ε Is Not All You Need: Empirical Privacy Calibration for Text Rewriting under Local Differential Privacy

Anonymous ACL submission

Abstract

As large language models become widely used infrastructure for processing sensitive user inputs, privacy protection and evaluation have become increasingly important. A growing line of NLP research addresses this challenge through text rewriting under Local Differential Privacy (LDP), which obfuscates inputs before they are shared with remote services. These mechanisms provide formal guarantees through a privacy budget ε that bounds worst-case privacy loss. However, nominal ε values are often difficult to interpret and compare across text rewriting mechanisms, which operate at different granularities and may preserve input-specific information in surface form and semantics. We argue that evaluating LDP text rewriting should complement formal guarantees with empirical calibration tailored to text. We propose Text Distinguishability Audit (TeDA), an empirical calibration framework that translates the indistinguishability interpretation of differential privacy into source attribution tests on privatized text. Across six representative LDP mechanisms, we show that similar nominal ε bounds can yield substantially different attribution success rates, revealing differences in empirical distinguishability. Our work positions empirical calibration as an evaluation layer for LDP text rewriting that complements formal guarantees by enabling mechanism comparison, privacy-utility interpretation, and diagnosis of mechanism-specific behavior.

1 Introduction

As large language models (LLMs) become widely deployed, user text has become a primary interface with language technology. It often carries personal, sensitive, or proprietary information, making privacy protection and evaluation practical concerns. A growing line of work addresses this risk through text rewriting under Local Differential Privacy (LDP) (Duchi et al., 2013), where text is obfuscated before being shared with a remote service (Igamberdiev and Habernal, 2023; Meisenbacher et al., 2024). As these mechanisms move toward deployment, a central question is how their privacy behavior should be assessed in practice and compared across rewriting methods.

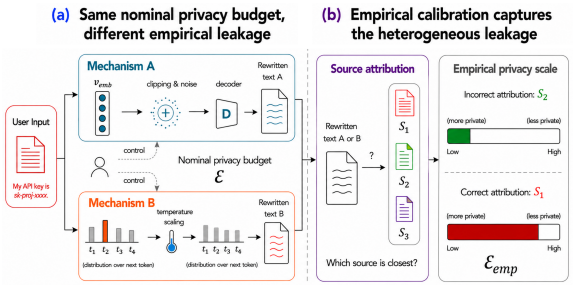

Figure 1: Overview of TeDA. Different LDP text rewriting mechanisms with the same nominal ε can yield different empirical leakage; TeDA calibrates this gap by asking whether privatized texts remain attributable to their sources among candidate inputs.

These mechanisms inherit the appeal of Differential Privacy (DP), which formalizes privacy through indistinguishability (Dwork et al., 2006). Rather than requiring complete disclosure prevention, which is unattainable when released data retain utility and adversaries possess auxiliary information (Dwork, 2006), DP asks neighboring inputs to induce similar output distributions so an observer cannot distinguish which input produced an output. This guarantee is parameterized by a privacy budget ε, which bounds worst-case distinguishability: smaller ε implies stronger privacy guarantees. LDP text rewriting adopts this convention by reporting nominal ε values as its primary privacy metric.

However, this convention is difficult to use as a practical decision metric. A user or system designer choosing an LDP rewriting method must choose both a mechanism and a privacy budget for an acceptable privacy-utility tradeoff, a choice known to be difficult in DP deployments (Dwork et al., 2019). Nominal ε is a worst-case bound, but it does not show how much source information remains ob-

	Method	$\varepsilon_{\text{sent}}$	Rewriting
Ref: I've only used it once and it helped! PLEASE DO NOT SHARE MY NAME OR E-MAIL ADDRESS!			
		Reported $\varepsilon = 1000$	
⇓	ADePT	1000	sues run barcelona because now provided 80 if drive believe flood smart situation sold sign
	DP-BART	1000	Click here to view the latest edition of SHARE THIS STORY Helen
	DP-Para.	26000	I Democratic People helm the White House. foggy skies violence. I dysphagia. I classroom.
	DP-ST	1000	I have used it.
	DP-MLM	20000	I ve only considered it once and it acted! Please Off Stop Whatsapp Pc * With Help Port!
	DP-Prompt	29000	I've only used it once and it helped! Please do not share my name or E-mail address!
		Reported $\varepsilon = 10$	
⇓	ADePT	10	more growing app
	DP-BART	10	(empty output)
	DP-Prompt	290	Imagesated was 70 contacts too People distinguish maximize° Kunstow shifting valve Elizabeth unlikelyfekt Vienna
	DP-Para.	260	it didinder does.aroo-mute used one person ',it works in an empty store like room a few years while getting on
	DP-MLM	200	Fascinated enges only surprised it once and it ant! Vacc Ash Hum Appear Executive Morgan 50 Messenger Classification!
	DP-ST	10	I am being used by Asos.

Table 1: Example text rewritings from Trustpilot (Hovy et al., 2015) under the same reported ε by ADePT (Krishna et al., 2021), DP-BART (Igamberdiev and Habernal, 2023), DP-Paraphrase (DP-Para.) (Mattern et al., 2022), DP-Prompt (Utpala et al., 2023), DP-MLM (Meisenbacher et al., 2024) and DP-ST (Meisenbacher et al., 2025). $\varepsilon_{\text{sent}}$ is the aggregated sentence-level budget: sentence-level mechanisms are unchanged, and token-level mechanisms use basic composition (Dwork and Roth, 2014). Methods are ordered by ascending empirical privacy loss ε_{emp} on Trustpilot (Figure 2; higher indicates greater distinguishability), as indicated by the red downward arrows (⇓).

servable, or whether the same value gives comparable protection across mechanisms. This is challenging as LDP rewriting mechanisms define privacy at different granularities, such as tokens (Meisenbacher et al., 2024) or sentences (Igamberdiev and Habernal, 2023), and even same-granularity guarantees can differ in tightness. As shown in Table 1, identical nominal ε can produce outputs with markedly different relationships to the source text. This motivates empirical calibration (Figure 1): measuring source distinguishability under concrete rewriting mechanisms to support mechanism comparison and privacy-utility interpretation.

This calibration perspective has precedent in empirical privacy measurement, which makes formal privacy claims more interpretable and comparable. In centralized DP, Jayaraman and Evans (2019) used membership inference attacks (Shokri et al., 2017; Yeom et al., 2018) to show that identical nominal ε values can correspond to different empirical leakage. Subsequent auditing work estimates empirical lower bounds on privacy loss with stronger adversarial tests (Jagielski et al., 2020; Nasr et al., 2021, 2023; Maddock et al., 2023; Steinke et al., 2023). For LDP protocols, Arcolezi and Gambs (2024) proposed distinguishability-based auditing for frequency estimation, where an adversary infers the source input from an obfuscated output. Because DP itself constrains distinguishability (Dwork and Roth, 2014), this connects empirical attacks to the formal guarantee.

For LDP text rewriting, however, empirical calibration is not a plug-and-play transfer from tabular auditing. It remains largely unexplored, leaving open how rewriting mechanisms should be assessed and compared in practice. This gap matters for NLP because rewriting operates on discrete language, where source information may persist through wording, meaning, style, or decoding behavior, and existing mechanisms rely on diverse modeling choices such as token sampling, embedding perturbation, and generation. Calibration must also remain computationally practical, since modern rewriting mechanisms often invoke large pretrained models. These features make calibration an emerging interdisciplinary research problem of privacy and NLP evaluation.

To address this gap, we propose Text Distinguishability Audit (TeDA), an empirical calibration framework for LDP text rewriting. TeDA complements nominal privacy budgets by estimating source distinguishability through text-specific attribution tests, candidate construction, and efficient estimation. Our key contributions are:

- We formulate empirical calibration as a complementary evaluation methodology for LDP text

rewriting, translating distinguishability-based auditing into surface- and embedding-level attribution tests for natural language (Section 3.1).

- We instantiate this framework with candidate selection strategies modeling adversaries with varying auxiliary information (Section 3.2) and an aggregate estimator for practical calibration of resource-intensive rewriting mechanisms (Section 3.3).
- We calibrate six representative LDP text rewriting mechanisms, showing that similar nominal ε bounds can imply different empirical privacy behavior and that calibration informs privacy-utility analysis, post-processing effects, and mechanism-specific auditing (Section 4).

2 Related Work

Differential Privacy and Privacy Reporting. DP (Dwork et al., 2006) bounds privacy loss in data release and analysis. Its guarantees are commonly reported through a privacy budget ε, enabling mechanism design and comparison under specified neighboring relations. In privacy-preserving machine learning, DP-SGD (Abadi et al., 2016) and refined accounting methods (Mironov, 2017; Balle and Wang, 2018) have made such reporting central to training on sensitive data. Empirical privacy measurement has shown, however, that nominal ε values alone may not fully characterize practical leakage or mechanism behavior (Jagielski et al., 2020; Nasr et al., 2023; Steinke et al., 2023). We study this issue for LDP text rewriting, where ε values are widely reported but difficult to compare.

Privacy-Preserving Text Rewriting. LDP (Duchi et al., 2013) is a DP variant adapted to decentralized settings, where users perturb their own data before transmission. This property has motivated inference-time privatization for sharing text with remote services (Feyisetan et al., 2019; Hu et al., 2024; Vu et al., 2024; Thareja et al., 2026). Existing mechanisms differ in the linguistic unit used for privacy accounting: some report token-level guarantees (Mattern et al., 2022; Utpala et al., 2023; Meisenbacher et al., 2024), while others report sentence-level guarantees (Krishna et al., 2021; Igamberdiev and Habernal, 2023). These works provide formal guarantees, but reported budgets are not always comparable because granularity, accounting tightness, and observable leakage differ.

Empirical Calibration and Auditing. Empirical calibration makes formal DP claims more interpretable across mechanisms. In central DP, Jayaraman and Evans (2019) used membership inference attacks to show that mechanisms with the same nominal ε can exhibit different empirical leakage. Subsequent calibration work has considered reconstruction attacks and divergence-based formulations for relating empirical attacks to privacy guarantees (Faustini et al., 2025; Kaissis et al., 2024; Kulynych et al., 2025). This line of work has largely focused on central DP.

