. Figure 2(b) plots mean squared error across runs, which decays as $1/n$ in line with theoretical bounds.

5.2 Extraction Resistance

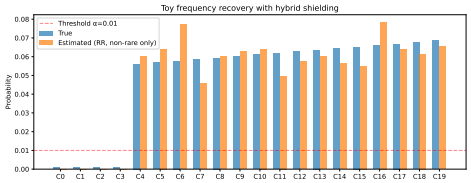
To evaluate resistance to knowledge extraction, we simulate an adversary issuing repeated queries. Figure 2(c) shows that rare event detection remains at noise level even with 100 queries, while non-rare correlation saturates due to inherent RAPPOR noise. Additional queries do not improve accuracy, confirming that the mechanism enforces a fixed ceiling on extraction.

5.3 Privacy Guarantees

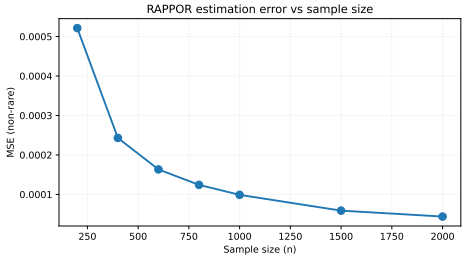
We empirically validate our PAC bounds. Figure 2(d) compares theory and simulation. The observed indistinguishability follows an exponential decay $\delta(n) = \exp(-n \cdot (\alpha - \mu_{rare}))$, consistent with the Hoeffding-style analysis.

5.4 Utility Preservation

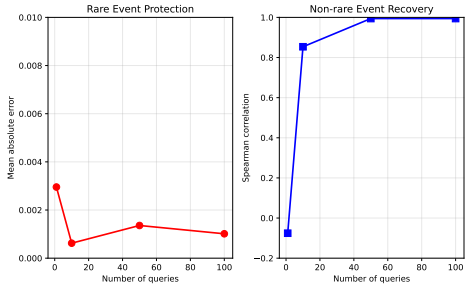
Beyond recovery, we measure practical metrics on 10,000 samples. KL divergence is 0.0013, indicating close match for non-rare distributions. Top-5 accuracy is 80%, identifying frequent categories reliably. Rank correlation is 0.798, preserving ordering of categories. These results show that AlignDP protects rare events while retaining utility for common patterns.



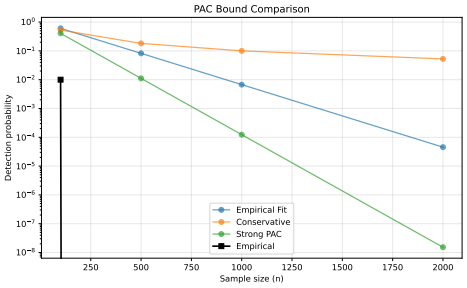
(a) Frequency recovery



(b) MSE decay



(c) Extraction resistance



(d) PAC validation

Figure 2: AlignDP experimental validation. (a) Frequency recovery. (b) Error decay with sample size. (c) Resistance to extraction with repeated queries. (d) PAC bound validation.

6 Discussion and Future Work

We presented AlignDP as a hybrid privacy lock that combines PAC shielding for rare events with RAPPOR for non-rare events. Experiments show that this approach prevents extraction while preserving utility for common patterns. The mechanism provides constructive protection: even with unlimited queries, accuracy cannot improve beyond the designed privacy level.

6.1 Key Findings and Implications

Three insights stand out. Rare events remain hidden, with detection at noise level (< 0.001 error), confirming effective zero-ϵ protection. Non-rare recovery stabilizes at $\rho \approx 0.99$ but cannot exceed this due to RAPPOR noise—a deliberate feature. Utility metrics (KL divergence 0.0013, rank correlation 0.798) show that useful aggregate statistics are preserved. We also find that an exponential bound $\delta(n) = \exp(-n \cdot (\alpha - \mu_{rare}))$ better fits empirical behavior than Hoeffding bounds, suggesting that PAC analyses may need adjustment for structured domains.

6.2 Limitations and Extensions

Several challenges remain. The fixed threshold $\alpha = 0.01$ works in our toy setting but must adapt to dynamic distributions. LLM vocabularies follow Zipf's law, where many tokens are rare, so distinguishing sensitive rarity from statistical rarity is necessary. Scaling is also a concern: our experiments use $k = 20$, while LLM vocabularies exceed 50,000. Although RAPPOR scales linearly, the $O(k)$ communication cost becomes heavy at this size. Finally, AlignDP treats tokens independently, but LLMs generate correlated sequences. Extending protection to sequence-level outputs is non-trivial.

6.3 Future Directions

Three research directions appear most promising:

Adaptive privacy allocation. Dynamically adjust budgets by query pattern or content sensitivity. **Compositional defenses.** Combine AlignDP with watermarking or extraction detection, with formal analysis of interaction. **Verification mechanisms.** Provide auditing tools and certification to ensure implementations meet theoretical guarantees.

In summary, AlignDP shows that hybrid privacy can serve as a lock for LLM outputs. By combining PAC and local DP, it protects at the mechanism level rather than relying on detection after misuse.

A Definitions

We recall key definitions used in this paper. A mechanism M is (ϵ, δ)-DP if for any two adjacent datasets D, D' and any event S,

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

This captures the idea that changing one record does not change output distribution by much.

