LoSanR: Free Lunches for Differentially Private Text Sanitization

Anonymous ACL submission

Abstract

Differentially private (DP) text sanitization replaces sensitive documents with privacy-preserving alternatives, but methods based on embedding perturbation suffer from the curse of dimensionality. The noise required for privacy grows with the embedding dimension, overwhelming the semantic signal. We identify two *free lunch* techniques that improve utility at zero privacy cost in retrieval-based DP sanitization, where each sensitive sentence is replaced by a match retrieved from a public corpus. First, projecting embeddings into a low-dimensional subspace fitted on *public* data drastically reduces the noise needed without harming privacy. Second, on top of these noise-reduced embeddings, scoring each replacement for coherence with already-selected surrounding sentences (rather than just nearest in embedding space) further improves output quality at no extra privacy cost. We combine these techniques in a system called LoSanR, which across three datasets substantially outperforms prior DP methods on downstream classification at every privacy budget tested, while empirically resisting re-identification under model-free attacks (98.5% at the tightest).

1 Introduction

Differentially private (DP) text sanitization aims to replace sensitive documents with privacy-preserving alternatives that retain utility for downstream tasks. For example, a private movie review such as *"I rented this for family movie night with my husband and two kids; only our 8-year-old stayed awake"* might be replaced with *"This movie was so slow most of my family fell asleep before the ending"*; the sanitized version preserves the sentiment while removing personal context. A growing body of work tackles this problem at various levels of granularity, from individual words or tokens (Yue et al., 2021; Chen et al., 2023; Meisenbacher et al., 2024), to sentences (Meisenbacher et al., 2025c), to entire documents (Igamberdiev and Habernal, 2023).

These methods typically encode text as numerical vectors (embeddings), perturb these vectors under DP, then (attempt to) recover useful, coherent text from the noisy result. Yet, a fundamental bottleneck plagues this approach: the **curse of dimensionality**. The noise required for privacy grows with the dimension of the embedding, and in the high-dimensional spaces used by modern encoders, can overwhelm the semantic signal.

In this paper, we ask: *can we improve the utility of sanitized text without weakening the privacy guarantee?* We propose two techniques that achieve this at zero additional privacy cost:

1. **PCA on *public* data.** A natural direction to solve the curse of dimensionality is to reduce the dimension of the embedding vector before adding noise, e.g., via a linear projection. But the projection must not leak information about the private document itself. Thus, we fit PCA on *public* (non-sensitive) data from the same domain, then project embeddings into the resulting low-dimensional subspace before adding noise. Because PCA is fitted entirely on public data, it consumes no privacy budget, yet it drastically reduces the noise needed for a given privacy guarantee. Such domain-matched public data is readily available in practice: public product reviews, clinical notes, or court filings can serve as pools for their private counterparts.

2. **Retrieval with coherence scoring.** Rather than decoding noisy embeddings directly back into text, we replace each sensitive sentence with a 'best match' from our public corpus. Using the public corpus allows us to take advantage of the relationship between sentences in the same document. We

score replacement candidates for coherence with the already-selected previous replacements and the noisy embeddings of nearby sentences. Since we only use public data and already-privatized embeddings, this constitutes post-processing and again requires no privacy cost.

These free lunches compose: PCA provides most of the headline gain on downstream classification, and context-aware scoring adds further utility on top of PCA's noise reduction while also improving inter-sentence coherence (Section 3).

Beyond these two techniques, we examine the privacy implications of data-dependent budget allocation (Meisenbacher et al., 2025c) and study sentence selection as a complementary way to manage composition cost (Section 4).

Unlike generation-based approaches that require model training and truncate inputs to 20 tokens for tractability (Igamberdiev and Habernal, 2023), our method is training-free, imposes no restriction on input length, and produces interpretable output: each sanitized sentence is a sentence from the public corpus. We combine these ideas in a system called LoSanR (**Lo**w-dimensional **San**itization via **R**etrieval). At the same document-level privacy budgets used by prior methods, LoSanR substantially outperforms prior DP methods on downstream classification across three datasets, while empirically resisting re-identification under model-free attacks (Sec. 3).

2 Method

Our 'free lunches' apply to *retrieval-based* DP text sanitization: pipelines that embed sentences, perturb the embedding under DP, and retrieve a replacement sentence from a public corpus by nearest-neighbor matching. To measure more directly the power of these free lunches, we instantiate with the Analytic Gaussian Mechanism (Balle and Wang, 2018), on sentence embeddings from all-MiniLM-L6-v2, showing that *even with this straightforward mechanism*, the two free lunches close most of the gap to the non-sanitized baseline and substantially outperform prior DP methods at the same document-level privacy budgets (Section 3). We treat as future work (Sec. 6) to apply our methods to other mechanisms.

The free lunches require:

- A public corpus from a similar domain to the private data, used to fit PCA and to source replacement sentences.
- Sentence-level decomposition of documents, so that retrieval and context-aware scoring operate on coherent units.
- A retrieval-based downstream consumer: the perturbed vector is matched against public-corpus embeddings, not fed to a decoder. We note that while our methods require no training, they can be complementary to mechanisms that use LLMs to e.g., rewrite our output sanitized text (see Sec. 5).

2.1 Threat Model and Problem Setup

We consider a *local differential privacy* (LDP) setting (Kasiviswanathan et al., 2011): a data holder wishes to release sanitized versions of sensitive documents satisfying the following guarantee.

Definition 1 ((ε, δ)-LDP for Documents)**.** *A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-LDP if for any two inputs x, x' (two documents) and any measurable set $S \subseteq$ range$(\mathcal{M})$, $\Pr[\mathcal{M}(x) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(x') \in S] + \delta$.*

Intuitively, this bounds how much the output of $\mathcal{M}$ can reveal about any single input: smaller ε means stronger privacy, and δ allows a small failure probability. We use document-level granularity to compare our methods with previous work under matched privacy guarantees.

Our method assumes access to a public reference corpus $\mathcal{P}$ of non-sensitive text from the same domain as the private data, which can be used without privacy cost. Each document is split into sentences $\mathbf{s} = (s_1, \ldots, s_L)$. We then embed each sentence, apply a DP mechanism $\mathcal{M}$ to the embedding, and retrieve a replacement sentence from the public corpus $\mathcal{P}$ using a context-aware scoring function (Section 2.6). The sanitized document is assembled from these replacements. Under basic sequential composition (Dwork et al., 2006), if each sentence satisfies $(\varepsilon_s, \delta_s)$-LDP, the full document satisfies $(L \cdot \varepsilon_s, L \cdot \delta_s)$-DP. For document-level privacy at budget $(\varepsilon_{\text{doc}}, \delta_{\text{doc}})$, we set $\varepsilon_s = \varepsilon_{\text{doc}}/L$ and $\delta_s = \delta_{\text{doc}}/L$.

2.2 Embedding and Perturbation

We describe the Analytic Gaussian Mechanism (AGM) below, though our methods generalize to any DP mechanism for text embeddings.

Let $f : \mathcal{S} \to \mathbb{R}^d$ be a pretrained sentence encoder that maps each sentence to a d-

ε_s	σ	Noise/$\|\mathbf{x}\|$	$\widehat{\cos}$
10	1.00	19.6×	0.051
50	0.30	5.9×	0.168
100	0.19	3.7×	0.259

Table 1: Signal-to-noise ratio in d=384 dimensions at per-sentence budget (ε_s, δ_s=10^{-5}) with sensitivity Δ_2=2. σ is the AGM noise scale; Noise/$\|\mathbf{x}\|$ is $\sigma\sqrt{d}$; $\widehat{\cos}$ is from Eq. 2.

dimensional vector. Each sentence is encoded and ℓ_2-normalized to the unit sphere: $\mathbf{x}_i = f(s_i)/\|f(s_i)\|_2 \in \mathbb{S}^{d-1}$.