Empirical auditing is closely related, but typically estimates empirical lower bounds on privacy loss. In central DP, this line includes stronger adversarial tests for auditing privacy guarantees (Jagielski et al., 2020; Nasr et al., 2023). For LDP, auditing has focused mainly on tabular frequency estimation protocols (Gursoy et al., 2022; Arcolezi et al., 2023; Arcolezi and Gambs, 2024). In particular, Arcolezi and Gambs (2024) formalize distinguishability attacks that infer the source input from a privatized output, aligning empirical auditing with DP's indistinguishability interpretation. While this LDP auditing work suggests a route to calibration, tabular frameworks do not transfer directly to text rewriting mechanisms. We adapt distinguishability-based auditing through a text-specific calibration framework for LDP text rewriting.

3 Empirical Privacy Calibration

We present **T**ext **D**istinguishability **A**udit (TeDA), an empirical calibration framework for LDP text rewriting. We formulate calibration as source distinguishability grounded in LDP's indistinguishability view and design a text-tailored pipeline to estimate empirical privacy loss. The main methodological output is a complete calibration procedure for LDP text rewriting, summarized in Algorithm 1.

3.1 Calibration as Source Distinguishability

LDP aims to ensure that a privatized output does not allow an adversary to distinguish the true input from *any other* neighboring input (Igamberdiev and Habernal, 2023; Meisenbacher et al., 2025).[1] Formally, an (ε, δ)-LDP mechanism $\mathcal{M}$ satisfies, for any input texts x, x' and output z,

$$\Pr[\mathcal{M}(x) = z] \leq e^{\varepsilon} \Pr[\mathcal{M}(x') = z] + \delta, \quad (1)$$

where $\delta \geq 0$ permits deviations from the pure ε-LDP guarantee. When $\delta = 0$, this reduces to pure

[1] This differs from metric DP (Chatzikokolakis et al., 2013), where distinguishability may depend on distance under a specified metric.

ε-LDP. The reported ε is an upper bound and can be conservative: a mechanism may report a large budget while remaining empirically difficult to distinguish. Bound tightness can also vary with the accounting procedure, such as the composition theorem used (Dwork and Roth, 2014), complicating comparison across mechanisms. Because LDP is defined in terms of input distinguishability, source attribution provides a direct empirical calibration test: can an adversary identify the source of a privatized output?

From tabular to text. Existing LDP auditing has largely focused on tabular frequency estimation, where mechanisms operate over discrete categorical values and the possible candidates can be enumerated (Arcolezi and Gambs, 2024). Text rewriting operates in a different regime. Modern methods either inject noise into continuous embedding spaces before decoding (Krishna et al., 2021; Igamberdiev and Habernal, 2023) or apply controlled token- or sentence-level perturbations (Mattern et al., 2022; Utpala et al., 2023; Meisenbacher et al., 2024, 2025). The text space is large and open-ended to enumerate, and the lexical or semantic proximity among candidate texts change the difficulty of source attribution.

Candidate-based attribution. To make calibration tractable in the text space, we cast source distinguishability as attribution within a plausible candidate set. Let $\mathcal{M}$ produce a privatized output $y = \mathcal{M}(v_i)$ from source text v_i. Given y and a candidate set $\mathcal{S} = \{v_1, \ldots, v_k\}$ with $v_i \in \mathcal{S}$, an adversary $\mathcal{A}$ predicts which candidate generated y. A successful attribution indicates that the source remains empirically distinguishable within $\mathcal{S}$. This avoids enumeration of the full text space while preserving the indistinguishability question central to LDP.[2] The threat model assumes an adversary with auxiliary information about possible inputs (Dwork and Roth, 2014, §2.2). We instantiate this information as a candidate set containing v_i and distractor samples; in our experiments, these samples are drawn from the same domain as the source, although they could come from elsewhere. We discuss candidate construction in Section 3.2.

[2]Candidate-based evaluation has been applied when reconstructing the original input is difficult even for humans, as in reconstructed-image privacy (Sun et al., 2023) and unsupervised summarization (Zhuang et al., 2022, §4.5). It evaluates relative attribution rather than an isolated reconstruction score, which matters when surface similarity may not reflect actual privacy leakage (Sun et al., 2023; Faustini et al., 2025).

Text-specific attribution. Given a privatized text and a set of candidate sources, attribution requires comparing the output with each candidate. For text, this comparison can be made at two levels. **Text-based attribution** treats source identification as a linguistic judgment over wording, phrasing, and syntactic cues. We use an LLM-based scoring function $f(\cdot, \cdot)$ for source plausibility:

$$\mathcal{A}_{\text{text}}(y) = \arg\max_{v \in \mathcal{S}} f(y, v). \tag{2}$$

Embedding-based attribution instead compares candidates in a semantic vector space. Pretrained embedding models map each text to a vector, enabling distance-based attribution. Let $\phi(\cdot)$ be an embedding function and $d(\cdot, \cdot)$ a distance metric:

$$\mathcal{A}_{\text{emb}}(y) = \arg\min_{v \in \mathcal{S}} d(\phi(y), \phi(v)) \,. \tag{3}$$

An attribution test succeeds when it selects the true source v_i over alternative candidates. The two tests provide complementary calibration perspectives: text-based attribution captures surface-level and human-readable cues, while embedding-based attribution probes semantic structure that may remain after surface form changes. We expect them to broadly align in comparing mechanisms.

3.2 Candidate Sets for Text Calibration

The candidate-based view makes candidate construction central to calibration: the distractors determine which neighboring inputs the adversary must distinguish from the source. In LDP, every element of the input domain is considered adjacent to every other, so the mechanism must make any two inputs indistinguishable (Igamberdiev and Habernal, 2023; Meisenbacher et al., 2025). This all-pairs requirement has different consequences across domains. For frequency estimation over tabular data, mechanisms such as randomized response treat alternative categorical values symmetrically, so prior LDP auditing does not require identifying worst-case inputs (Arcolezi and Gambs, 2024). In text, however, candidate inputs differ in lexical overlap, semantic similarity, and embedding distance, which affect attribution difficulty and therefore the estimated empirical privacy loss.

Candidate selection should therefore be treated as part of text privacy calibration, not merely an implementation detail. Treating candidate construction as arbitrary may yield overly optimistic estimates that underestimate the distinguishability a

realistic observer could achieve. This reflects the auditing principle that stronger attacks reveal more about privacy loss (Jagielski et al., 2020).

We consider three candidate-selection regimes: *uniform sampling*, which ignores semantic proximity; *more-similar sampling*, which draws candidates close to the source and imposes a harder attribution task; and *more-diverse sampling*, which draws dissimilar candidates, making attribution easier and increasing distinguishability pressure. We implement these regimes with *Probabilistic Transition Sampling* (details in Appendix A), which uses a temperature parameter λ to interpolate among them: $\lambda < 0$ gives more-diverse sampling, $\lambda > 0$ gives more-similar sampling, and $\lambda \to 0$ recovers uniform sampling (Arcolezi and Gambs, 2024).

We use more-diverse sampling as the primary configuration because it maximizes distinguishability pressure. We verify in Section 4.5 that the main calibration trends are consistent across candidate-selection regimes, suggesting that our conclusions are robust to the specific selection strategy.

3.3 Estimating Empirical Privacy Loss

Prior distinguishability-based auditing estimates privacy loss through hypothesis tests that track true positive (TP) events, where the adversary identifies the true source, and false positive (FP) events, where it falsely selects that source (Jagielski et al., 2020; Arcolezi and Gambs, 2024). Such procedures often require 10^4 to 10^6 Monte Carlo trials, a cost that is manageable for tabular mechanisms but expensive for LDP text rewriting, where each trial may run large pretrained models.

We use a reduced-form estimator based on aggregated attribution success. An FP event under one hypothesis corresponds to a TP event under the swapped hypothesis, so correct source attribution captures the relevant distinguishability information without separate TP/FP estimation. For a fixed mechanism and attribution test, let p denote success probability; larger p indicates greater empirical distinguishability. We estimate a lower confidence bound $\hat{p}_0$ using the Clopper-Pearson interval (Clopper and Pearson, 1934), yielding

$$\varepsilon_{\text{emp}} = \ln\left(\frac{(k-1)(\hat{p}_0 - \delta)}{1 - \hat{p}_0}\right). \tag{4}$$

The derivation of Equation (4) is deferred to Appendix B; the complete calibration pipeline is given in Algorithm 1 (Appendix C). For $k = 2$, this reduces to $\ln((\hat{p}_0 - \delta)/(1 - \hat{p}_0))$, so larger attribution success yields larger empirical privacy loss. This estimator preserves the hypothesis-testing interpretation while reducing redundant model queries, making calibration practical for resource-intensive rewriting mechanisms.

4 Experimental Setup and Results

4.1 Experimental Setup

Datasets. We test on four datasets commonly used in LDP text rewriting research: ATIS (Dahl et al., 1994), SST-2 (Socher et al., 2013), SNIPS (Coucke et al., 2018), and Trustpilot (Hovy et al., 2015) (licenses and artifact use in Appendix E.1). These datasets span different domains and sentence lengths, providing a diverse calibration testbed.

LDP Text Rewriting Mechanisms. We evaluate six representative mechanisms: ADePT (Krishna et al., 2021; Habernal, 2021), DP-BART (Igamberdiev and Habernal, 2023), DP-Paraphrase (Mattern et al., 2022), DP-Prompt (Utpala et al., 2023), DP-MLM (Meisenbacher et al., 2024), and DP-ST (Meisenbacher et al., 2025). Together, they cover sentence-, token-, and semantic-structure-level rewriting. Mechanism details and implementations are provided in Appendix E.3.

Distinguishability Attacks. We instantiate the attribution tests from Section 3.1 as three attacks: an **LLM** attack using Qwen3-8B (Yang et al., 2025), an **external embedding** attack using `all-mpnet-base-v2` (Song et al., 2020) with cosine distance, and an **internal embedding** attack for ADePT and DP-BART, which apply DP at the native embedding level. Attack details and the LLM prompt are provided in Appendix E.4.

