Evidence-bearing Insights under Differential Privacy: Beyond the Limits of Private Text Generation

Anonymous Authors[1]

Abstract

Differentially private (DP) text generation protects against memorization and leakage, but auditing and reporting require outputs to serve as evidence about a private dataset, not merely fluent private text. We study *evidence-bearing insight generation*, where reported statements are backed by privacy-preserving support evidence. We formalize statement-level support and show that a single free-text DP output is locally limited as evidence: attribution and support estimation are statistically constrained by DP indistinguishability, independently of the language model. We introduce proposal-and-filter reporting: a data-independent proposer generates candidate statements, and a DP support test emits only those whose noisy lower confidence bound exceeds a threshold. The mechanism provides DP support certificates, abstentions, per-candidate one-sided honesty bounding unsupported emissions by β, and an EmitRate trade-off. Experiments on realistic reporting tasks show that free-text DP baselines often emit unsupported claims, whereas proposal-and-filter controls UnsupportedEmit and matches the predicted EmitRate interval. These results suggest grounding trust in explicit DP support certificates rather than text generation alone.

1. Introduction

Large language models (LLMs) are increasingly deployed over sensitive data, including retrieval-augmented generation systems (Lewis et al., 2020) over confidential corpora, enterprise assistants grounded in private organizational data, and reporting pipelines that summarize records (Tamkin et al., 2024; Liu et al., 2025). In these settings, memorization is not only a training-time concern: generated reports may expose, amplify, or appear to substantiate patterns derived from private records. Even when systems run under confidential computing (Apple Security Research, 2024; Confidential Computing Consortium, 2023; Google, 2025), intermediate computations and logs may be inaccessible to deployers and auditors. Downstream readers often see only the report, not the computation that produced it. This creates an *observability blackout*: trust must be carried by the released output itself.

Differential privacy (DP) limits the influence of any single record on the output of a randomized mechanism (Dwork et al., 2006; Dwork & Roth, 2014). Recent work extends DP to text generation through private training, synthetic data generation, and inference-time private decoding (Abadi et al., 2016; Xie et al., 2024; Vinod et al., 2025; Yue et al., 2022), aiming to reduce memorization and leakage while preserving useful natural-language outputs. However, suppressing memorization is not the same as producing trustworthy reports. A DP mechanism can spend privacy budget and still output a fluent sentence whose evidential basis is unknown: the claim may reflect a dataset-level trend, a rare memorized or near-memorized anecdote amplified by noise, or merely a model prior. For example, from the statement "*Users frequently discuss medication side effects*," a reader cannot tell whether the claim is strongly supported, weakly supported, or unsupported by the private data. The missing object is not fluency, but evidence.

We therefore study *evidence-bearing insight generation*: reporting mechanisms that output statements together with privacy-preserving evidence that those statements are supported by the private dataset. We formalize reporting at the level of statements. Each statement s has a support value $f_s(R) \in [0, 1]$, measuring how strongly the private dataset R supports it. A *free-text* reporting mechanism outputs only natural language, whereas an *evidence-bearing* mechanism outputs statements together with DP support certificates and abstention decisions. This distinction separates two questions that are often conflated: generating private text that avoids memorization leakage, and certifying that reported claims are supported by the data.

Contributions. We make three contributions:

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

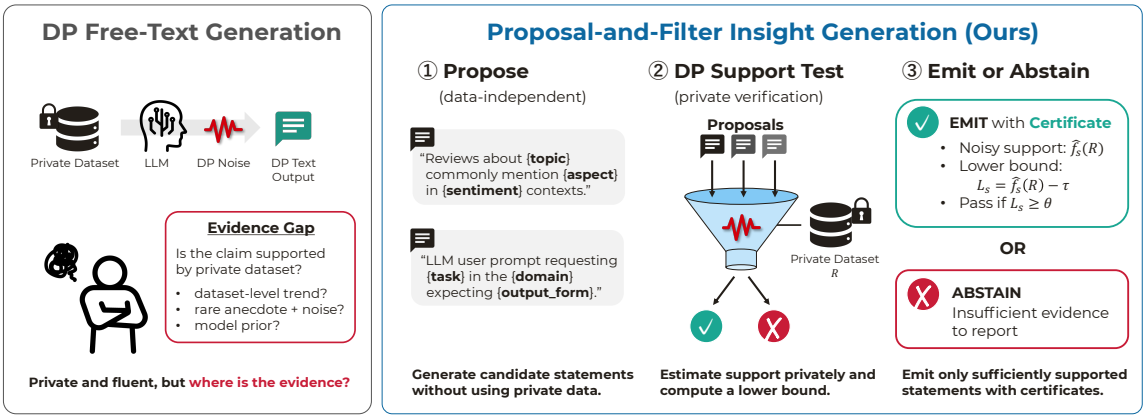

Figure 1. **From private text to evidence-bearing insight.** Proposal-and-filter separates statement proposal from DP support verification, emitting only claims that pass a privacy-preserving evidence test.