The ℓ_2-*sensitivity* required by the AGM (App. L) is the maximum ℓ_2 distance between two encoded sentences $\mathbf{x}, \mathbf{x}'$. Since each $\mathbf{x}_i$ lies on the unit sphere, $\Delta_2 = \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x} - \mathbf{x}'\|_2 = 2$ (the diameter).

We apply the AGM (Balle and Wang, 2018), which calibrates σ to satisfy $(\varepsilon_s, \delta_s)$-DP:

$$\mathbf{y}_i = \mathbf{x}_i + \boldsymbol{\eta}_i, \quad \boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d) \tag{1}$$

The AGM calibration determines σ numerically from ε_s, δ_s, and Δ_2 (smaller ε_s or δ_s require larger σ). The noisy embedding is then re-normalized: $\hat{\mathbf{y}}_i = \mathbf{y}_i/\|\mathbf{y}_i\|_2$.

2.3 The Curse of Dimensionality

The expected norm of the noise vector is $\mathbb{E}[\|\boldsymbol{\eta}\|_2] = \sigma\sqrt{d}$, while the signal (the original embedding) has unit norm. The expected cosine similarity between the original and noisy embedding is approximated as (derivation in App. M):

$$\mathbb{E}\left[\cos(\mathbf{x}, \mathbf{y})\right] \approx \widehat{\cos} := \frac{1}{\sqrt{1 + \sigma^2 d}} \tag{2}$$

This expression reveals the fundamental tension: for a fixed privacy budget ε (which determines σ), increasing d degrades the signal-to-noise ratio. Table 1 illustrates this for d=384 dimensions: even at per-sentence ε_s=100 (quite a loose budget), δ_s=10^{-5}, the noise exceeds the signal by 3.7× and the expected cosine similarity is only 0.26. Nearest-neighbor retrieval under these conditions is poor or close to random. DP-BART (Igamberdiev and Habernal, 2023) faces the same problem in its 15,360-dimensional encoder latent space, achieving only F_1=0.53 (best variant) in a downstream binary classification task at comparable privacy levels.

2.4 Free Lunch: PCA on Public Data

Our key observation is that dimensionality reduction on *public* data is privacy-free. We choose PCA over random projections (Blocki et al., 2012) because it selects the directions that capture the most variation in the data, producing a more informative low-dimensional representation when public data is available.

Let $\mathbf{W}_k \in \mathbb{R}^{d\times k}$ be the top-k principal components fitted on the public corpus embeddings $\{f(p) : p \in \mathcal{P}\}$. Since $\mathcal{P}$ is public, this costs zero privacy budget (Dwork et al., 2006). This assumes that the public and private corpora occupy similar regions of embedding space. In practice, this is common: public product reviews serve as a pool for sanitizing proprietary customer feedback, published clinical notes for hospital records, or public court filings for confidential legal documents.

Rather than perturbing $\mathbf{x}_i$ directly in $\mathbb{R}^d$ as in Sec. 2.2, the pipeline becomes:

1. Project: $\tilde{\mathbf{x}}_i = \mathbf{W}_k^\top \mathbf{x}_i \in \mathbb{R}^k$
2. Normalize: $\bar{\mathbf{x}}_i = \tilde{\mathbf{x}}_i/\|\tilde{\mathbf{x}}_i\|_2 \in \mathbb{S}^{k-1}$ (the unit sphere in $\mathbb{R}^k$)
3. Perturb: $\mathbf{y}_i = \bar{\mathbf{x}}_i + \boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_k)$
4. Re-normalize: $\hat{\mathbf{y}}_i = \mathbf{y}_i/\|\mathbf{y}_i\|_2$

The Gaussian mechanism requires a finite bound on the ℓ_2-sensitivity Δ_2 to calibrate σ. Step 2 provides such a bound by normalizing to the unit sphere, giving Δ_2=2 (the diameter). Alternatives such as norm clipping would also work; we choose unit-sphere normalization because it gives a tight, data-independent bound on Δ_2. Step 4 is a convenience: it lets us use dot products as cosine similarities during retrieval. The sensitivity on $\mathbb{S}^{k-1}$ remains Δ_2=2, so the same Analytic Gaussian calibration applies. However, the noise norm is now $\sigma\sqrt{k}$ instead of $\sigma\sqrt{d}$, yielding a noise reduction factor of $\sqrt{d/k}$. For k=16, this is $\sqrt{384/16} = 4.9\times$. Table 2 shows the resulting improvement in signal-to-noise ratio. This noise-reduction effect is not encoder-specific: App. E validates the same pattern on all-mpnet-base-v2 (d=768), with PCA-16 winning at every ε tested and a 17.5% relative F_1 gain at ε=100.

Why does PCA preserve task-relevant signal despite discarding most variance? Many downstream tasks depend on low-dimensional structure in embedding space. We find empirically that 16 components suffice to recover near-baseline accuracy on downstream tasks (Section 3.2).

ε_s	d=384	k=32	k=16
10	19.6×	5.7×	4.0×
50	5.9×	1.7×	1.2×
100	3.7×	1.1×	**0.76**×

Table 2: Noise-to-signal ratio ($\sigma\sqrt{d}$ or $\sigma\sqrt{k}$) across dimensions. At k=16, ε_s=100, noise is *smaller* than the signal, enabling meaningful nearest-neighbor retrieval.

2.5 Document-Level Composition

For a document with L sentences and a document-level privacy budget ε_{doc}, we allocate $\varepsilon_s = \varepsilon_{\text{doc}}/L$ and $\delta_s = \delta_{\text{doc}}/L$ per sentence. By basic sequential composition (Dwork et al., 2006), the document satisfies $(\varepsilon_{\text{doc}}, \delta_{\text{doc}})$-DP.

This composition accounting is comparable to that of Meisenbacher et al. (2025c), who also split the document budget across sentences (albeit with non-uniform per-sentence allocation rather than equal division). Many prior works (Yue et al., 2021; Chen et al., 2023) report per-word or per-sentence ε without composition, which understates the true document-level privacy cost. On IMDB at ε=100, the gap between per-sentence and proper doc-level accounting is 6–15 F_1 points (App. H). We use composed document-level accounting throughout. As σ depends on $\varepsilon_s = \varepsilon_{\text{doc}}/L$, documents with more sentences require more noise per sentence.

2.6 Free Lunch: Context-Aware Retrieval

Given the noisy embedding $\hat{\mathbf{y}}_i$ for sentence i, we retrieve a replacement from the public pool. To promote document-level coherence, we use a scoring function that balances similarity with contextual coherence.

Candidate selection. Since the scoring function (below) evaluates each candidate in context, we need a manageable candidate set rather than searching the full pool. We cluster the public pool into K groups using K-means on their (PCA-projected, normalized) embeddings, then select the top-r (where r is a constant, which we set in our experiments to be 3) clusters closest to $\hat{\mathbf{y}}_i$ as the candidate set $\mathcal{C}_i$.

Scoring. For each candidate $\mathbf{c} \in \mathcal{C}_i$ (the normalized embedding of a pool sentence), we compute:

$$\begin{aligned}\text{score}(\mathbf{c}) &= w_{\text{sim}} \cdot \hat{\mathbf{y}}_i^\top \mathbf{c} + w_{\text{prev}} \cdot \mathbf{x}'^{\top}_{i-1}\mathbf{c} \\ &+ w_{\text{next}} \cdot \hat{\mathbf{y}}_{i+1}^\top \mathbf{c} + w_{\text{cl}} \cdot \mathbf{1}[\mathbf{c} \in \text{top-1 cluster}]\end{aligned} \tag{3}$$

The four terms capture complementary signals. The first term scores semantic similarity to the current (noisy) sentence. The second promotes coherence with the preceding context: $\mathbf{x}'_{i-1}$ is the clean embedding of the public replacement already chosen for sentence $i-1$, so candidates that fit naturally after it score higher. The third provides a lookahead: $\hat{\mathbf{y}}_{i+1}$ is the noisy embedding of the *next* sentence, favoring candidates that transition well into what follows. The fourth is a small bonus for candidates from the nearest cluster, encouraging topical consistency. All dot products are cosine similarities between unit vectors in $\mathbb{R}^k$, and $\mathbf{1}[\cdot]$ is the indicator function. The replacement for sentence i is $s'_i = \arg\max_{\mathbf{c}\in\mathcal{C}_i} \text{score}(\mathbf{c})$.