Calibration Configuration. We vary two hyperparameters: the number of Monte Carlo trials, $T \in \{10^4, 10^6\}$, for statistical estimation, and the candidate set size, $k \in \{2, 4, 8, 16\}$, for different amounts of adversarial auxiliary information. We find consistent trends across these settings and use $T = 10^4, k = 2$ as the primary configuration. Full calibration settings are provided in Appendix E.5.

Downstream Utility and Privacy Evaluation. To interpret empirical calibration alongside task behavior, we evaluate downstream utility and private-attribute protection following Meisenbacher et al. (2025). We report SNIPS intent classification, Trustpilot sentiment classification, and Trustpilot gender attribute inference, where lower F1 indicates stronger private-attribute protection. Details are provided in Appendix E.6.

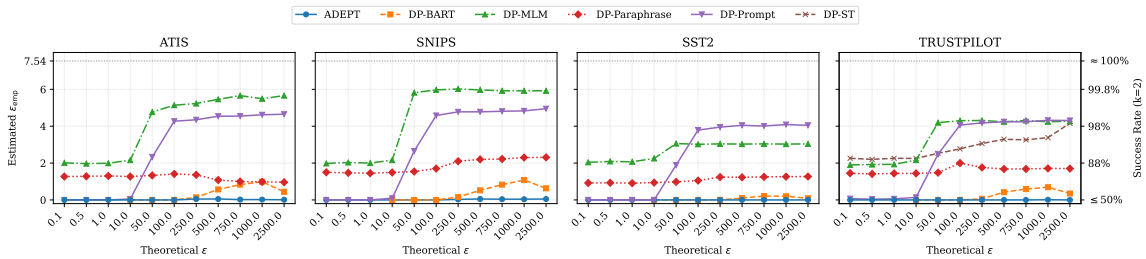

Figure 2: Empirical calibration results under the LLM distinguishability attack across datasets. The x-axis shows nominal ε; the left y-axis shows estimated empirical privacy loss ε_{emp}; and the right y-axis maps ε_{emp} to attribution success rate. The dashed horizontal line marks the finite-sample ceiling ($\approx$7.54, $k = 2$, $T = 10^4$).

4.2 Cross-Mechanism Calibration

Table 1 illustrates a privacy leakage evaluation problem: mechanisms with the same nominal ε can produce privatized texts with different empirical leakage. This gap between reported privacy budgets and empirical leakage motivates our calibration framework as a complementary evaluation layer for LDP text rewriting.

Finding 1

Empirical calibration turns DP's distinguishability perspective into a shared scale for measuring and comparing privacy leakage.

Figure 2 reports empirical calibration results using LLM-based source attribution. For each mechanism, we apply the attribution test to privatized outputs, estimate success rate p over $T = 10^4$ trials, derive the Clopper-Pearson lower bound $\hat{p}_0$, and compute ε_{emp} via Equation (4). Values near zero indicate attribution no better than random guessing. The dashed ceiling is finite because even 100% attribution success gives a Clopper-Pearson lower bound $\hat{p}_0 < 1$, keeping Equation (4) finite under fixed T. On this shared scale, ADePT and DP-BART generally remain near zero, whereas DP-MLM, DP-Paraphrase, DP-Prompt, and DP-ST are more distinguishable and tend to increase with nominal ε. This ranking matches patterns in Table 1.

The empirical separation partly reflects differences in reporting granularity. ADePT and DP-BART report sentence-level budgets, whereas DP-Paraphrase, DP-Prompt, and DP-MLM report token-level budgets. These conventions can be partially aligned by composing token-level guarantees into sentence-level upper bounds: under basic composition, rewriting an n-token sentence gives a bound of roughly $n\varepsilon$ (Dwork and Roth, 2014; Meisenbacher et al., 2024). This conversion remains a worst-case bound and can be loose. In Figure 5 in Appendix G.1, converting token-level guarantees on ATIS using mean sentence length preserves the main ranking, suggesting that nominal rescaling alone is insufficient to explain the separation. We attribute the remaining differences to how mechanisms apply privacy to different linguistic units: sentence-level, token-level, and semantic-triple mechanisms can leave different source cues after privatization. These unit choices also reflect practical tradeoffs; for example, token-level operation can reduce computational cost and generalize more easily across datasets with different sentence-length distributions (Mattern et al., 2022).

Variation appears within a shared granularity. Among token-level mechanisms, the same nominal ε can still lead to different empirical leakage because implementations tokenize inputs and apply token-level privatization differently. As illustrated in Table 5 in Appendix G.2, the same sentence can involve different numbers of privatized tokens or sub-tokens across DP-Paraphrase, DP-MLM, and DP-Prompt. Thus, nominal token-level budgets do not translate into identical sentence-level behavior even within the same broad accounting category. Because both linguistic-unit choices and method-specific implementations shape empirical leakage, empirical calibration provides a shared scale for measuring and comparing privacy leakage.

4.3 Privacy-Utility Interpretability

Nominal ε often guides privacy-utility interpretation and parameter selection. In LDP text rewriting, this can be unreliable when nominal budgets do not align well with empirical leakage.

Finding 2

Empirical calibration provides a more informative basis for privacy-utility analysis than nominal budgets alone.

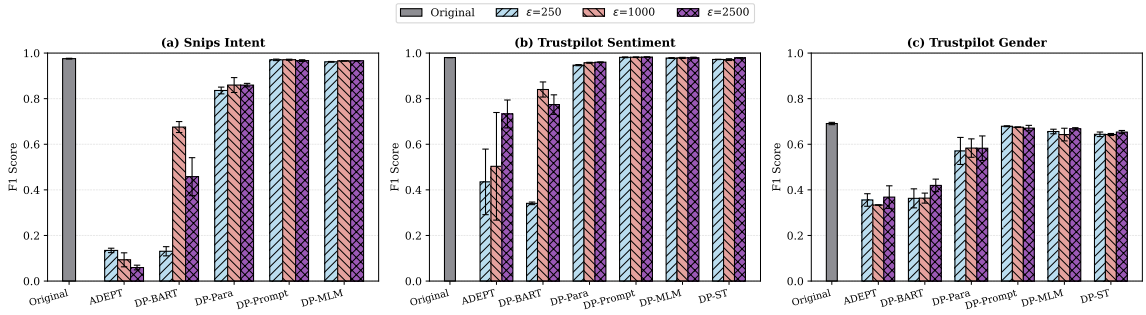

Figure 3: Downstream utility and privacy attribute protection across LDP text rewriting methods, evaluated at $\varepsilon \in \{250, 1000, 2500\}$ with error bars over 3 random seeds. (a) SNIPS intent classification and (b) Trustpilot sentiment report utility F1 (higher is better). (c) Trustpilot gender attribute inference F1 measures resistance to private attribute inference (lower is better).

Figure 3 shows that nominal ε alone gives an incomplete view of the privacy-utility tradeoff. ADePT and DP-BART operate at nominal budgets as large as 1000 or 2500, yet often incur substantial downstream utility loss. Token-level mechanisms preserve stronger intent or sentiment performance at smaller reported budgets. Read only through nominal ε, these results are difficult to place on a single privacy-utility curve.

Empirical calibration clarifies the pattern by aligning the privacy axis with observed distinguishability. ADePT and DP-BART have near-zero ε_{emp}, and this strong empirical indistinguishability is accompanied by degraded downstream utility. Token-level mechanisms show the opposite pattern: they preserve more task-relevant information, but their higher ε_{emp} indicates weaker source privacy, and their Trustpilot gender-inference F1 remains comparatively high. For LDP text rewriting, privacy-utility analysis is therefore more interpretable when the formal worst-case budget is paired with calibrated distinguishability and downstream task behavior.

4.4 Diagnostic Insights from Calibration

Beyond comparison and tradeoff interpretation, empirical calibration is also diagnostic. Varying the attribution view and inspecting curve shapes can show how mechanism choices affect observable source distinguishability.

Finding 3

Empirical calibration diagnoses mechanism behavior, revealing cue-specific attribution, post-processing effects, and model-prior artifacts hidden by nominal budgets.

Complementary Attribution Views. Embedding-space attribution in Figure 6 in Appendix G.1 broadly matches the LLM-based ranking in Figure 2, with one notable reversal: on ATIS, DP-MLM ranks higher than DP-Prompt under the LLM attack but lower under the external embedding attack. Comparing representative outputs helps explain this reversal (Table 2): DP-MLM can leave lexical traces while distorting semantics, which the LLM can use even when embedding similarity weakens. DP-Prompt changes wording more freely but can retain salient semantic cues, giving the embedding attack stronger signal.

Case 1: LLM correct, embedding wrong (DP-MLM)	
Rewritten	moving from india to salt washing
True input	flights from indianapolis to seattle washington
Distractor	what is bna
Case 2: LLM wrong, embedding correct (DP-Prompt)	
Rewritten	the following restrictions are not included in the text
True input	explain restriction ap 57
Distractor	what is the latest flight from boston to denver that serves a meal

Table 2: Complementary behavior of LLM and embedding attacks on ATIS.

Decoding as Post-Processing. For ADePT and DP-BART, DP is applied to sentence embeddings, with decoding applied afterward as post-processing. We therefore run an internal embedding attack on the pre-decoding mechanism output. As shown in the bottom panel of Figure 6 in Appendix G.1, latent-level distinguishability is monotonic in ε, and both mechanisms reach near-ceiling distinguishability before decoding. This differs sharply from LLM and external embedding attacks on decoded text, especially for DP-BART at $\varepsilon = 2500$. While DP

guarantees are immune to post-processing (Dwork and Roth, 2014), the pre/post-decoding gap suggests that decoding may reduce observable distinguishability in practice.

Curve Anomalies. Calibration flags unexpected curve shapes. For DP-BART, ε_{emp} decreases as ε increases from 1000 to 2500. Table 4 in Appendix G.2 shows that at $\varepsilon = 2500$, outputs have high variance and often diverge into generic content, while at $\varepsilon = 1000$ outputs remain more consistent. Figure 7 in Appendix G.2 suggests a possible explanation: DP-BART clips embeddings to a fixed L2 norm before noise injection, so at high ε the perturbed embedding remains close to the clipped representation. One hypothesis is that if clipping shifts representations away from regions that preserve source information well, reducing noise need not improve source fidelity and may instead produce more generic outputs. Such outputs are less tied to the source and harder to attribute, consistent with the utility drop at $\varepsilon = 2500$ in Figure 3.