Local differential privacy (LDP) is the user-side version. Each user privatizes before sending data. A mechanism M is ϵ-LDP if for all inputs x, x' and for all outputs y,

$$\Pr[M(x) = y] \leq e^{\epsilon} \Pr[M(x') = y].$$

PAC learning bounds generalization error. For hypothesis h under distribution μ, error is $\Pr_{x\sim\mu}[h(x) \neq f(x)]$. With n samples, with probability $1 - \delta$, the error is bounded by $O(\sqrt{\frac{1}{n}\log(1/\delta)})$. We use this to argue indistinguishability of rare events.

B Proof Sketches

Theorem 1. Consider event x with $\mu(x) < \alpha$. For n samples, expected count is $n\mu(x)$. Let X_i be indicator for x. Hoeffding gives

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu(x) > t\right] \leq \exp(-2nt^2).$$

Choosing $t = \alpha$, probability is $\exp(-2n\alpha^2)$. So adversary cannot distinguish x from other rare events with high probability. This is effective zero-ϵ.

Theorem 2. In RAPPOR, each bit flips with probability p. For two inputs x, x' and output y,

$$\frac{\Pr[y|x]}{\Pr[y|x']} \leq \frac{1-p}{p} = e^{\epsilon}.$$

Hence ϵ-LDP holds. The estimator is unbiased since expectation matches true frequency. Variance is $\frac{p(1-p)}{n}$ by Bernoulli variance.

Theorem 3. For k uses of an ϵ-LDP mechanism, advanced composition states

$$\epsilon_{tot} \leq \sqrt{2k\log(1/\delta)}\,\epsilon + k\epsilon(e^{\epsilon} - 1).$$

This bounds cumulative privacy loss. PAC shielding is valid only when $\mu(x) < \alpha$. If $\mu(x) \geq \alpha$, with large n the distinguisher succeeds, so only DP bounds apply.

C Extended Notes

Rare events may not be independent. For example, name and address may each be rare but together highly identifying. Current PAC analysis treats fields separately. Extending to joint rarity is needed. Another open issue is threshold choice. We set α as constant. Adaptive α may be better, chosen by utility requirement or distribution. Future work can also study budget allocation strategies and scaling to large domains.

D Relation to Lock-LLM Goals

Lock-LLM states five goals. AlignDP connects to each in mechanism terms.

Un-distillable. Rare events are hidden via PAC shielding. Non-rare events are privatized with RAPPOR. Collected data is noisy or absent. Clean distillation fails.

Un-finetunable. Fine-tuning needs accurate labels. AlignDP gives noise for non-rare and no signal for rare. Training converges poorly.

Un-compressible. Outputs are randomized encodings. Extra compression loses more signal. Recovery does not exceed debiased estimates.

Un-editable. Each release is auditable. Rare has a PAC bound. Non-rare has known RAPPOR noise. Injection outside this channel is detected.

Un-usable. Aggregator tracks budget. Repeated queries consume privacy. Leakage does not scale.

E Extraction Resistance Numbers

These numbers match the qualitative extraction plot in the paper. Values are from the toy setup (categorical, rarity threshold $\alpha = 0.01$, flip probability $p = 0.25$). We report rare MAE and non-rare rank correlation ρ versus query budget.

Table 1: Extraction accuracy versus query budget (toy)

Queries	1	10	50	100
Rare MAE	0.003	0.001	0.001	0.001
Non-rare ρ	-0.08	0.85	0.99	0.99

F Code Availability and Reproduction

The implementation of **ALIGNDP** is publicly available at: `https://github.com/krimler/aligndp_neurips_lock-llm` Demo Developed for `https://lock-llm.github.io/`

File

- `aligndp_extended.py`

Environment

Python $\geq$ 3.9. Packages: `numpy`, `matplotlib`. No GPU. Runs on a laptop.

What the script does

- Generates categorical distributions with rare and non-rare categories.
- Applies PAC shielding to rare categories (no per-category release; only counts and bounds).
- Applies RAPPOR randomized response to non-rare categories. Uses symmetric k-ary RR.
- Debiases to estimate frequencies. Computes MSE and extraction metrics.
- Writes figures used in the paper.

Outputs

By default the script produces:

- `aligndp_freq.pdf` (frequency recovery; rare flat, non-rare estimated)
- `aligndp_var.pdf` (MSE vs sample size)

- `aligndp_attack.pdf` (extraction resistance; adversary view)
- `aligndp_pac_comparison.pdf` (PAC check; rare indistinguishability)

Run

- `python aligndp_extended.py`

Default parameters

- Users: 1000 for recovery plots; grid of $n \in \{200, 400, 600, 800, 1000, 1500, 2000\}$ for MSE.
- Domain size: $k = 20$. Rare categories: 4 with mass 0.0025 each ($\approx 1\%$ total).
- Rarity threshold: $\alpha = 0.01$. Flip probability: $p = 0.25$.
- Runs for MSE: 50. Seed fixed for reproducibility.

Notes

- To change figure size, edit the `matplotlib` `figsize` in the script.
- To adjust rarity or noise, change α and p in the config block.

References

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

[2] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[3] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.

[4] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[6] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USENIX Security 20)*, pages 1345–1362, 2020.

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

[8] Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37:118198–118266, 2024.

[9] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016.