Weights, number of clusters, and number of top clusters to consider were set to match the inter-sentence coherence of the public corpus; downstream classification results are robust to reasonable variations of these parameters (see Section 3.2.2). For boundary sentences (the first has no previous replacement; the last has no next embedding), the applicable weights are renormalized to sum to 1.

Privacy. The scoring function uses only (a) noisy embeddings $\hat{\mathbf{y}}_i$ (already privatized), (b) previously selected public sentences $\mathbf{x}'_{i-1}$ (public pool data), and (c) cluster assignments derived from public data. No additional privacy budget is consumed. Like PCA, context-aware retrieval is a free lunch: it improves replacement quality through post-processing at zero privacy cost.

Algorithm 1 summarizes the full pipeline for reference; the privacy guarantee follows next.

Theorem 1 (Proof in App. N)**.** *Algorithm 1 satisfies* $(\varepsilon_{doc}, \delta_{doc})$*-LDP.*

3 Experiments

3.1 Setup

Datasets. We evaluate on three classification benchmarks: **IMDB** (Maas et al., 2011) (binary sentiment on movie reviews, ∼10 sentences/review), **Amazon Polarity** (He and McAuley, 2016) (binary sentiment on product reviews, ∼4.7 sentences/review), and **AG News** (Zhang et al., 2015) (4-class topic classification on news articles, ∼1.5 sentences/article). Each dataset's standard test set is used only for downstream evaluation. From the training portion, we use 12,500 documents as the *public pool* (for

Algorithm 1 LoSanR document sanitization

Require: document d, sentence encoder f, public pool $\mathcal{P}$, privacy parameters $(\varepsilon_{\text{doc}}, \delta_{\text{doc}})$

// Encode (deterministic)
Fit PCA matrix $\mathbf{W}_k$ on public pool embeddings $\{f(p) : p \in \mathcal{P}\}$
Split d into sentences $(s_1, \ldots, s_L)$
Allocate $\varepsilon_s \leftarrow \varepsilon_{\text{doc}}/L$, $\delta_s \leftarrow \delta_{\text{doc}}/L$
for $i = 1, \ldots, L$ **do**
  $\mathbf{x}_i \leftarrow f(s_i)/\|f(s_i)\|_2$
  $\bar{\mathbf{x}}_i \leftarrow \mathbf{W}_k^\top \mathbf{x}_i / \|\mathbf{W}_k^\top \mathbf{x}_i\|_2$
// Privatize (AGM per sentence)
for $i = 1, \ldots, L$ **do**
  $\mathbf{y}_i \leftarrow \bar{\mathbf{x}}_i + \boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_k)$ with σ calibrated for $(\varepsilon_s, \delta_s)$ via AGM
  $\hat{\mathbf{y}}_i \leftarrow \mathbf{y}_i / \|\mathbf{y}_i\|_2$
// Retrieve (post-processing)
for $i = 1, \ldots, L$ **do**
  $s'_i \leftarrow \arg\max_{\mathbf{c} \in \mathcal{C}_i} \text{score}(\mathbf{c})$ ($\mathbf{c}$ is a candidate pool embedding; score from Eq. 3)
return sanitized document $(s'_1, \ldots, s'_L)$

PCA fitting and the retrieval corpus) and draw 200 from the remaining training documents as the *private set*. The small private set simulates a realistic scenario where only a limited collection of sensitive documents is available.

Mechanism. Sentences are embedded with all-MiniLM-L6-v2 (d=384). App. E validates the same pattern with all-mpnet-base-v2 (d=768). PCA is fitted on the public pool embeddings and projects to k dimensions; we sweep $k \in \{2, 4, 8, 16, 32, 64\}$ and compare against the full 384-dim baseline (no PCA). We evaluate document-level $\varepsilon_{\text{doc}} \in \{100, 250, 500, 1000\}$. For δ_{doc} we match DP-BART's per-dataset choices: δ_{doc}=10^{-5} on IMDB and AG News, δ_{doc}=10^{-7} on Amazon. Each sentence receives $\varepsilon_s = \varepsilon_{\text{doc}}/L$ and $\delta_s = \delta_{\text{doc}}/L$. Retrieval clustering and scoring weight details are in App. K.

Downstream evaluation. We evaluate LoSanR under two downstream classifier protocols. *LoSanR (LR)* uses logistic regression on per-document representations (mean of each document's sanitized sentence embeddings) over the 200 documents in our private set, consistent with the limited-data scenario described above. *LoSanR (BERT)* relaxes the 200-doc limit and uses bert-base-uncased fine-tuned on 12,500 LoSanR-sanitized documents (the full private half of each train split), matching DP-BART's evaluation pipeline for a more direct comparison. The primary metric is macro-averaged F_1. Training documents are LoSanR-sanitized; test sets are unsanitized. BERT fine-tuning details are in App. J.

Baselines. We compare against published state-of-the-art results:

- **DP-BART** (Igamberdiev and Habernal, 2023): DP-BART is the de facto baseline for document-level DP via embedding perturbation, providing rigorous LDP guarantees and an open-source implementation that subsequent work compares against and extends (Meisenbacher et al., 2025c). We compare against the best F_1 across DP-BART variants (CLV, PR, PR+) and ADePT from Igamberdiev and Habernal (2023).
- **Budget Wisely** (Meisenbacher et al., 2025c): Extends DP-BART to sentence-level with non-uniform per-sentence budget allocation. We compare against their IMDB $F_1 \approx 0.49$ at ε=500, evaluated with DeBERTa-v3-base.

LoSanR (BERT) outperforms DP-BART and Budget Wisely at every comparable cell (Tables 3 and 4). The minimal LR setup also outperforms DP-BART on IMDB and at low ε on Amazon.

3.2 Main Results

Table 3 shows the full PCA dimension sweep on IMDB, and Table 4 summarizes F_1 across all three datasets under both LR and BERT protocols. The central finding: under the matched BERT protocol, LoSanR outperforms DP-BART at every comparable cell, by margins ranging from 2 F_1 points (Amazon at ε=1000) to 45 points (Amazon at ε=100). We analyze PCA dimensions and baselines below, then evaluate context-aware retrieval.

3.2.1 Analysis

Effect of PCA dimension. On IMDB (Table 3), k=16 achieves the highest F_1 across all ε values, reaching 0.702 at ε=500. The benefit is largest when noise dominates. At ε=100, PCA-16 improves F_1 from 0.497 (no PCA) to 0.639, a 29% relative gain. The optimal k varies across datasets (App. D), but k=16 stays within 3 F_1 points of the best on every dataset and budget, making it a robust default.

Method	ε_{doc} 100	250	500	1000
k=2	.546	.550	.550	.535
k=4	.572	.614	.620	.624
k=8	.599	.643	.688	.686
k=16	**.639**	**.679**	**.702**	**.703**
k=32	.585	.642	.657	.700
k=64	.607	.639	.689	.667
No PCA (d=384)	.497	.573	.605	.694
DP-BART†	.46	.55	.53	.60
Bdg. Wisely†	—	—	.49	.50
No-DP baseline		*.719*		

Table 3: F_1 (macro) on IMDB under document-level DP. LoSanR (LR) rows shown across PCA dimensions k (or no PCA at d=384). See Table 4 for matched-protocol BERT comparison. †Best across DP-BART variants (CLV, PR, PR+) and ADePT from Igamberdiev and Habernal (2023).