For DP-Paraphrase, ε_{emp} remains flat across all nominal ε values. Table 5 in Appendix G.2 shows that its outputs are consistently disconnected from the source but share similar stylistic patterns across $\varepsilon \in \{250, 1000, 2500\}$. We attribute this to the GPT-2 paraphraser fine-tuned on SNLI entailment pairs (Meisenbacher et al., 2024): as shown in Figure 8 in Appendix G.2, source-relevant tokens have negligible probability mass while SNLI-style tokens dominate at every temperature setting, making outputs insensitive to privacy-budget changes.

4.5 Robustness Analysis

We further test whether calibration conclusions depend on procedural choices: Monte Carlo trial count, candidate set size, candidate selection strategy, and estimator form.

Monte Carlo Trial Count. Figure 4 shows consistent trends under $T = 10^4$ and $T = 10^6$ trials for DP-MLM on ATIS. We use $T = 10^4$ as the primary setting because it reduces runtime from about 18 hours to under one hour across mechanisms while preserving the comparative trends.

Candidate Set Size. Figure 9 in Appendix G.3 reports two checks: varying k for DP-MLM and comparing mechanisms at $k = 4$. Larger k increases the finite-sample ceiling through the $(k-1)$ factor in Equation (4), but it also makes attribution harder and can lower $\hat{p}_0$. Empirically, the qualitative trends and mechanism ranking remain stable.

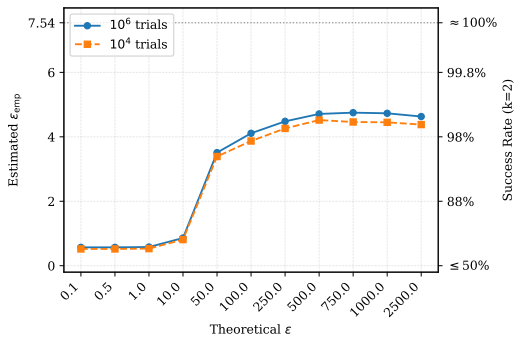

Figure 4: Empirical calibration results under $T = 10^4$ and 10^6 Monte Carlo trials with DP-MLM on ATIS, showing consistent trends across ε values.

Candidate Selection Strategy. Figure 10 in Appendix G.3 compares more-similar ($\lambda = 10{,}000$), uniform ($\lambda = 0$), and more-diverse ($\lambda = -10{,}000$) sampling, reflecting different adversarial auxiliary information. Absolute ε_{emp} values differ because the candidate set changes the attribution task and its difficulty, but the ranking remains consistent.

Efficient Estimation. Figure 11 in Appendix G.3 compares the reduced-form estimator from Equation (4) with the baseline of Arcolezi and Gambs (2024). Under the same more-diverse sampling strategy, the two estimates nearly coincide; estimator choice has much smaller effect than candidate selection, confirming that the efficient estimator does not materially change calibration results.

Together, these checks support our primary configuration ($T = 10^4$, $k = 2$, more-diverse sampling, and the reduced-form estimator) while showing that the main calibration rankings remain stable across reasonable alternatives.

5 Conclusion

We proposed TeDA, an empirical calibration framework for evaluating LDP text rewriting beyond nominal privacy budgets. TeDA translates the indistinguishability interpretation of DP into text-specific source-attribution tests over surface and embedding representations. Across six representative mechanisms, we showed that identical nominal ε values can yield substantially different empirical distinguishability, reflecting differences in privacy granularity, accounting, and rewriting behavior. These findings position empirical calibration as a practical evaluation layer for LDP text rewriting: it complements formal guarantees while supporting interpretable privacy claims through mechanism comparison, privacy-utility analysis, and diagnosing mechanism-specific behavior.

Limitations

This work takes a first step toward empirical calibration for LDP text rewriting by translating DP's distinguishability perspective into text-domain source attribution tests. The proposed framework is intended as a complementary evaluation layer rather than a replacement for formal guarantees, and its current scope reflects two methodological choices.

First, empirical calibration is inherently tied to the attack formulations used for evaluation. Prior privacy studies have explored different attack paradigms, including membership inference and data reconstruction attacks in centralized DP (Jayaraman and Evans, 2019; Faustini et al., 2025). For LDP protocols, distinguishability-based attribution is, to our knowledge, the only empirical auditing formulation developed for privacy-loss estimation, introduced for tabular frequency estimation (Arcolezi and Gambs, 2024). We build on this LDP-specific attribution perspective and extend it to text rewriting, while acknowledging that stronger or more realistic text attacks may provide complementary views of empirical privacy. We recognize this as an important future research direction.

Second, calibration depends on the construction of candidate source sets. As discussed in Section 3.1, an adversary may in principle compare a privatized output against candidates drawn from a broader source pool. Our experiments construct candidates from the same dataset and use more-diverse sampling to model a strong auxiliary-information setting. This provides a controlled and reproducible evaluation protocol under one practically relevant adversarial setting. Future work may examine candidate construction using broader corpora, user-specific priors, or domain-specific auxiliary information.

Ethics Statement

This work focuses on evaluating privacy-preserving text rewriting methods rather than collecting, releasing, or inferring private user data. Our experiments are conducted on existing benchmark datasets under controlled evaluation settings.

The goal of this work is to improve the interpretability and comparability of privacy evaluations for LDP text rewriting mechanisms. We do not anticipate direct negative societal impacts from the proposed empirical calibration framework, whose purpose is to support more transparent evaluation of privacy-preserving NLP methods.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Héber H Arcolezi and Sébastien Gambs. 2024. Revealing the true cost of locally differentially private protocols: An auditing perspective. *Proceedings on Privacy Enhancing Technologies*, 2024(4):123–141.

Héber H. Arcolezi, Sébastien Gambs, Jean-François Couchot, and Catuscia Palamidessi. 2023. On the risks of collecting multidimensional data under local differential privacy. *Proc. VLDB Endow.*, 16(5):1126–1139.

Borja Balle and Yu-Xiang Wang. 2018. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 394–403. PMLR.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the Scope of Differential Privacy Using Metrics. In *Privacy Enhancing Technologies*, pages 82–102, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53.

C. J. Clopper and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément

Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. *arXiv preprint*. ArXiv:1805.10190 [cs.CL].

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438.

Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. 2019. Differential Privacy in Practice: Expose your Epsilons! *Journal of Privacy and Confidentiality*, 9(2).

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Pedro Faustini, Natasha Fernandes, Annabelle Mciver, and Mark Dras. 2025. Empirical calibration and metric differential privacy in language models. *IEEE Access*, 13:196024–196039.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

M. Emre Gursoy, Ling Liu, Ka-Ho Chow, Stacey Truex, and Wenqi Wei. 2022. An adversarial approach to protocol analysis and selection in local differential privacy. *Trans. Info. For. Sec.*, 17:1785–1799.

Ivan Habernal. 2021. When differential privacy meets NLP: The devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 452–461, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. Differentially private natural language models: Recent advances and future directions. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 478–499, St. Julian's, Malta. Association for Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private sgd? In *Advances in Neural Information Processing Systems*, volume 33, pages 22205–22216. Curran Associates, Inc.

Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912, Santa Clara, CA. USENIX Association.

Georgios Kaissis, Stefan Kolek, Borja Balle, Jamie Hayes, and Daniel Rueckert. 2024. Beyond the calibration point: Mechanism comparison in differential privacy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22840–22860. PMLR.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Bogdan Kulynych, Juan Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio Calmon, and Jean Raisaro. 2025. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. In *Advances in Neural Information Processing Systems*, volume 38, pages 111967–112003. Curran Associates, Inc.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs.CL].

Samuel Maddock, Alexandre Sablayrolles, and Pierre Stock. 2023. CANIFE: Crafting canaries for empirical privacy measurement in federated learning. In *The Eleventh International Conference on Learning Representations*.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2025. Leveraging semantic triples for private document generation with local differential privacy guarantees. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8976–8992, Suzhou, China. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9314–9328, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. 2023. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1631–1648, Anaheim, CA. USENIX Association.

Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 866–882.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Phillip Schneider, Manuel Klettner, Elena Simperl, and Florian Matthes. 2024. A comparative analysis of conversational large language models in knowledge-based text generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–367, St. Julian's, Malta. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. 2023. Privacy auditing with one (1) training run. In *Advances in Neural Information Processing Systems*, volume 36, pages 49268–49280. Curran Associates, Inc.

Xiaoxiao Sun, Nidham Gazagnadou, Vivek Sharma, Lingjuan Lyu, Hongdong Li, and Liang Zheng. 2023. Privacy assessment on reconstructed images: Are existing evaluation metrics faithful to human perception? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, and Nils Lukas. 2026. DP-fusion: Token-level differentially private inference for large language models. In *The Fourteenth International Conference on Learning Representations*.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457, Singapore. Association for Computational Linguistics.

Doan Nam Long Vu, Timour Igamberdiev, and Ivan Habernal. 2024. Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 507–527, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs.CL].

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282.

Haojie Zhuang, Wei Emma Zhang, Jian Yang, Congbo Ma, Yutong Qu, and Quan Z. Sheng. 2022. Learning from the source document: Unsupervised abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4194–4205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Probabilistic Transition Sampling

To instantiate the candidate sets discussed in Section 3.2, we use a stochastic sequence generation process: each new candidate is drawn according to its transition probability from the currently selected set. Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ denote the dataset. We iteratively sample a subset $\mathcal{S}_k = \{v_1, \ldots, v_k\}$, initializing v_1 uniformly at random. At each subsequent step t ($2 \leq t \leq k$), the selection probability is defined through the following construction.

Pairwise Transition Probability. We define the probability of transitioning from a single existing sample $s \in \mathcal{S}_{t-1}$ to a new candidate x as a categorical distribution over $\mathcal{X}$, where probability decays exponentially with distance:

$$P(x \mid s) = \frac{\exp\left(-d(x, s)\right)}{\sum_{x' \in \mathcal{X}} \exp\left(-d(x', s)\right)}, \quad (5)$$

where $d(\cdot, \cdot)$ is a pairwise distance metric (*e.g.,* cosine distance).