1. We formalize evidence gaps in free-text DP reporting through statement-level attribution and support estimation, and show that single free-text DP outputs are locally limited along both axes by DP indistinguishability (Theorems 4.1 and 4.2).
2. We introduce proposal-and-filter reporting, which separates candidate generation from DP support verification, provides per-candidate one-sided honesty (Theorem 5.1), and characterizes the EmitRate trade-off (Theorem 5.2).
3. On TAB and WildChat, we show that free-text DP baselines frequently emit unsupported parsed claims, while proposal-and-filter controls UnsupportedEmit and matches the predicted EmitRate interval.

Together, they recast DP reporting as evidence verification: language proposes claims; DP certificates ground trust.

2. Related Work

DP text generation and private reporting. DP text generation has been studied through private training (Abadi et al., 2016; Yue et al., 2022), inference-time decoding (Vinod et al., 2025), and synthetic data generation (Xie et al., 2024). Aggregate reporting systems such as CLIO (Tamkin et al., 2024) and Urania (Liu et al., 2025) summarize usage patterns, with Urania adding formal DP guarantees. These works generate private text or aggregate insights; we ask when reported natural-language claims are explicitly certified as supported by the private data.

Relation to sparse vector techniques. Proposal-and-filter is related to sparse vector techniques (SVT) (Dwork & Roth, 2014; Lyu et al., 2017), which privately filter queries against a threshold. Our contribution is an evidence-bearing reporting interface: each statement is released with a DP lower-bound support certificate or an explicit abstention.

3. Problem Formulation

We study DP mechanisms that produce *reports* about a private corpus $R = \{r_1, \ldots, r_n\}$ for a reporting task q. A mechanism $\mathcal{A}(q, R)$ is ε-DP with respect to R if, for every fixed q, event E, and adjacent $R \sim R'$,

$$\Pr[\mathcal{A}(q, R) \in E] \leq e^{\varepsilon} \Pr[\mathcal{A}(q, R') \in E].$$

By group privacy, if $d(R, R') \leq k$, then the factor becomes $e^{k\varepsilon}$. This stability underlies the lower bounds in Section 4.

Statements and support. A statement $s \in \mathcal{S}$ is a semantic claim about the dataset, such as "Medication side effects are commonly discussed." Each statement has a support value $f_s(R) \in [0, 1]$, measuring how strongly R supports s. For a fixed per-record predicate $\sigma_s(r) \in \{0, 1\}$, we use the empirical-frequency support

$$f_s(R) = \frac{1}{|R|} \sum_{r \in R} \sigma_s(r), \tag{1}$$

which has sensitivity $\Delta_f \leq 1/|R|$ under replace-one adjacency. More general support models, such as classifiers or semantic matchers, require privacy calibration to their own sensitivity; if the support model depends on the private corpus, its privacy accounting must be handled separately.

Free-text and evidence-bearing reports. A *free-text* mechanism outputs only natural language, $y \sim \mathcal{A}(q, R)$, with no explicit support information. An *evidence-bearing* mechanism outputs tuples $\{(s_j, c_j, a_j)\}_{j=1}^{m}$, where s_j is a candidate statement, c_j is a DP support certificate, and $a_j \in \{\text{emit}, \text{abstain}\}$ is the reporting decision. Thus, statement generation is separated from support certification.

Reporting criteria. We evaluate reports by three criteria.

Support attribution tests whether a statement is supported: given s and $\theta_0 < \theta_1$, decide between $H_0 : f_s(R) \leq \theta_0$ and $H_1 : f_s(R) \geq \theta_1$.

Support estimation estimates $f_s(R)$ from the output; free-text reports require an estimator $\hat{f}_s : \mathcal{Y} \to [0,1]$, while evidence-bearing reports may expose support through c_j.

One-sided honesty bounds unsupported emissions: for an evaluated candidate s, a mechanism is honest at threshold θ with failure probability β if, whenever $f_s(R) < \theta$, $\Pr_Z[\mathrm{Emit}_s] \le \beta$. This per-candidate false-positive guarantee is conditional on s being evaluated and does not guarantee recall.