Dataset	Method	ε_{doc} 100	250	500	1000
IMDB	LoSanR (LR)	.639	.679	.702	.703
	LoSanR (BERT)	**.815**	**.822**	**.835**	**.834**
	DP-BART†	.46	.55	.53	.60
Amazon	LoSanR (LR)	.643	.685	.694	.712
	LoSanR (BERT)	**.820**	**.846**	**.866**	**.869**
	DP-BART†	.37	.82	.83	.85
AG News	LoSanR (LR)	.787	.788	.800	.795
	LoSanR (BERT)	**.878**	**.888**	**.894**	**.890**

Table 4: F_1 (macro) across datasets and downstream protocols. LoSanR (BERT) outperforms DP-BART† at every comparable cell. BERT numbers are single-seed final-epoch point estimates (App. J). †Best across DP-BART variants (CLV, PR, PR+) and ADePT from Igamberdiev and Habernal (2023).

Cross-dataset comparison with baselines. LoSanR (BERT) reaches F_1=.82−.89 at ε=500 across all three datasets (Table 4), spanning binary sentiment (IMDB, Amazon) and 4-class topic classification (AG News), and outperforms DP-BART at every comparable cell. The advantage is largest at low ε: on Amazon at ε=100, DP-BART falls to F_1=.37 while LoSanR (BERT) reaches .82 (a 45-point gap); on IMDB at ε=100, .46 vs. .82. At high ε on Amazon, both methods reach F_1>.83 and the gap narrows to 2–4 points, but LoSanR retains a clear margin throughout. DP-BART's degradation at low ε extends beyond F_1: at ε=100 it produces whitespace-only output for 99.9% of inputs, and outputs remain essentially unreadable until ε=500 (App. C).

	ε_{doc}	F_1 Plain NN	F_1 Context	Coherence Plain NN	Coherence Context
IMDB	100	.532	**.562**	.011	.388
	250	.698	**.707**	.053	.410
	500	.715	**.723**	.101	.443
	1000	.726	**.731**	.153	.480
	original	—		*.259*	
Amzn	100	**.675**	.650	.088	.424
	250	.682	**.691**	.201	.499
	500	.704	**.710**	.266	.565
	1000	.709	**.721**	.314	.595
	original	—		*.390*	

Table 5: Retrieval ablation on IMDB and Amazon. Context-aware scoring vs. plain nearest-neighbor, both with PCA-16. Coherence is mean cosine similarity between consecutive replacement sentences (a proxy metric); F_1 gain is shown alongside. *Original* shows coherence of the private documents before sanitization.

Taken together. This pattern of widening gap at low ε is exactly what our curse-of-dimensionality framing (Section 2.3) predicts: the 4.9× noise reduction from PCA-16 matters most where noise dominates the signal. The gap over DP-BART is therefore structural rather than a constant offset, with the widest margins at the tightest budgets.

3.2.2 Retrieval Ablation

To isolate the contribution of context-aware retrieval (Section 2.6), we compare against plain nearest-neighbor retrieval, which selects the public sentence with highest cosine similarity to the noisy embedding without considering surrounding context.

Table 5 shows the effect on IMDB and Amazon (AG News is excluded because 60% of its documents are single-sentence, leaving no consecutive pairs to measure coherence on). The coherence effect is striking. Plain NN produces documents where consecutive sentences are essentially unrelated (coherence near zero at low ε), while context-aware scoring restores embedding-level coherence to within or above the original-document range. The values exceeding the original partly reflect smoothing toward corpus-typical phrasing; we treat the metric as a proxy for inter-sentence cohesion, not faithful reproduction. The F_1 gain is consistent on IMDB (+0.5 to +3.0 points). On Amazon, context-aware scoring improves F_1 at $\varepsilon \geq 250$ but slightly underperforms plain NN at ε=100, where the shorter documents (∼4.7 sentences) leave less room for contextual signals to help.

PCA	Retrieval	ε_{doc} 100	250	500	1000
No	Plain NN	.599	.647	.622	.680
No	Context	.430	.549	.597	.667
Yes	Plain NN	.532	.698	.715	.726
Yes	Context	**.562**	**.707**	**.723**	**.731**

Table 6: Disentangling the two free lunches on IMDB. Context-aware scoring additionally improves coherence (Table 5).

To further disentangle the contributions of PCA and context-aware retrieval, Table 6 shows a 2×2 ablation on IMDB. PCA is the main improvement, adding 9–13 F_1 points at ε=500 regardless of retrieval method. Context-aware scoring provides additional gains with PCA (+0.5 to +3 points) but can hurt without it, particularly at low ε where noise in 384 dimensions overwhelms the contextual signals. The primary role of context-aware scoring is inter-sentence coherence (Table 5), a quality not captured by F_1; the F_1 gains are an additional benefit.

4 Budget Allocation

Data-Dependent Allocation: Privacy Risks. Within sentence-level sanitization, a natural question is how to allocate the privacy budget across sentences. Budget Wisely (Meisenbacher et al., 2025c) assigns per-sentence privacy budgets $\varepsilon_i = f(s_i)$ based on properties of each sentence s_i (POS composition, importance scores, etc.), with the document-level guarantee claimed as $\varepsilon_{\text{doc}} = \sum_i \varepsilon_i$. However, the allocation function f is *deterministic and data-dependent*: it examines the private text to determine how much budget each sentence receives. An observer who sees the allocation pattern (or can infer it from the sanitized output) learns that sentence i was "important" when ε_i is large. This leakage is not explicitly accounted for in their privacy analysis. Even in a threat model that does not reveal the allocation of privacy budgets, they affect the mechanism's output distribution. This effect, if not accounted for, may have unintended consequences for privacy leakage.

A more careful formulation would (1) privately (i.e., with noise) compute privacy budget $\varepsilon_i = f(s_i)$ for each sentence, then (2) perturb each s_i with budget ε_i. To do (1), one must thus spend part of the total budget to privately release the allocation itself, then use the remaining budget for perturbation. This additional cost can offset the gains from non-uniform allocation (and when considering expected retrieval quality, *uniform* allocation is in any case optimal, see below).

Sentence Selection. LoSanR's uniform allocation $\varepsilon_s = \varepsilon_{\text{doc}}/L$ avoids these issues entirely: the per-sentence budget depends only on L (the number of sentences), which is a structural property revealed by the sanitized output regardless. Standard basic composition applies directly.

Could *non-uniform* allocation improve utility? We observe that for a downstream classifier that averages sentence embeddings, non-uniform allocation is provably suboptimal. Let $u(\varepsilon)$ denote the expected retrieval quality at per-sentence budget ε. From Eq. 2, $u(\varepsilon) \approx 1/\sqrt{1 + c/\varepsilon^2}$ for a constant c depending on d, Δ_2, and δ, which is *concave* in ε. By Jensen's inequality, for any allocation $(\varepsilon_1, \ldots, \varepsilon_L)$ with $\sum_i \varepsilon_i = \varepsilon_{\text{doc}}$, $\frac{1}{L}\sum_i u(\varepsilon_i) \leq u(\varepsilon_{\text{doc}}/L)$. That is, uniform allocation maximizes the average quality.

However, averaging L sentence embeddings dilutes the signal. At $\varepsilon_{\text{doc}} = 100$ with $L = 10$ sentences, each gets $\varepsilon_s = 10$, yielding poor retrieval quality. Instead of allocating budget differently, we can *select fewer sentences*. If the downstream classifier uses only the first $N < L$ sentences, we need only sanitize those N sentences, each receiving ε_s=$\varepsilon_{\text{doc}}/N$. This is a fixed, data-independent policy. N mechanism calls at $(\varepsilon_{\text{doc}}/N, \delta_{\text{doc}}/N)$-DP each compose to $(\varepsilon_{\text{doc}}, \delta_{\text{doc}})$-DP.

Empirically (App. G), sentence selection substantially improves F_1 at low ε, with +16.7 F_1 points at ε=100 by sanitizing only N=1 sentence, and N=10 best at ε=500. The optimal N trades per-sentence budget against signal coverage, and the technique is complementary to PCA. PCA reduces per-sentence noise by $\sqrt{d/k}$, and sentence selection reduces composition cost by L/N.