Joint Transition Score. We aggregate pairwise transition terms into an unnormalized score for candidate x given the current set $\mathcal{S}_{t-1}$:

$$Q(x \mid \mathcal{S}_{t-1}) = \prod_{s \in \mathcal{S}_{t-1}} P(x \mid s). \quad (6)$$

For numerical stability, we work with its log-likelihood as the transition score:

$$\mathcal{L}(x, \mathcal{S}_{t-1}) = \sum_{s \in \mathcal{S}_{t-1}} \ln P(x \mid s). \quad (7)$$

This score captures the global alignment of candidate x with the current set in a numerically stable, additive form.

Sampling Policy. We define the selection probability for the next candidate v_t with a softmax over transition scores, where the sign of λ determines whether high-scoring candidates are favored or penalized:

$$\pi(v_t = x \mid \mathcal{S}_{t-1}) = \frac{\exp\left(\lambda \cdot \mathcal{L}(x, \mathcal{S}_{t-1})\right)}{Z_t}, \quad (8)$$

where Z_t is the partition function over the remaining candidate pool $\mathcal{X} \setminus \mathcal{S}_{t-1}$:

$$Z_t = \sum_{x' \in \mathcal{X} \setminus \mathcal{S}_{t-1}} \exp\left(\lambda \cdot \mathcal{L}(x', \mathcal{S}_{t-1})\right). \quad (9)$$

We set $\lambda < 0$ as our primary configuration, which penalizes candidates close to the current set and favors diverse, dissimilar candidates that increase distinguishability pressure. The formulation also naturally accommodates alternative configurations: a positive λ produces semantically concentrated candidate sets that impose a harder distinguishability task on the adversary (more-similar sampling), while $\lambda \to 0$ recovers uniform random sampling, as adopted in prior work (Arcolezi and Gambs, 2024).

B Proof of the Reduced-Form Privacy Loss Estimator

We provide a formal justification for the reduced-form estimator derived in Section 3.3. The key insight is that a false positive (FP) event under one hypothesis corresponds exactly to a true positive (TP) event under the swapped hypothesis, meaning all relevant distinguishability information can be captured through aggregated success events alone, without separately tracking TP and FP rates. We first establish the notation used throughout this proof.

Data and mechanisms. $\mathcal{X} = \{x_1, \ldots, x_N\}$ denotes the input dataset. $\mathcal{M}$ denotes the LDP text rewriting mechanism and $\mathcal{A}$ denotes the distinguishability attack. $S = \{s_1, \ldots, s_k\}$ denotes a specific candidate set of size k, drawn via the probabilistic transition sampling in Appendix A. Let $\mathcal{C}_k = \{S \subseteq \mathcal{X} : |S| = k\}$ denote the collection of all size-k subsets of $\mathcal{X}$.

Random variables. Cand denotes the candidate set, sampled as explained in Section 3.2. IN denotes the true input, the target drawn uniformly from S in line 11 of Algorithm 1, *i.e.,*

$$\mathbb{P}(\mathrm{IN} = v \mid \mathrm{Cand} = S) = \frac{1}{k} \quad \text{for all } v \in S. \quad (10)$$

$\mathrm{guess} := \mathcal{A}(\mathcal{M}(\mathrm{IN}))$ denotes the output of $\mathcal{A}$ applied to the privatized output $\mathcal{M}(\mathrm{IN})$, taking values in S.

Events. For any $v, v' \in S$ with $v \neq v'$, a true positive (TP) event is $(\mathrm{Cand} = S \wedge \mathrm{IN} = v \wedge \mathrm{guess} = v)$, *i.e.,* the adversary correctly identifies the true source. A false positive (FP) event is $(\mathrm{Cand} = S \wedge \mathrm{IN} = v' \wedge \mathrm{guess} = v)$, *i.e.,* the adversary attributes v when the true source is v'.

Setup. We define the overall success and failure events as unions over all possible candidate sets

$S \in \mathcal{C}_k$ and all possible true inputs $v \in S$:

$$\begin{aligned} A &= \text{success} \\ &= \bigcup_{S \in \mathcal{C}_k} \bigcup_{v \in S} (\text{Cand} = S \wedge \text{IN} = v \wedge \text{guess} = v), \\ \bar{A} &= \text{failure} \\ &= \bigcup_{S \in \mathcal{C}_k} \bigcup_{v \in S} (\text{Cand} = S \wedge \text{IN} = v \wedge \text{guess} \neq v). \end{aligned}$$

By construction, A and $\bar{A}$ are complementary, *i.e.,* $\mathbb{P}(A) + \mathbb{P}(\bar{A}) = 1$.

Step 1: Rewriting the failure event. We rewrite $\bar{A}$ in triple-union form. First, we expand $\text{guess} \neq v$ as a union over all distinct $v' \in S \setminus \{v\}$:

$$\bar{A} = \bigcup_{S \in \mathcal{C}_k} \bigcup_{v \in S} \bigcup_{v' \in S \setminus \{v\}} (\text{Cand} = S \wedge \text{IN} = v \wedge \text{guess} = v'). \quad (11)$$

Swapping the order of the two inner unions:

$$\bar{A} = \bigcup_{S \in \mathcal{C}_k} \bigcup_{v' \in S} \bigcup_{v \in S \setminus \{v'\}} (\text{Cand} = S \wedge \text{IN} = v \wedge \text{guess} = v'). \quad (12)$$

Relabeling $v \leftrightarrow v'$ under the total union yields:

$$\bar{A} = \bigcup_{S \in \mathcal{C}_k} \bigcup_{v \in S} \bigcup_{v' \in S \setminus \{v\}} (\text{Cand} = S \wedge \text{IN} = v' \wedge \text{guess} = v). \quad (13)$$

This shows that $\bar{A}$ is precisely the union of all FP events $(\text{Cand} = S \wedge \text{IN} = v' \wedge \text{guess} = v)$. Since $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$, the probability of all FP events is fully determined by $\mathbb{P}(A)$ alone, without any separate tracking of FP events.

Step 2: Bounding a single success event. The overall success probability decomposes as:

$$\mathbb{P}(A) = \sum_{S \in \mathcal{C}_k} \sum_{v \in S} \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v \wedge \mathcal{A}(\mathcal{M}(v)) = v). \quad (14)$$

We bound a single success event. For any $v \in S$ and any neighboring input $v' \in S \setminus \{v\}$, applying the chain rule of probability:

$$\begin{aligned} &\mathbb{P}(\text{Cand} = S \wedge \text{IN} = v \wedge \mathcal{A}(\mathcal{M}(v)) = v) \\ &= \mathbb{P}(\mathcal{A}(\mathcal{M}(v)) = v \mid \text{Cand} = S \wedge \text{IN} = v) \\ &\quad \cdot \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v). \end{aligned}$$

Applying the (ε, δ)-LDP guarantee of $\mathcal{M}$ and thus $\mathcal{A} \circ \mathcal{M}$ by immunity to post-processing and auxiliary information:

$$\begin{aligned} &\leq \left(e^{\varepsilon} \mathbb{P}(\mathcal{A}(\mathcal{M}(v')) = v \mid \text{Cand} = S \wedge \text{IN} = v) + \delta\right) \\ &\quad \cdot \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v). \end{aligned}$$

Given a candidate set S, for all $v \in S$, the random variables $\mathcal{A}(\mathcal{M}(v))$ and IN are independent, so we have:

$$\begin{aligned} &= \left(e^{\varepsilon} \mathbb{P}(\mathcal{A}(\mathcal{M}(v')) = v \mid \text{Cand} = S \wedge \text{IN} = v') + \delta\right) \\ &\quad \cdot \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v). \end{aligned}$$

Using the uniform prior equation 10, we have:

$$\begin{aligned} &= e^{\varepsilon} \mathbb{P}(\mathcal{A}(\mathcal{M}(v')) = v \mid \text{Cand} = S \wedge \text{IN} = v') \\ &\qquad \cdot \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v') \\ &\quad + \delta \cdot \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v) \\ &= e^{\varepsilon} \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v' \wedge \mathcal{A}(\mathcal{M}(v')) = v) \\ &\quad + \delta \cdot \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v). \end{aligned}$$

Therefore, for any $v \in S$ and $v' \in S \setminus \{v\}$:

$$\begin{aligned} &\mathbb{P}(\text{Cand} = S \wedge \text{IN} = v \wedge \text{guess} = v) \\ &\quad \leq e^{\varepsilon} \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v' \wedge \text{guess} = v) \\ &\qquad + \delta \cdot \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v). \end{aligned} \quad (15)$$

Step 3: Aggregating over all success events. Multiplying both sides of equation 14 by $(k-1)$ and expanding the factor $(k-1)$ as a sum over $v' \in S \setminus \{v\}$:

$$\begin{aligned} &(k-1) \cdot \mathbb{P}(A) \\ &= \sum_{S} \sum_{v \in S} \sum_{v' \in S \setminus \{v\}} \\ &\qquad \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v \wedge \text{guess} = v). \end{aligned}$$

Applying equation 15 and using

$$\begin{aligned} &\mathbb{P}(\text{Cand} = S \wedge \text{IN} = v) \\ &= \mathbb{P}(\text{IN} = v \mid \text{Cand} = S) \cdot \mathbb{P}(\text{Cand} = S). \end{aligned}$$

we obtain:

$$\begin{aligned} &\leq \sum_{S} \sum_{v \in S} \sum_{v' \in S \setminus \{v\}} \\ &\qquad \Big(e^{\varepsilon} \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v' \wedge \text{guess} = v) \\ &\qquad + \delta \cdot \mathbb{P}(\text{IN} = v \mid \text{Cand} = S) \cdot \mathbb{P}(\text{Cand} = S)\Big) \\ &= e^{\varepsilon} \sum_{S} \sum_{v \in S} \sum_{v' \in S \setminus \{v\}} \\ &\qquad \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v' \wedge \text{guess} = v) \\ &\qquad + \delta (k-1) \sum_{S} \sum_{v \in S} \mathbb{P}(\text{IN} = v \mid \text{Cand} = S) \\ &\qquad\qquad \cdot \mathbb{P}(\text{Cand} = S) \\ &= e^{\varepsilon} \mathbb{P}(\bar{A}) + \delta \cdot (k-1), \end{aligned}$$

where the last step uses two facts:

(i) By equation 13, $\bar{A}$ is the union of all FP events, so $\sum_S \sum_{v \in S} \sum_{v' \in S \setminus \{v\}} \mathbb{P}(\text{Cand} = S \wedge \text{IN} = v' \wedge \text{guess} = v) = \mathbb{P}(\bar{A})$.