4. Limits of Free-Text DP Reporting

We ask whether a single natural-language output from a DP mechanism can serve as reliable evidence about the dataset. In the local regime where DP stability is informative, the answer is generally negative: free-text outputs are limited both for attribution and for support estimation. These limits are not language-model-specific; they follow from DP indistinguishability. Our contribution is the reporting formulation, not the testing machinery.

Attribution from a single output. Fix a statement s and thresholds $\theta_0 < \theta_1$. Consider testing $H_0 : f_s(R) \le \theta_0$ versus $H_1 : f_s(R) \ge \theta_1$ from one output $Y \sim \mathcal{A}(q, R)$.

Theorem 4.1 (Testing lower bound under DP)**.** *Let $\mathcal{A}$ be any ε-DP mechanism. Fix q, and let R_0, R_1 satisfy $d(R_0, R_1) \le k$. For $P_i := \mathcal{L}(\mathcal{A}(q, R_i))$,*

$$d_{\mathrm{TV}}(P_0, P_1) \le \frac{e^{k\varepsilon} - 1}{e^{k\varepsilon} + 1}.$$

Consequently, for any randomized test $\varphi : \mathcal{Y} \to [0,1]$,

$$\mathbb{E}_{P_0}[\varphi(Y)] + \mathbb{E}_{P_1}[1 - \varphi(Y)] \ge \frac{2}{e^{k\varepsilon} + 1}.$$

Proof sketch. By group privacy, P_0 and P_1 satisfy the likelihood-ratio bound corresponding to $k\varepsilon$-DP. This implies the stated total-variation bound by the binary testing interpretation of DP (Kairouz et al., 2015). The testing lower bound follows from the standard identity that the optimal sum of type-I and type-II errors is $1 - d_{\mathrm{TV}}(P_0, P_1)$. □

If R_0, R_1 additionally satisfy $f_s(R_0) \le \theta_0$ and $f_s(R_1) \ge \theta_1$, the same lower bound applies to any support-attribution rule based on a single free-text output. Thus, the presence of a statement in a generated report is not, by itself, a certificate that the dataset supports it. The bound is most informative when $k\varepsilon \ll 1$; for empirical-frequency support, a support gap Δ typically requires changing $O(n\Delta)$ records, so the result is a local indistinguishability statement rather than a claim that large support differences are always hidden (Wasserman & Zhou, 2010; Duchi et al., 2013).

Support estimation. We next ask whether the support value itself can be recovered from a single free-text output.

Theorem 4.2 (Support estimation lower bound)**.** *Let $\mathcal{A}$ be any ε-DP mechanism. Fix q, s, and datasets R_0, R_1 with $d(R_0, R_1) \le k$. Let $P_i := \mathcal{L}(\mathcal{A}(q, R_i))$, $a_i := f_s(R_i)$, and $\Delta := |a_1 - a_0|$. Then for every estimator $\hat{f}_s : \mathcal{Y} \to [0,1]$,*

$$\max_{i \in \{0,1\}} \mathbb{E}_{Y \sim P_i}\left[|\hat{f}_s(Y) - a_i|\right] \ge \frac{\Delta}{4}\left(1 - d_{\mathrm{TV}}(P_0, P_1)\right) \ge \frac{\Delta}{2(e^{k\varepsilon} + 1)}.$$

Proof sketch. Suppose an estimator achieves small error on both R_0 and R_1. Thresholding $\hat{f}_s(Y)$ at $(a_0 + a_1)/2$ yields a test distinguishing P_0 from P_1. A standard two-point reduction then gives the first inequality, and Theorem 4.1 gives the second. □

Theorem 4.2 formalizes the support gap: a free-text DP report may contain a statement, but the report alone does not reveal how strongly that statement is supported. A rare anecdotal pattern and a widespread dataset-level trend should therefore not be reported with the same evidential status.

Together, these results show that a free-text DP report may be fluent and private, yet still fail to function as a trustworthy report in the sense of Section 3. We therefore turn from inferring support from text to selectively emitting statements with explicit DP support certificates.

5. Method

We instantiate evidence-bearing reporting by separating proposal from verification. A data-independent proposal distribution Π_q generates candidates $s_1, \ldots, s_m$. Each candidate is verified by a DP support test: $\hat{f}_{s_j}(R) = f_{s_j}(R) + Z_j$, $L_{s_j} = \hat{f}_{s_j}(R) - \tau$. The mechanism emits s_j with certificate L_{s_j} iff $L_{s_j} \ge \theta$, and otherwise abstains. The final report is post-processing of emitted statements and certificates; correctness is enforced by DP support verification, not by the proposer.