5 Related Work

DP text generation. DP-BART (Igamberdiev and Habernal, 2023) applies the Analytic Gaussian Mechanism to the BART encoder output, truncating inputs to 20 tokens due to the high dimensionality of the latent space. Meisenbacher et al. (2025c) extend DP-BART to sentence-level processing with non-uniform per-sentence budget allocation. DP-Prompt (Utpala et al., 2023) uses an LLM as a utility function for the exponential mechanism to generate a replacement document word-by-word. PrivRewrite (Kim, 2025) extends

the exponential-mechanism to black-box LLM access, with a refined sensitivity bound on the selection utility. Just Rewrite It Again (Meisenbacher and Matthes, 2024) post-processes sanitized text with an LLM. DP-GTR (Li et al., 2025) uses LLM-based group text rewriting in a three-stage DP framework. In contrast, LoSanR presents improvements that are orthogonal (and complementary) to LLM rewrites in the loop.

DP word and token perturbation. SanText (Yue et al., 2021) applies the Exponential Mechanism at the word level with metric-based utility over word embeddings. CusText (Chen et al., 2023) improves upon SanText with customized word groupings. CluSanT (Awon et al., 2025) introduces token clustering with LLM-based substitution to balance privacy and semantic coherence. DP-MLM (Meisenbacher et al., 2024) uses masked language models for token-level replacement. Other word-level methods include TEM (Carvalho et al., 2023), Dyntext (Zhang et al., 2025), ADePT (Krishna et al., 2021; Habernal, 2021), and NADP (Bollegala et al., 2023), each varying the noise mechanism or word grouping strategy. Mattern et al. (2022) identify fundamental limitations of word-level DP, including the inability to obfuscate syntactic information and the failure to privatize full utterances.

DP sentence-level methods. Meehan et al. (2022) formalize sentence-level privacy for document embeddings. DP-ST (Meisenbacher et al., 2025a) decomposes documents into semantic triples, applies the exponential mechanism, and reconstructs text with an LLM. In contrast, LoSanR requires no training, but can be complementary to mechanisms rewriting text with LLMs. Bollegala et al. (2025) propose CMAG, a metric-DP mechanism for sentence embeddings using the Mahalanobis metric. Metric DP is a relaxation of standard DP that scales protection with input distance; our work operates under standard LDP.

Dimensionality reduction and DP. Random projections via the Johnson-Lindenstrauss transform preserve DP as the projection is data-independent (Blocki et al., 2012; Kenthapadi et al., 2013). In addition, Du et al. (2023) use trainable linear maps for dimension reduction of sentence embeddings (in the context of *metric* local DP). LoSanR uses PCA instead, which selects the highest-variance directions from public data and retains more task-relevant signal per dimension. We chose PCA because domain-matched public data is assumed to be available in our setting, and PCA concentrates more signal into fewer dimensions than random projections (App. F).

Privacy attacks on DP-sanitized text. Recent work probes whether DP-sanitized text can be reverse-engineered. Pang et al. (2024) show DP-sanitized text can be vulnerable to reconstruction attacks, and Meisenbacher et al. (2025b) identify a contextual vulnerability whereby surrounding context amplifies reconstruction risk. We examine LoSanR's reconstruction surface empirically (App. A) and find no content-bearing leakage.

6 Discussion and Conclusion

LoSanR's two free lunches yield F_1=.81−.89 across the three datasets, outperforming SOTA. Furthermore, 98.5% of private documents resist embedding-similarity re-identification at the tight-budget end (ε=100; App. B).

Empirical privacy at the utility frontier. The formal $(\varepsilon_{\text{doc}}, \delta_{\text{doc}})$-LDP bound is e^{ε}, which is too loose to be operationally meaningful at the budgets used in DP-NLP papers (a point already acknowledged by Igamberdiev and Habernal (2023) for DP-BART). The relevant question is empirical: how much does the sanitized output actually leak? We measure this for LoSanR under a model-free embedding-similarity MIA (App. B). At ε=100, 98.5% of private documents resist top-1 identification out of 200 candidates (top-1 re-id 1.5% vs. 0.5% random), despite the formal bound being vacuous. Resistance scales with the budget: at looser ε, more signal is preserved for downstream utility (Table 4) and the embedding-similarity adversary correspondingly recovers more documents. This is the operational utility-privacy trade-off intrinsic to retrieval-based DP at high ε: usable output requires preserving signal that an adversary can exploit, and the empirical curve quantifies the trade.

Our results show the potential of LoSanR: **future work** includes applying the same free-lunch techniques to other related retrieval-based DP pipelines, or as a retrieval head on top of embedding sanitization methods such as CluSanT (Awon et al., 2025) or Sentence-LDP (Du et al., 2023).

Code. `https://anonymous.4open.science/r/losanr-091E`, MIT licence.

Limitations

Public corpus. LoSanR requires domain-matched public data for both PCA fitting and retrieval; retrieval quality degrades if the corpus does not cover the private data's topics and vocabulary. This is a stronger requirement than methods that rely only on pretrained public encoders (DP-BART, SanText, DP-MLM).

Evaluation scope. The sentence selection results (Section 4) assume that early sentences carry task-relevant signal, which holds for reviews but may not generalize. We report empirical privacy checks against two attack surfaces: sentence-level reconstruction (App. A), flagged by Pang et al. (2024) and Meisenbacher et al. (2025b), and model-free membership inference (App. B). The MIA result shows that empirical re-identification grows monotonically with ε, as the theory of DP at large ε predicts; we have not evaluated stronger attacks (shadow-classifier MIAs in the style of Du et al. (2023), or LLM-based reconstruction in the style of Pang et al. (2024)).

Coherence above the original. Context-aware retrieval produces inter-sentence coherence that can exceed the original documents (Table 5).

We used an AI coding assistant (Claude Code) to help with experiment scripting and paper editing; all scientific contributions were made by the authors.

References

Ahmed Musa Awon, Yun Lu, Shera Potka, and Alex Thomo. 2025. CluSanT: Differentially private and semantically coherent text sanitization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

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 (ICML)*, pages 403–412. PMLR.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland. Association for Computational Linguistics.

Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. 2012. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 410–419.

Danushka Bollegala, Shuichi Otake, Tomoya Machide, and Ken-ichi Kawarabayashi. 2023. A Neighbourhood-Aware Differential Privacy Mechanism for Static Word Embeddings. *Preprint*, arXiv:2309.10551.

Danushka Bollegala, Shuichi Otake, Tomoya Machide, and Ken-ichi Kawarabayashi. 2025. A Metric Differential Privacy Mechanism for Sentence Embeddings. *ACM Trans. Priv. Secur.*, 28(2):20:1–20:34.

Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: High Utility Metric Differential Privacy on Text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 883–890. Society for Industrial and Applied Mathematics.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758.

Minxin Du, Xiang Yue, Sherman S. M. Chow, and Huan Sun. 2023. Sanitizing sentence embeddings (and labels) for local differential privacy. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2349–2359, New York, NY, USA. Association for Computing Machinery.

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

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.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 507–517.

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.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith.

2011. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826.

Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. 2013. Privacy via the Johnson-Lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1):39–71.

Jongwook Kim. 2025. PrivRewrite: Differentially private text rewriting under black-box access with refined sensitivity guarantees. *Applied Sciences*, 15:11930.

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.

Mingchen Li, Heng Fan, Song Fu, Junhua Ding, and Yunhe Feng. 2025. DP-GTR: Differentially Private Prompt Protection via Group Text Rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1573–1585.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 142–150.

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.

Casey Meehan, Khalil Mrini, and Kamalika Chaudhuri. 2022. Sentence-level privacy for document embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3367–3380.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. 2025a. 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 (EMNLP)*, pages 8976–8992.

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*.