(ii) $\sum_S \sum_{v \in S} \mathbb{P}(\text{IN} = v \mid \text{Cand} = S) \cdot \mathbb{P}(\text{Cand} = S) = 1$ and $\sum_S \mathbb{P}(\text{Cand} = S) = 1$.

For $\mathbb{P}(A) < 1$, using $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$, rearranging gives:

$$\varepsilon \geq \ln \frac{(k-1)(\mathbb{P}(A) - \delta)}{1 - \mathbb{P}(A)}. \tag{16}$$

Since the expression in Equation 16 is non-decreasing in $\mathbb{P}(A)$, substituting the Clopper-Pearson lower bound $\hat{p}_0$ (Clopper and Pearson, 1934) for $\mathbb{P}(A)$ gives the estimator used in Algorithm 1:

$$\varepsilon_{\text{emp}} = \ln \left(\frac{(k-1)(\hat{p}_0 - \delta)}{1 - \hat{p}_0} \right). \tag{17}$$

When $k = 2$, the estimator reduces to:

$$\varepsilon_{\text{emp}} = \ln \left(\frac{\hat{p}_0 - \delta}{1 - \hat{p}_0} \right). \tag{18}$$

C Empirical Auditing Pipeline

As described in Section 3, we summarize our complete empirical privacy loss estimation pipeline in Algorithm 1.

Algorithm 1 Empirical Privacy Loss Estimation

Input: Dataset $\mathcal{X}$; LDP Mechanism $\mathcal{M}$; Distinguishability Attack $\mathcal{A}$; Subset Size k; Trial Count T; Confidence Level α_{conf}; Privacy Slack δ; Sampling Temperature λ.
Output: Estimated privacy loss ε_{emp}.

```
TP ← 0
for t ← 1 to T do
    S ← {v_1}, with v_1 ~ Unif(X)
    for j ← 2 to k do
        # Compute log-likelihoods for candidates
        L(x) ← Σ_{v∈S} ln P(x | v),   ∀x ∉ S
        Compute policy π(x) ∝ exp(λ · L(x))
        Sample v_j ~ π
        Update S ← S ∪ {v_j}
    end for
    Draw target index i ~ Uniform{1, ..., k}
    y ← M(v_i)
    if A(y, S) = v_i then
        TP ← TP + 1
    end if
end for
p̂_0 ← ClopperPearsonLower(TP, T, α_conf)
return ε_emp ← ln((k−1)(p̂_0−δ) / (1−p̂_0))
```

D AI Assistants

We used ChatGPT and Claude for writing and formatting support, including proofreading, grammar checking, and surface-level language edits.

E Experimental Setup Details

As discussed in Section 4.1, we provide additional details on the datasets, mechanisms, implementation, calibration configuration, and downstream evaluation protocol.

E.1 Datasets, Licenses, and Artifact Use

We use the training splits of ATIS (Dahl et al., 1994), SNIPS (Coucke et al., 2018), SST-2 (Socher et al., 2013), and Trustpilot (Hovy et al., 2015) for academic, non-commercial research after preprocessing and deduplication. We do not redistribute original or processed datasets. SNIPS is used under the CC0 1.0 license; for the HuggingFace-hosted ATIS, SST-2, and Trustpilot versions, we did not find explicit public license information.

We use public pretrained models and released checkpoints under their respective licenses: Apache License 2.0 for Qwen3-8B (Yang et al., 2025), `all-mpnet-base-v2` (Song et al., 2020), FLAN-T5 (Chung et al., 2024), and DP-BART checkpoints (Igamberdiev and Habernal, 2023); MIT License for RoBERTa (Liu et al., 2019) and GPT-2 (Radford et al., 2019); and the Llama 3.2 Community License Agreement for Llama-3.2 (Grattafiori et al., 2024). We use these artifacts only for research evaluation.

E.2 Computational Resources

All experiments were run on NVIDIA V100 32GB GPUs. Calibration with $T = 10^4$ varies by mechanism and attack but typically completes within one GPU hour, while the $T = 10^6$ run for DP-MLM on ATIS required approximately 18 GPU hours.

E.3 LDP Text Rewriting Mechanisms

Mechanism Overview. The six mechanisms span two main paradigms. ADePT and DP-BART privatize text by injecting Laplace or Gaussian noise into continuous embedding representations, while the remaining methods realize privacy via the Exponential Mechanism (EM): DP-Paraphrase, DP-Prompt, and DP-MLM control token-level generation, and DP-ST replaces semantic triples with candidates drawn from an auxiliary corpus.

- **ADePT** (Krishna et al., 2021; Habernal, 2021) is an RNN-based autoencoder that privatizes text by injecting calibrated noise into the encoder's latent representation before decoding.
- **DP-BART** (Igamberdiev and Habernal, 2023) adds noise to the sentence-level latent embedding

of a BART autoencoder (Lewis et al., 2020); we use the DP-BART-PR+ variant.

- **DP-Paraphrase** (Mattern et al., 2022) applies temperature sampling during autoregressive generation with GPT-2 (Radford et al., 2019) to provide token-level guarantees.
- **DP-Prompt** (Utpala et al., 2023) treats temperature sampling during FLAN-T5 (Chung et al., 2024) generation as an exponential mechanism.
- **DP-MLM** (Meisenbacher et al., 2024) performs token-level privatization through the exponential mechanism over candidate tokens generated by RoBERTa (Liu et al., 2019).
- **DP-ST** (Meisenbacher et al., 2025) extracts semantic triples, privatizes them via the exponential mechanism over an auxiliary corpus, and reconstructs the output with LLM paraphrasing.

Implementation Details. We use released implementations whenever available. For ADePT (Krishna et al., 2021) and DP-BART (Igamberdiev and Habernal, 2023), we use the implementation and checkpoints from Igamberdiev and Habernal (2023). For DP-Paraphrase (Mattern et al., 2022), DP-Prompt (Utpala et al., 2023), and DP-MLM (Meisenbacher et al., 2024), we use the unified implementation from Meisenbacher et al. (2024). For DP-ST (Meisenbacher et al., 2025), we use the official code from Meisenbacher et al. (2025). In all cases, we retain the default hyperparameters reported in the respective works unless otherwise noted below.

ADePT. We adopt the implementation from Igamberdiev and Habernal (2023), which fixes the sensitivity calculation issue pointed out by Habernal (2021). We use the released scripts to fine-tune the rewriting RNN model on the OpenWebText dataset (Gokaslan and Cohen, 2019) and apply it for rewriting.

DP-BART. We adopt the official implementation from Igamberdiev and Habernal (2023), selecting the DP-BART-PR+ variant among DP-BART-CLV, DP-BART-PR, and DP-BART-PR+, which is reported to achieve the best privacy-utility trade-off. We use the released checkpoints provided by Igamberdiev and Habernal (2023).[3] The analytical Gaussian mechanism is applied for both ADePT and DP-BART.

DP-Paraphrase. As no official implementation is released, we follow the replication from Meisenbacher et al. (2024), in which the authors fine-tune a GPT-2 model (Radford et al., 2019) on the SNLI dataset (Bowman et al., 2015). We use this open-source GPT-2 model and follow their modification of changing the logit clipping operation from normalization to a fixed range, as discussed in Appendix B of Meisenbacher et al. (2024).

DP-Prompt and DP-MLM. We use the released code from Meisenbacher et al. (2024) with their default configurations. For DP-Prompt, the paraphraser is FLAN-T5-Base as specified in their implementation. For DP-MLM, the masked language model is RoBERTa-Base.

DP-ST. We use the code released by Meisenbacher et al. (2025) and follow the instructions to extract 15 million subject-verb-object (SVO) triples (Schneider et al., 2024) from FineWeb (Penedo et al., 2024) as the auxiliary corpus. Private SVO triples are replaced via the exponential mechanism using triples sampled from the nearest cluster, then paraphrased with Llama-3.2-1B-Instruct (Grattafiori et al., 2024); the 1B model is reported to provide better empirical privacy than the 3B variant (Meisenbacher et al., 2025, §5). We adopt a clustering size of 50K, which achieves better privacy-utility trade-offs. Because DP-ST requires reliable SVO extraction from the source input, we report DP-ST results only on Trustpilot, where such extraction is consistently feasible.

E.4 Distinguishability Attacks

For the LLM attack, we use Qwen3-8B (Yang et al., 2025) as the judge. The model receives a privatized output and k candidate inputs and predicts the most plausible source. For the external embedding attack, we encode all texts with `all-mpnet-base-v2` (Song et al., 2020) and select the nearest candidate under cosine distance. For ADePT and DP-BART, which operate explicitly in embedding space, we also evaluate internal embeddings by directly comparing pre- and post-noise latent representations. This internal attack validates the direct mechanism-level relation between noise and distinguishability before decoding post-processing: higher noise (lower ε) should monotonically reduce distinguishability.

LLM Distinguishability Prompt. The following prompt is used for the LLM distinguishability attack. The rewritten text and candidate original texts are substituted at inference time.

[3] https://huggingface.co/TrustHLT

LLM Distinguishability Prompt

You are given a rewritten text Y and several candidate original texts. Your task is to identify which candidate is the most likely original text that was rewritten to produce Y.

Output format. Answer with exactly the following format on a single line: **answer: [[N]]**, where $N \in \{1, \ldots, K\}$. Do not include analysis or any additional text.

Rewritten text Y:
{y_text}

Candidate original texts:
1. *{candidate_1}*
2. *{candidate_2}*
⋮
K. *{candidate_K}*

Which candidate is most likely the original?