Proposer. The proposer affects utility but not privacy or per-candidate honesty, since Π_q is independent of R. Its coverage is captured by

$$p_a(R) := \Pr_{s \sim \Pi_q}[f_s(R) \ge a].$$

A useful proposer is data-independent, matcher-aligned, and places mass above the reporting threshold. We use a *tag-template proposer*: a public schema over topic, aspect, and sentiment is rendered into fixed statements, e.g., "Reviews about {`topic`} commonly mention {`aspect`} in {`sentiment`} contexts." The schema is fixed before observing R, and its controlled form makes statements reliably evaluable.

Privacy. If each support estimate is ε_j-DP, proposal-and-filter is $\left(\sum_{j=1}^{m} \varepsilon_j\right)$-DP; with per-query budget ε_0, it is $m\varepsilon_0$-DP. This follows from data-independent proposal, DP support estimates, sequential composition, and post-processing.

Theorem 5.1 (Per-candidate one-sided honesty)**.** *Fix a candidate statement s. Suppose $\hat{f}_s(R) = f_s(R) + Z$ and $\Pr[Z \geq \tau] \leq \beta$. If the mechanism emits s only when $\hat{f}_s(R) - \tau \geq \theta$, then for every R with $f_s(R) < \theta$, $\Pr_Z[\text{Emit}_s] \leq \beta$.*

Proof sketch. If s is emitted while $f_s(R) < \theta$, then $Z \geq \tau + (\theta - f_s(R)) > \tau$. Thus unsupported emission has probability at most β. A union bound gives family-wise failure at most $m\beta$. □

Theorem 5.2 (EmitRate trade-off)**.** *Let $s_1, \ldots, s_m \overset{\text{i.i.d.}}{\sim} \Pi_q$, and suppose $\hat{f}_{s_j}(R) = f_{s_j}(R) + Z_j$ with $\Pr[Z_j \geq \tau] \leq \beta$ and $\Pr[Z_j \leq -\tau] \leq \beta$. For*

$$\mathcal{O} := \{j : \hat{f}_{s_j}(R) - \tau \geq \theta\}, \qquad \text{EmitRate} := \mathbb{E}\left[\frac{|\mathcal{O}|}{m}\right],$$

we have

$$(1-\beta)p_{\theta+2\tau}(R) \leq \text{EmitRate} \leq p_\theta(R) + \beta(1 - p_\theta(R)) \leq p_\theta(R) + \beta.$$

Proof sketch. For the upper bound, candidates with $f_s(R) \geq \theta$ are emitted with probability at most 1, while those with $f_s(R) < \theta$ are emitted with probability at most β by Theorem 5.1. For the lower bound, any candidate with $f_s(R) \geq \theta + 2\tau$ is emitted whenever $Z \geq -\tau$, which occurs with probability at least $1 - \beta$. Averaging over $s \sim \Pi_q$ gives the bounds. □

Thus, the filter controls unsupported emission, while coverage is determined by proposer mass above $\theta + 2\tau$ and the privacy-induced noise margin.

6. Empirical Evaluation

We evaluate whether, on realistic reporting tasks, proposal-and-filter controls unsupported emission and its EmitRate follows Theorem 5.2, unlike free-text DP baselines.

Metrics. For proposal-and-filter, UnsupportedEmit is the empirical aggregate of the per-candidate failure event controlled by Theorem 5.1, and EmitRate is the emitted fraction controlled by Theorem 5.2. We also report

$$\text{FalseEmission} := \Pr[f_s(R) < \theta \mid s \text{ emitted}],$$

which is not theorem-backed but is useful descriptively. MeanSupport is the average classifier-defined support of emitted statements or parsed free-text claims. For free-text baselines, which have no explicit candidate set or abstention decision, we parse generated text into asserted claims and report FalseEmission and MeanSupport.

Table 1. **Realistic reporting results.** Free-text DP baselines often emit low-support claims, whereas proposal-and-filter emits higher-support statements with lower FalseEmission.