Stephen Meisenbacher, Alexandra Klymenko, Andreea-Elena Bodea, and Florian Matthes. 2025b. The double-edged sword of LLM-based data reconstruction: Understanding and mitigating contextual vulnerability in word-level differential privacy text sanitization. In *Proceedings of the 24th Workshop on Privacy in the Electronic Society (WPES@CCS)*.

Stephen Meisenbacher, Chaeeun Joy Lee, and Florian Matthes. 2025c. Spend your budget wisely: Towards an intelligent distribution of the privacy budget in differentially private text rewriting. In *Proceedings of the 15th ACM Conference on Data and Application Security and Privacy (CODASPY)*, pages 84–95.

Stephen Meisenbacher and Florian Matthes. 2024. Just Rewrite It Again: A Post-Processing Method for Enhanced Semantic Similarity and Privacy Preservation of Differentially Private Rewritten Text. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ARES '24, pages 1–11, New York, NY, USA. Association for Computing Machinery.

Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, Minhui Xue, and Bo Zhang. 2024. Reconstruction of differentially private text sanitization via large language models. *arXiv preprint arXiv:2410.12443*.

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. IEEE.

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.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866.

Juhua Zhang, Zhiliang Tian, Minghang Zhu, Yiping Song, Taishu Sheng, Siyi Yang, Qiunan Du, Xinwang Liu, Minlie Huang, and Dongsheng Li. 2025. DYNTEXT: Semantic-Aware Dynamic Text Sanitization for Privacy-Preserving LLM Inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20243–20255, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, pages 649–657.

Appendices

Appendices provide additional empirical results (§A–§H), qualitative examples (§I), implementation details (§K), and theoretical derivations and proofs (§L–§N) supporting the main paper.

A Reconstruction Risk Analysis

This appendix examines reconstruction risk at two scales: surface-level n-gram overlap (BLEU, Table 7) and sentence-level semantic similarity to the original (Table 8). The sentence-level analysis probes the attack surface flagged for DP-sanitized text by Pang et al. (2024) and Meisenbacher et al. (2025b). Both analyses indicate that similarity between sanitized and original text reflects sentiment alignment, not content reconstruction.

Surface-level overlap (BLEU). Table 7 reports BLEU on the sanitized output. LoSanR's BLEU stays below 0.01 on IMDB and Amazon at every ε, and below 0.10 on AG News (where shorter documents and topical news vocabulary produce more incidental n-gram overlap). For context, paraphrase generation systems typically reach BLEU around 0.11–0.34 on standard benchmarks (Bandel et al., 2022); our IMDB and Amazon values are an order of magnitude lower, and AG News is below the lower end (though closer to it due to its very short documents). The mechanism behind this is direct: each replacement sentence is retrieved from the public corpus, so any shared n-grams with the original arise incidentally.

Dataset	ε=100	ε=250	ε=500	ε=1000
IMDB	.006	.007	.007	.009
Amazon	.007	.008	.009	.011
AG News	.019	.044	.082	.096

Table 7: LoSanR sanitized-output BLEU (PCA-16, context-aware retrieval). Values are NLTK sentence-BLEU averaged over the private documents, on a 0–1 scale.

Sentence-level nearest-original similarity. For each sanitized sentence we measured the cosine similarity in raw MiniLM space (d=384) to its position-matched original, over 2,079 sentences across the 200 IMDB private documents. Table 8 reports the distribution across the full ε grid.

Mean aligned similarity grows monotonically with ε, staying at or below the within-document baseline (0.291) through ε=500 and rising modestly above it (by 0.045) at ε=1000. The 95th percentile remains below 0.7 at every ε, and only 1.25% of pairs cross 0.7 at ε=1000 (versus 0.24% at ε=500).

ε_{doc}	Mean	Median	95th	Max	≥ 0.7	≥ 0.95
100	0.206	0.196	0.406	0.676	0	0
250	0.249	0.245	0.460	0.722	1	0
500	0.290	0.284	0.518	1.000	5	1
1000	0.336	0.328	0.571	1.000	26	3

Table 8: Sentence-level nearest-original cosine similarity on IMDB (2,079 pairs per ε). Within-document consecutive-sentence baseline: 0.291. "$\geq$ 0.7" and "$\geq$ 0.95" columns count pairs exceeding the threshold.

We inspected the highest-similarity pairs. The four pairs with similarity $\geq$ 0.95 across $\varepsilon \in \{500, 1000\}$ are punctuation runs (`!!!!!!!!!!!!`, `!!!!!!!` matched to `!!!!!!!!!`) or 2–4 word generic sentiment templates (*"Worth seeing."*, *"This show is WONDERFUL!"*). Pairs in $[0.64, 0.95)$ are uniformly generic sentiment phrases shared by many public-corpus reviews (variants of *"I highly recommend this movie"* or *"worst movie I've ever seen"*). At ε=100, the maximum aligned similarity is 0.676 and top pairs share only topical register: the original *"a dreadfully dull waste of time"* pairs with a sanitized *"terrible acting... boring and predictable"* (both negative but lexically distinct). The only attribute reconstructable from these near-matches is sentiment polarity, which is already the public classification label. We observe no reconstruction of content-bearing private information at any tested ε.

B Empirical Membership Inference

The formal $(\varepsilon_{\text{doc}}, \delta_{\text{doc}})$-LDP guarantee bounds pairwise indistinguishability by e^{ε}, which is vacuous at the budgets used in this paper. The empirical question is whether actual leakage is much smaller than the formal bound.

Following the membership inference framework of Shokri et al. (2017), we measure this with a model-free embedding-similarity MIA on IMDB. For each sanitized document S_i we compute the mean of its sentence embeddings (raw MiniLM, d=384) and score it against the 200 originals $\{O_j\}$ (same mean-pooled representation) via cosine similarity. We use two metrics. The *200-way re-identification* metric asks,

Method	Top-1	Top-5	Top-10	Rank	AUC	Win
Unsanitized (ceiling)	100%	100%	100%	1.0	1.00	100%
Random-replace (floor)	0.5%	3.0%	5.5%	96.3	0.52	51%
LoSanR ε=100	1.5%	16.0%	24.5%	58.6	0.66	73.5%
LoSanR ε=250	17.0%	39.0%	46.5%	28.7	0.79	85.5%
LoSanR ε=500	26.5%	55.0%	69.0%	13.9	0.86	95.0%
LoSanR ε=1000	46.0%	77.0%	87.5%	6.1	0.91	97.5%

Table 9: Empirical MIA on IMDB. Top-k: fraction of sanitized docs whose true original ranks in the top-k of 200 candidates (random baselines: 0.5%, 2.5%, 5%). Rank: mean rank of true original (random = 100.5; perfect = 1). Win: fraction of i where $\cos(S_i, O_i) > \cos(S_i, D_i)$ for a uniform-random decoy D_i (random = 50%).

for each i, whether $j{=}i$ is ranked first among the 200 candidates. The *decoy-based pairwise* metric asks whether $\cos(S_i, O_i) > \cos(S_i, D_i)$, where D_i is a uniform random decoy from the 12,500 public pool. We compare LoSanR across $\varepsilon \in \{100, 250, 500, 1000\}$ against two reference rows: unsanitized originals (no-privacy ceiling) and random-replacement (each private sentence replaced by a uniform-random public sentence, the perfect-privacy floor). Results are reported in Table 9.

The floor and ceiling are calibrated. Random-replacement sits at the chance baseline on every metric (top-1 0.5%, AUC 0.52, win 51%), confirming the attack itself has no built-in bias. At ε=100, LoSanR provides measurable empirical resistance despite the formal LDP bound being vacuous. 98.5% of private documents resist top-1 re-identification, and the embedding-similarity adversary's pairwise win of 73.5% reflects only coarse sentiment-and-topic signal rather than document-identifying content, since LoSanR's cluster-matched replacements preserve sentiment/topic by construction independent of the specific source document.