E.5 Calibration Configuration

For each method, we evaluate rewriting across $\varepsilon \in \{0.1, 0.5, 1.0, 10.0, 50.0, 100.0, 250.0, 750.0, 1000.0, 2500.0\}$. For DP-BART, we restrict to $\varepsilon \in \{10.0, 50.0, 100.0, 250.0, 750.0, 1000.0, 2500.0\}$, as Igamberdiev and Habernal (2023) note that at small ε values noise is too high for meaningful decoding, producing predominantly empty outputs.

For our framework, we evaluate both $T = 10^4$ and $T = 10^6$ Monte Carlo trials and find consistent results. We report $T = 10^4$ to keep computation tractable. Each trial uses a distinct random seed to capture rewriting randomness. Our pipeline extends beyond the standard two-candidate setting to any number of candidates k, representing adversaries with varying prior knowledge. We primarily report $k = 2$ for comparability with prior work (Arcolezi and Gambs, 2024), and verify $k = 4, 8, 16$, finding consistent trends.

For sampling temperature, as introduced in Appendix A, $\lambda < 0$ yields more-diverse sampling, $\lambda > 0$ more-similar sampling, and $\lambda \to 0$ uniform sampling. We set $\lambda = \pm 10{,}000$ based on alignment with top-1 and top-10 nearest/furthest neighbor sampling using `sklearn` (Pedregosa et al., 2011) with cosine distance on embeddings. Section 4.5 and Figure 10 show consistent trends across λ values.

E.6 Downstream Evaluation

To assess privacy-utility tradeoffs, we follow Meisenbacher et al. (2025). For each mechanism, we fine-tune a DeBERTa-v3-base (He et al., 2021) classifier for one epoch on the privatized training set and evaluate on the original clean test set. On Trustpilot, we evaluate binary sentiment classification as the utility task and binary gender attribute inference as the privacy task, where lower F1 on the latter indicates stronger resistance to attribute inference. On SNIPS, we evaluate multi-class intent classification as the utility task. All experiments are repeated over 3 random seeds, and we report mean F1 with standard deviation.

F Additional Text Rewriting Examples

As discussed in Section 1, Table 3 provides additional sample system outputs for all six mechanisms on the Trustpilot (Hovy et al., 2015) dataset across three privacy budget levels.

G Additional Results

This section provides supporting evidence for the calibration analysis in Section 4.2, the diagnostic analyses in Section 4.4, and the robustness checks in Section 4.5.

G.1 Additional Calibration Results

Figure 5 compares mechanisms after converting token-level budgets to sentence-level equivalents, supporting the main conclusion that empirical privacy ordering is not merely an artifact of plotting mechanisms under different nominal-budget conventions. Figure 6 extends the calibration analysis to external and internal embedding attacks: the external attack provides a second observable view of source distinguishability, while the internal attack isolates pre-decoding behavior for mechanisms whose DP noise is applied in latent space.

G.2 Mechanism-Specific Diagnostics

Table 4 and Figure 7 diagnose the non-monotonic behavior of DP-BART, while Table 5 and Figure 8 explain token-level variation and the flat response of DP-Paraphrase across nominal privacy budgets. Together, these diagnostics support the claim that calibration can reveal mechanism-specific behavior hidden by nominal budgets.

G.3 Robustness Checks

Figure 9 varies the candidate set size, Figure 10 varies the candidate-selection strategy, and Figure 11 compares the reduced-form estimator with the baseline auditing procedure.

Method	Rewriting
Ref: I had previously used and bought from Boilerjuice, so it had on record somewhere a USER name and EMAIL ADDRESS for me	
$\varepsilon = 1000$	
ADePT	zoo retired results now package counts comeback such because
DP-BART	Quote from Boiler Room "I just can't believe it
DP-Paraphrase	I Stapped on Bo resurgence. I tribeed on Boilerju pioneered. I————-
DP-Prompt	I had previously used and bought from Boilerjuice , so it had on record somewhere a USER name and EMAIL ADDRESS for me
DP-MLM	I had normally purchased and purchased from The, so it had on record both a Us name and Email Catalogue for me
DP-ST	We tried to put leftovers in a bottle, but it was on a record.
$\varepsilon = 250$	
ADePT	debate
DP-BART	Martby Smm Srm NESS Sle
DP-Paraphrase	I SPENT BOTH OF Durant's money EPISODING Boilerju arrivals—— I disappears punished Traitor with a bunch Encyclopedia links Football
DP-Prompt	I had previously used and bought from Boilerjuice , so it had on record somewhere a USER name and EMAIL ADDRESS for me
DP-MLM	I had simply usage and ate from The, so it had on record both a Us name and Email Address for me
DP-ST	We have bottles on record.
$\varepsilon = 10$	
ADePT	minnesota eggs obstacle gearbox : online time any mart app upgrades
DP-BART	(empty output)
DP-Paraphrase	Bovince a USA to disconnect communication forgeting undermined communication resolutionsWhy gospel name ENCARDONSK Officiel onboifer iyvesd too.pl
DP-Prompt	galaxyblurest retail steel supplies biography replenish tested council Smith $200 none VersicherungsStack spreading signage Barack spectrum cassette lou Press loin Strawberryawa Joseph neatnet 1954wbours Except phone solltenNERwith dagegen =prox progress Trouble Quarry108 regional
DP-MLM	艹 had congregation given and pro from Bell, so it had on inventor executed a Crib hobbit and Error Bold for me
DP-ST	Free sodas are typically held in records.

Table 3: Example outputs of six LDP text rewriting mechanisms under the same theoretical ε on Trustpilot reviews (Hovy et al., 2015).

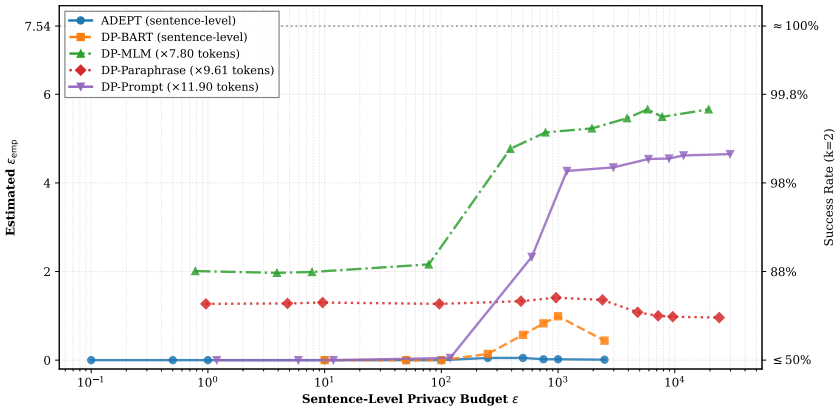

Figure 5: Cross-convention comparison on the ATIS dataset for the LLM distinguishability attack (x-axis in log scale); token-level methods (DP-MLM, DP-Paraphrase, DP-Prompt) are converted to sentence-level equivalents via the basic composition theorem (Dwork and Roth, 2014) using mean sentence token length.

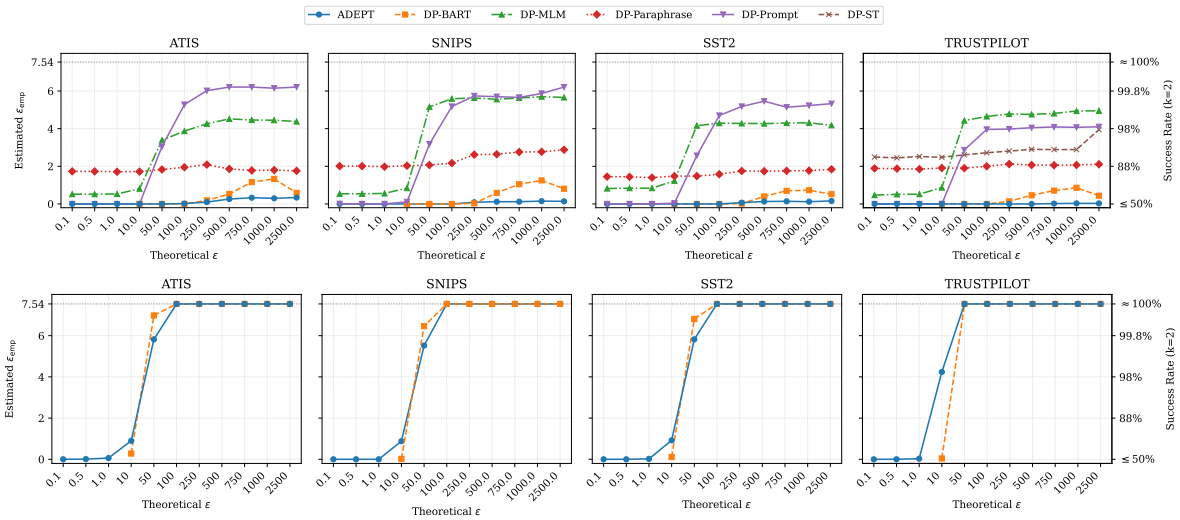

Figure 6: Empirical calibration results under the **external** (*top*) and **internal** (*bottom*) embedding distinguishability attacks across datasets. Axes and ceiling as in Figure 2.

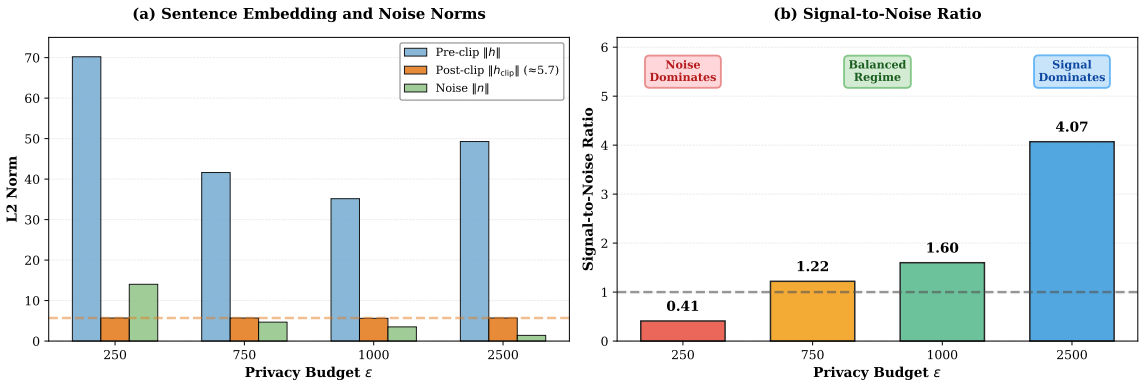

Figure 7: Signal-to-noise ratio (SNR) analysis of DP-BART across different ε values. (a) The clipping operation forces all post-clip sentence embeddings to have approximately constant L2 norm (≈5.7), while noise magnitude decreases as ε increases. (b) The resulting SNR varies significantly: high SNR at $\varepsilon = 2500$ (4.07), balanced SNR at $\varepsilon = 1000$ (1.60), and low SNR at $\varepsilon = 250$ (0.41), corresponding to different output quality regimes.