Dataset	Method	FalseEmit ↓	MeanSupport ↑
TAB	Unfiltered LLM	1.000	0.000
	free-text DP (doc)	0.894	0.022
	INVISIBLEINK	0.958	0.008
	Proposal-and-Filter	**0.010**	**0.545**
WildChat	Unfiltered LLM	0.947	0.014
	free-text DP (doc)	0.729	0.051
	INVISIBLEINK	0.647	0.058
	Proposal-and-Filter	**0.014**	**0.288**

Setup. We evaluate TAB (Pilán et al., 2022) and WildChat (Zhao et al., 2024). Support uses Eq. (1) with a frozen public NLI cross-encoder, `cross-encoder/nli-deberta-v3-base` (He et al., 2021), fixed before accessing R, so $\Delta_f \leq 1/|R|$. We compare against no-data LLM generation, document-level free-text DP, and token-level DP decoding via INVISIBLEINK (Vinod et al., 2025). All private methods use $(\varepsilon_{\text{tot}}, \delta) = (10, 10^{-5})$; proposal-and-filter allocates the budget across support tests by basic composition.

Table 1 shows that free-text DP baselines are poorly aligned with classifier-defined support. For INVISIBLEINK, FalseEmission is 0.958 on TAB and 0.647 on WildChat, whereas proposal-and-filter achieves much lower FalseEmission and higher MeanSupport.

Proposal-and-filter also controls omitted theorem-backed quantities: UnsupportedEmit is 0.008 on TAB and 0.004 on WildChat, both below $\beta = 0.05$, and EmitRate falls within the predicted intervals: TAB $0.404 \in [0.356, 0.557]$, WildChat $0.220 \in [0.162, 0.350]$. Thus, trustworthy private reporting requires DP support verification, not merely private text generation.

7. Conclusion

Free-text DP generation can produce fluent private text, but it does not certify dataset support. We formalized this evidence gap through local limits on attribution and support estimation, and introduced proposal-and-filter reporting, which separates data-independent statement proposal from DP support verification. The mechanism provides per-candidate honesty with a predictable EmitRate trade-off and empirically controls unsupported emissions where free-text DP baselines do not. Trust in DP reporting should therefore come from explicit support certificates, not text generation alone.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318. ACM, 2016. doi: 10.1145/2976749.2978318.

Apple Security Research. Private cloud compute security guide. https://security.apple.com/documentation/private-cloud-compute/, 2024. Accessed 2026.

Confidential Computing Consortium. A technical analysis of confidential computing. Technical report, Confidential Computing Consortium, 2023. https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 429–438. IEEE Computer Society, 2013. doi: 10.1109/FOCS.2013.53.

Dwork, C. and Roth, A. *The Algorithmic Foundations of Differential Privacy*, volume 9 of *Foundations and Trends in Theoretical Computer Science*. Now Publishers, 2014. doi: 10.1561/0400000042.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference (TCC)*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14.

Google. Private ai compute in the cloud. https://services.google.com/fh/files/misc/private_ai_compute_technical_brief.pdf, 2025. Accessed 2026.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Kairouz, P., Oh, S., and Viswanath, P. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1376–1385. PMLR, 2015.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Liu, D., Cohen, E., Ghazi, B., Kairouz, P., Kamath, P., Knop, A., Kumar, R., Manurangsi, P., Sealfon, A., Yu, D., and Zhang, C. Urania: Differentially private insights into AI use. In *Conference on Language Modeling (COLM)*, 2025.

Lyu, M., Su, D., and Li, N. Understanding the sparse vector technique for differential privacy. *Proceedings of the VLDB Endowment*, 10(6):637–648, 2017.

Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., and Batet, M. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4): 1053–1101, 2022. doi: 10.1162/coli_a_00458.

Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S., Stern, M., Clarke, B., Goldberg, L., Sumers, T. R., Mueller, J., McEachen, W., Mitchell, W., Carter, S., Clark, J., Kaplan, J., and Ganguli, D. Clio: Privacy-preserving insights into real-world AI use. Technical report, Anthropic, 2024. https://arxiv.org/abs/2412.13678.

Vinod, V., Pillutla, K., and Thakurta, A. G. InvisibleInk: High-utility and low-cost text generation with differential privacy. In *Advances in Neural Information Processing Systems*, 2025.

Wasserman, L. and Zhou, S. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. doi: 10.1198/jasa.2009.tm08651.

Xie, C., Lin, Z., Backurs, A., Gopi, S., Yu, D., Inan, H. A., Nori, H., Jiang, H., Zhang, H., Lee, Y. T., Li, B., and Yekhanin, S. Differentially private synthetic data via foundation model APIs 2: Text. In *International Conference on Machine Learning (ICML)*, 2024.

Yue, X., Inan, H. A., Li, X., Kumar, G., McAnallen, J., Shajari, H., Sun, H., Levitan, D., and Sim, R. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348*, 2022.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024. https://openreview.net/forum?id=Bl8u7ZRlbM.