At looser budgets, the same signal preservation that supports the F_1=0.81–0.89 classification utility (Table 4) also weakens empirical resistance. Top-1 re-identification rises to 17, 26.5, and 46% at $\varepsilon \in \{250, 500, 1000\}$. The formal LDP guarantee continues to hold. The empirical curve quantifies actual leakage at budgets where the formal LDP bound is vacuous.

Caveat. The attack is model-free (mean MiniLM embedding plus cosine), and is therefore a *lower bound* on adversary capability.

C LLM-Judged Fluency

To complement the cosine-similarity coherence in Table 5 and the downstream F_1 in Table 4, we measured the fluency of sanitized output using Claude Opus 4.7 as judge. For each of three privacy budgets ($\varepsilon_{\text{doc}} \in \{100, 250, 500\}$) we sampled 30 IMDB private documents under three conditions: LoSanR (PCA-16 with context-aware retrieval), DP-BART, and the unsanitized original. The judge scored each document on a 1–5 Likert scale (1 = ungrammatical word salad; 5 = natural, well-formed English), with the condition order randomized per document and condition labels withheld from the judge.

ε_{doc}	LoSanR	DP-BART
100	4.67 ± 0.48	whitespace†
250	4.90 ± 0.31	1.10 ± 0.31
500	4.97 ± 0.18	2.70 ± 0.47

Table 10: LLM-judged document fluency on IMDB (1–5 Likert; mean $\pm$ std over 30 documents per cell). Unsanitized original anchor: 5.00 ± 0.00. †DP-BART at ε=100 produced whitespace-only output for 24,984/25,000 (99.9%) of inputs; not scored on the Likert scale.

LoSanR stays near the original-text ceiling at every privacy budget. DP-BART produces whitespace-only output at ε=100 (99.9% of inputs), hallucinated news-wire fragments unrelated to the input review at ε=250 (e.g., *"DALLAS, N.C. (DCCI) – a 33"* or *"ISMATVEMIA, Wis. (DTV) – Unactled in"*), and partly-readable fragments at ε=500. The qualitative gap is intrinsic to the two approaches: LoSanR returns real public-corpus sentences (fluent by construction), whereas DP-BART decodes from a noisy latent (fluency degrades with ε, and at low ε the decoder produces no readable text).

Caveats. The judgments use a single LLM judge (Claude Opus 4.7) with no inter-rater check. We focus on the fluency axis because the gap (more than two Likert points at every budget where DP-

	ε_{doc}			
k	100	250	500	1000
2	.499	.522	.462	.494
4	.549	.634	.615	.615
8	**.685**	.679	.703	.695
16	.667	.676	.709	.724
32	.623	.646	.667	.721
64	.620	**.692**	**.722**	**.725**
No PCA (d=384)	.611	.656	.690	.730
No-DP baseline		*.731*		

Table 11: F_1 (macro) on Amazon under document-level DP across PCA dimensions k. Best PCA-k per ε in bold. PCA's benefit over No-PCA is largest at low ε (gap of 7 F_1 points at ε=100) and shrinks at high ε (No-PCA catches up at ε=1000), consistent with the curse-of-dimensionality framing in Section 2.3. k=16 stays within 3 F_1 points of the best k at every budget.

	ε_{doc}			
k	100	250	500	1000
2	.584	.641	.623	.579
4	.765	.762	.764	.761
8	.781	.795	.799	.792
16	**.787**	.788	.799	.796
32	.737	**.811**	.778	.814
64	.705	.806	**.818**	**.815**
No PCA (d=384)	.696	.780	.803	.810
No-DP baseline		*.827*		

Table 12: F_1 (macro) on AG News (4-class topic classification) under document-level DP. Best PCA-k per ε in bold. PCA's benefit is largest at low ε (gap of 9 F_1 points at ε=100) and narrows at high ε (within 1 point of No-PCA at ε=1000). Larger k helps on this 4-class task, but k=16 stays within 3 F_1 points of the best k everywhere.

BART produces output) is large enough that plausible judge-model bias does not close it.

D PCA Dimension Sweep on Amazon and AG News

Table 3 shows the full k sweep on IMDB. Tables 11 and 12 extend this to Amazon and AG News.

The optimal k varies modestly across datasets. On IMDB, k=16 wins at every ε. On Amazon, k=8 wins at ε=100 and k=64 wins at $\varepsilon \geq 250$. On AG News, a 4-class topic-classification task, k=16 wins at ε=100, k=32 at ε=250, and k=64 at $\varepsilon \geq 500$. Across all three datasets and all four privacy budgets, k=16 remains within 3 F_1 points of the best k, supporting its use as a robust default. PCA's gain over No-PCA is largest at low ε on every dataset (7–14 F_1 points at ε=100) and narrows at high ε, which is exactly the pattern predicted by the $\sigma\sqrt{d/k}$ noise reduction.

E Cross-Encoder Validation

The main paper's empirical analysis uses all-MiniLM-L6-v2 (d=384). To validate that PCA's noise reduction is not encoder-specific, we rerun the IMDB experiment with all-mpnet-base-v2 (d=768, a larger sibling). Table 13 compares PCA-16 against no PCA on both encoders.

PCA-16 wins on every cell. The relative improvement at ε=100 is substantial on both encoders (28.6% MiniLM, 17.5% mpnet). At ε=1000 the gap shrinks to within seed noise on both, consistent with the curse-of-dimensionality argument: noise reduction matters most where noise is large relative to signal. mpnet's stronger raw representation means its no-PCA baseline starts higher, leaving less absolute room for PCA gain, but the relative gain remains substantial.

	MiniLM (d=384)		mpnet (d=768)	
ε_{doc}	No PCA	PCA-16	No PCA	PCA-16
100	.497	**.639**	.647	**.760**
250	.573	**.679**	.728	**.793**
500	.605	**.702**	.727	**.746**
1000	.694	**.703**	.778	**.788**

Table 13: Cross-encoder validation on IMDB. PCA-16 wins on every (ε, encoder) cell. Relative F_1 improvement at ε=100 is 28.6% on MiniLM and 17.5% on mpnet. The $\sigma\sqrt{d/k}$ noise-reduction effect transfers across encoders.

F PCA Variance Analysis

Table 14 reports the cumulative variance captured by PCA components on IMDB alongside the corresponding F_1 at ε=500. The k=16 choice in the main text reflects this trade-off: enough dimensions to preserve task-relevant signal, few enough that noise remains manageable after perturbation.

G Sentence Selection

Table 15 reports F_1 as a function of the number of sanitized sentences N, supporting the discussion in Section 4.

k	Variance Explained	F_1 at ε=500
2	7.6%	0.550
4	11.8%	0.620
8	18.5%	0.688
16	28.1%	**0.702**
32	40.5%	0.657
64	55.8%	0.689
384	100%	0.605

Table 14: Variance explained by PCA components and corresponding F_1 at ε_{doc}=500. The optimal trade-off occurs at k=16: enough dimensions to preserve sentiment signal, few enough that noise remains manageable.

N	ε_s at ε_{doc}=100	F_1 (ε=100)	F_1 (ε=500)
1	100.0	**0.699**	0.701
2	50.0	0.653	0.710
3	33.3	0.674	0.700
5	20.0	0.607	0.706
10	10.0	0.624	**0.718**
all	var.	0.532	0.715

Table 15: Sentence selection within sentence-level sanitization. Selecting fewer sentences drastically improves F_1 at low ε by concentrating the privacy budget. "All" uses every sentence in the document ($L \approx 10$ on average, varying per document). PCA-16, plain NN retrieval.

H Per-Sentence vs. Document-Level Composition

Table 16 compares per-sentence ε (no composition accounting) with composed document-level ε to illustrate the cost of proper composition.

I Sanitization Examples

We show an example of sanitization on an IMDB review at ε_{doc}=500, comparing LoSanR (PCA-16 with context-aware retrieval) with DP-BART on the same input.