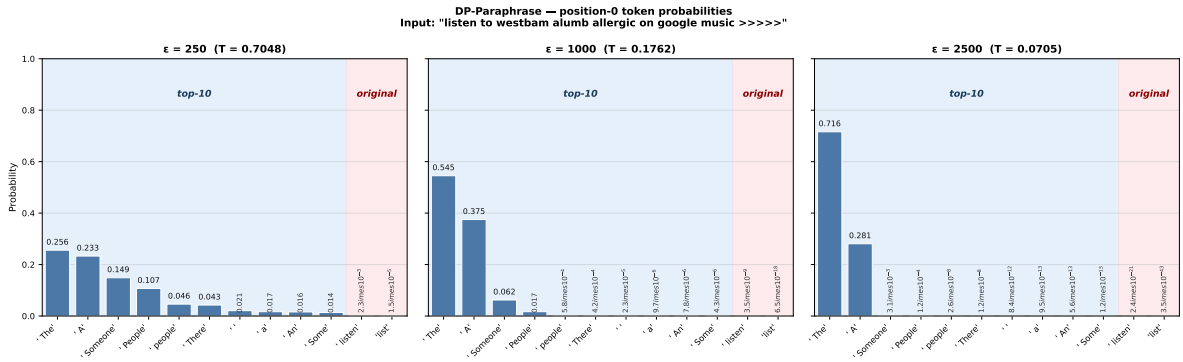

Figure 8: Token probabilities at the first generated position reveal that the GPT-2 paraphraser used by the DP-Paraphrase (as reproduced by Meisenbacher et al. (2024)), fine-tuned on SNLI (Bowman et al., 2015) entailment pairs, has an overwhelming prior toward SNLI-style sentence starters (*e.g.,* ‘ The’ at 71.6% for $\varepsilon = 2500$), while the original token ‘ listen’ has negligible probability ($\leq 2.3 \times 10^{-3}$) at various ε values. Increasing ε further concentrates the distribution on these prior-favored tokens, making the generated output insensitive to the adjustment of ε.

Seed	Output
Ref: After many tries, I got it to work. I hope it's fixed for next time!	
$\varepsilon = 2500$ *(high variance, divergent outputs)*	
42	A while back, I got a call from my friend, Nathan.
1234	The following is a rush transcript. Copy may not be in its final form.
101	Atheist "It's time to get started. I
$\varepsilon = 1000$ *(low variance, semantically consistent)*	
42	If you live in Northern California, it's a good time to try and
1234	If you live in Northern California, it's a good time to try and
101	If you live near Grenfell Tower, it's a good time to try and
Ref: It slows the game performance a bit, but it's totally worth it!	
$\varepsilon = 2500$ *(high variance, divergent outputs)*	
42	The game of the week is a blur, but it's worth a while
1234	The following is a rush transcript. Copy may not be in its final form.
101	The game of the week is a bit of a grind, especially if you'
$\varepsilon = 1000$ *(low variance, consistent pattern)*	
42	It's that time of the year again! But Super Mario Kart 5.
1234	It's that time of the year again! But Super Mario Kart 3.
101	It's that time of the year again! But if you enjoy slow paced

Table 4: DP-BART output variability across random seeds at different ε values. At $\varepsilon = 2500$, outputs exhibit high variance and diverge into generic or unrelated content (*e.g.,* media templates, religious terms), while at $\varepsilon = 1000$, outputs remain semantically consistent across seeds.

Method	Output
Original: listen to westbam alumb allergic on google music	
DP-Paraphrase (BPE (Sennrich et al., 2016) tokenization, 12 tokens)	
Tokenization	['list', 'en', ' to', ' west', 'b', 'am', ' al', 'umb', ' allergic', ' on', ' google', ' music']
$\varepsilon = 250$	A person is♦ trying to Magneto. A person is
$\varepsilon = 1000$	The peoplecats are on the internet. The people are on the
$\varepsilon = 2500$	The people are rainbow colored. The people are on the internet
DP-MLM (NLTK (Bird, 2006) word tokenization, 8 tokens)	
Tokenization	["listen", "to", "westbam", "alumb", "allergic", "on", "google", "music"]
$\varepsilon = 250$	track to adam al a on google music
$\varepsilon = 1000$	listen to adam al a on google music
$\varepsilon = 2500$	list to adam al a on google music
DP-Prompt (SentencePiece (Kudo and Richardson, 2018) tokenization, 12 tokens)	
Tokenization	["listen", "to", "west", "b", "am", "al", "umb", "allergic", "on", "google", "music", "</s>"]
$\varepsilon = 250$	listen to westbam alumb allergic on google music
$\varepsilon = 1000$	listen to westbam alumb allergic on google music
$\varepsilon = 2500$	listen to westbam alumb allergic on google music

Table 5: Comparison of three token-level LDP text rewriting methods on a SNIPS sentence (Coucke et al., 2018), illustrating differences in tokenization and semantic fidelity.

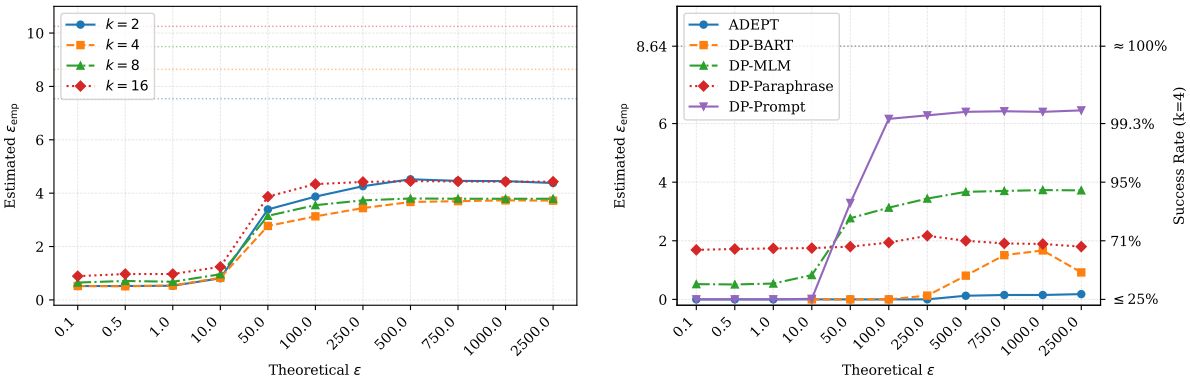

Figure 9: (Left) Effect of candidate set size $k \in \{2, 4, 8, 16\}$ on ε_{emp} for DP-MLM on ATIS under the external embedding attack. (Right) Calibration results under the external embedding attack with $k = 4$ on ATIS.

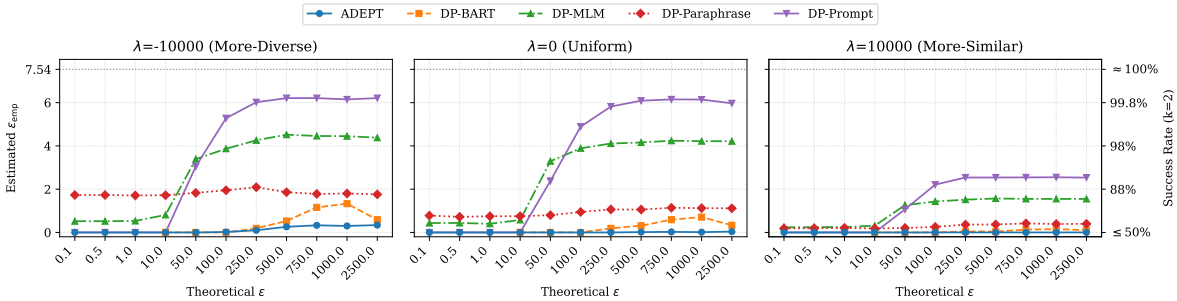

Figure 10: Results under three candidate selection strategies on the ATIS dataset: (a) more-diverse sampling (λ=-10,000), (b) uniform random sampling (λ=0), and (c) more-similar sampling (λ=10,000). While the empirical privacy loss varies depending on the auxiliary information available to the adversary, the relative ranking and trends across different LDP mechanisms remain consistent.

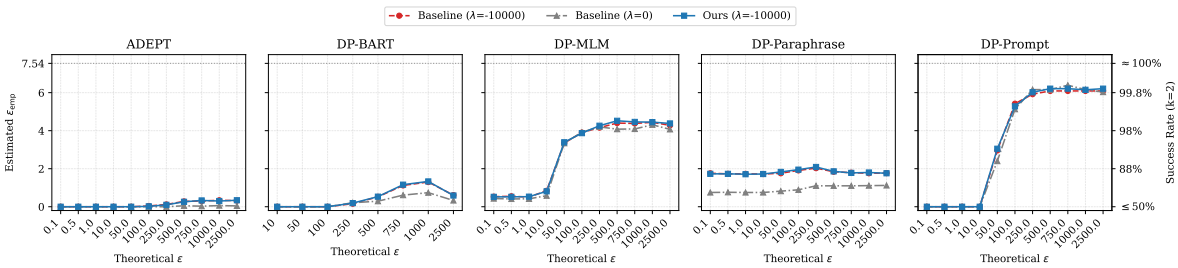

Figure 11: Empirical privacy loss under three configurations on ATIS: baseline (Arcolezi and Gambs, 2024) with more-diverse ($\lambda = -10,000$), uniform ($\lambda = 0$), and our method with more-diverse sampling.