Original (negative review, 5 sentences): *"This is a worthless sequel to a great action movie. Cheap looking, and worst of all, BORING ACTION SCENES! The only decent thing about the movie is the last fight sequence. Only 82 minutes, but it feels like it goes on forever! Even die-hard Van Damme fans (like myself) should avoid this one!"*

LoSanR (5 sentences): *"Though, the film does have its plot holes. This may seem like a rather thin plot, but this noisy movie is so chock-full of singing and music that there is barely any dialogue at all. Even the overlong ballet at the end of the film makes it a better film with it than without. It moves so slowly that I defy you not to touch the fast forward button, especially on the two dance routines! Chris Jericho 6/10 It's funny looking back at Cena's very first PPV, how he used to act, how he used to dress, and how he used to look (watch his interview, it's pretty funny)."*

DP-BART (full output): *"The 'It is a good time to talk about Guillem Stahl, the' "*

Discussion. LoSanR's output is composed of real sentences from the public IMDB corpus. Most retrieved sentences mirror the original's negative judgment ("plot holes", "thin plot", "moves so slowly that I defy you not to touch the fast forward button"). A few go the other way (sentence 3 here defends an "overlong ballet"). The document-mean embedding, which drives downstream classification, averages across these per-sentence retrievals and preserves the dominant negative signal. The retrieved sentences also describe a different genre, a musical rather than an action film. This is a privacy strength. Topical content is hidden, and the sentiment signal that drives downstream classification carries through.

DP-BART, in contrast, produces a short hallucinated news-wire fragment unrelated to the input, consistent with the fluency comparison in App. C.

Method	Accounting	ε=100	F_1
No PCA	Per-sentence	ε_s=100	0.642
PCA-16	Per-sentence	ε_s=100	0.701
No PCA	Doc-level	ε_{doc}=100	0.497
PCA-16	Doc-level	ε_{doc}=100	0.639

Table 16: Impact of composition accounting. Per-sentence ε_s=100 gives each sentence the full budget; doc-level ε_{doc}=100 divides by L sentences ($\sim$10 on average). PCA partially compensates for the increased noise under proper composition.

J Common Evaluation Protocol with BERT

BERT fine-tuning setup. We use bert-base-uncased, fine-tuned with the standard sequence-classification head. Hyperparameters: learning rate 2×10^{-5}, batch size 32, 3 epochs, max sequence length 256, AdamW optimizer with linear warmup over the first 10% of steps, seed 42. We train on 12,500 sanitized documents (the full private half of our train split) and use the final-epoch

checkpoint. Test sets are the standard splits in unsanitized form (25,000 docs for IMDB and Amazon, 7,600 for AG News). As a no-DP reference, we also train BERT with identical hyperparameters on the same 12,500 documents in their original (unsanitized) form, achieving F_1: IMDB .913, Amazon .932, AG News .918.

K Implementation Details

Sentence splitting. Documents are split into sentences using a regex-based splitter that handles common abbreviations.

Clustering. We use MiniBatch K-Means (K=50, batch size 4096, seed 42) on the public pool sentence embeddings (after PCA projection and normalization). Cluster centroids are ℓ_2-normalized after fitting so that dot product equals cosine similarity. Cluster sizes range from ∼1,200 to ∼4,400 sentences.

Retrieval scoring. We use top-r=3 clusters and scoring weights w_{sim}=0.5, w_{prev}=0.2, w_{next}=0.2, w_{cl}=0.1. Downstream classification results are robust to reasonable variations of these parameters.

Analytic Gaussian calibration. We use the analytic calibration of Balle and Wang (2018) with numerical root-finding (`scipy.optimize.brentq`). For large $\varepsilon > 500$, we use arbitrary-precision arithmetic (`mpmath`) to avoid float64 overflow in the Gaussian CDF.

Computation. LoSanR's sanitization runs on CPU. Embedding uses the `sentence-transformers` library and takes 15–30 minutes per dataset. Subsequent experiments run from cache in minutes. BERT fine-tuning (App. J) uses a single GPU.

L Analytic Gaussian Mechanism

Theorem 2 (Analytic Gaussian Mechanism (Balle and Wang, 2018))**.** *Let $f : \mathcal{S} \to \mathbb{R}^d$ be a function with global ℓ_2 sensitivity Δ_2. For any $\varepsilon \geq 0$ and $\delta \in [0, 1]$, the Gaussian output perturbation mechanism $M(x) = f(x) + Z$ with $Z \sim \mathcal{N}(0, \sigma^2 I)$ is (ε, δ)-DP iff*

$$\Phi\left(\frac{\Delta_2}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta_2}{2\sigma} - \frac{\varepsilon\sigma}{\Delta_2}\right) \leq \delta$$

where Φ is the CDF of the standard univariate Gaussian.

M Derivation of the Cosine Similarity Approximation

We derive the approximation in Eq. 2. Let $\mathbf{x} \in \mathbb{S}^{d-1}$ be a unit-norm embedding and $\mathbf{y} = \mathbf{x} + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$. The cosine similarity is:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top\mathbf{y}}{\|\mathbf{y}\|} = \frac{1 + \mathbf{x}^\top\boldsymbol{\eta}}{\|\mathbf{x} + \boldsymbol{\eta}\|} \tag{4}$$

since $\|\mathbf{x}\| = 1$ and $\mathbf{x}^\top\mathbf{x} = 1$. For the numerator, $\mathbf{x}^\top\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2)$, so $\mathbb{E}[\mathbf{x}^\top\boldsymbol{\eta}] = 0$ and the numerator has expected value 1.

For the denominator:

$$\|\mathbf{y}\|^2 = 1 + 2\mathbf{x}^\top\boldsymbol{\eta} + \|\boldsymbol{\eta}\|^2 \tag{5}$$

The term $\|\boldsymbol{\eta}\|^2 \sim \sigma^2\chi^2_d$ has mean $\sigma^2 d$ and standard deviation $\sigma^2\sqrt{2d}$. The cross term $2\mathbf{x}^\top\boldsymbol{\eta}$ has standard deviation 2σ. For large d, the cross term is negligible relative to $\|\boldsymbol{\eta}\|^2$ (order σ vs. order $\sigma^2 d$), and $\|\boldsymbol{\eta}\|^2$ concentrates tightly around its mean. Therefore $\|\mathbf{y}\|^2 \approx 1 + \sigma^2 d$, giving:

$$\mathbb{E}[\cos(\mathbf{x}, \mathbf{y})] \approx \frac{1}{\sqrt{1 + \sigma^2 d}} \tag{6}$$

At d=384 and σ=0.30 (ε_s=50), this gives $1/\sqrt{1 + 0.09 \times 384} = 1/\sqrt{35.56} \approx 0.168$, matching the value reported in Section 2.3.

N Proof of Theorem 1

Proof. We prove the privacy step-by-step.
Encoding, PCA projection, and normalization. (Line 2) Obtaining the principal components of the public corpus does not leak any privacy as it does not touch the document d. (Line 5-10) These steps first project individual sentences onto the lower-dimensional space and normalize them (call this process P), then add noise to each projected sentence (call this process N). In other words, at the end of line 11, each sentence x is transformed into $N(P(x))$. N is the Analytic Gaussian mechanism, with parameters chosen so that for all S, y, y',

$$\Pr(N(y) \in S) \leq e^{\varepsilon_{\text{doc}}/L}\Pr(N(y') \in S) + \delta_{\text{doc}}/L.$$

We obtain this same inequality by replacing y with $P(x)$ and y' with $P(x')$ for any x, x'.
Post-processing. (Lines 11-14) These steps are post-processing on the noised sentences y_i, and thus do not affect the above inequality.

Sequential composition over L sentences (Dwork et al., 2006) gives $(L \cdot \varepsilon_{\text{doc}}/L, L \cdot \delta_{\text{doc}}/L) = (\varepsilon_{\text{doc}}, \delta_{\text{doc}})$-LDP over the whole document. □