d_χ-STENCIL: Context-Aware Privacy for Text Sanitization

Anonymous ACL submission

Abstract

Token-level metric differential privacy mechanisms sanitize text by perturbing each token's embedding independently, but existing methods such as CUSTEXT$^+$ and the d_χ-noise mechanism discard contextual relationships between neighboring tokens, degrading downstream utility beyond what the privacy noise alone would cause. We propose d_χ-STENCIL, which computes a Gaussian-weighted combination of neighboring token embeddings before adding calibrated Laplacian noise, preserving inter-token coherence while satisfying 2ϵ-d_χ-privacy. To enable meaningful privacy evaluation of continuous-noise mechanisms, we extend a Bayesian reconstruction attack via Monte Carlo estimation of transition probabilities, the first application of mechanism-aware attacks to this class of methods. Evaluation on IMDB and QNLI with RoBERTa-base shows that d_χ-STENCIL with even-numbered window size achieves lower embedding inversion rates than both CUSTEXT$^+$ and the noise mechanism at matched accuracy, while maintaining comparable Bayesian reconstruction rate.[1]

1 Introduction

Natural language processing (NLP) models as a service, such as ChatGPT (OpenAI, 2021; Ouyang et al., 2022), present notable privacy challenges. These models often require users to send their data to external servers, which can result in potential exposure, misuse, or insufficient protection of personal and sensitive information (Sousa and Kern, 2023). For example, information can be leaked by model inversion (Li et al., 2017; Devlin et al., 2019), exploitation of feature memorization within the model (Carlini et al., 2021), and more. This issue affects not only users but also service providers, who are required to adhere to regulations such as the General Data Protection Regulation (GDPR; Voigt and Von dem Bussche, 2017) and the California Consumer Privacy Act (CCPA; Barrett, 2019). Moreover, the dependence on third-party services further complicates compliance, as users must rely on these providers to ensure that they follow these privacy standards and manage data responsibly.

One common approach to addressing privacy concerns in NLP is the application of differential privacy mechanisms (Dwork, 2006), which offer formal guarantees by limiting the amount of information that can be inferred about any individual user. These mechanisms can be employed during either the training or inference phases of an LLM, typically by perturbing the input data with carefully calibrated noise. This process effectively conceals the original input, ensuring that private details cannot be reconstructed from the model's outputs. Differential privacy can operate in both remote and local settings. However, given the inherent trust issues with remote servers, a more practical and secure alternative is to apply differential privacy *locally* (LDP; Arachchige et al., 2019). In this setting, users preprocess their data using differential privacy techniques before sharing sanitized results with a server. This approach ensures that raw data remains protected from potential breaches, significantly reducing privacy risks while still enabling useful contributions to the model's functionality.

Applying standard LDP to text poses a fundamental challenge: since each user's entire text constitutes a single data point, text-level LDP requires that any two texts produce nearly indistinguishable outputs, which at $n = 1$ demands noise levels that destroy utility for any meaningful privacy parameter (Brown et al., 2022; Mattern et al., 2022). To address this, previous work has proposed *metric* (d_χ) relaxations of LDP (Feyisetan et al., 2020; Chen et al., 2023), which operate at the token level and require output distributions to be close only for inputs that are close in an embedding space. This

[1] Our code is available at: https://anonymous.4open.science/status/Dchi-Stencil-BE03/.

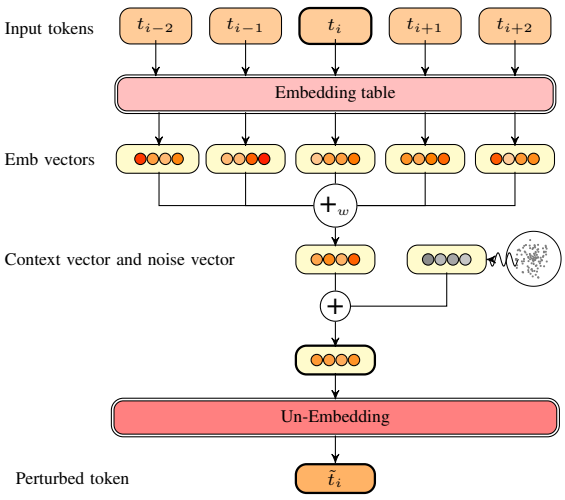

Figure 1: Schematic overview of our proposed method, d_χ-STENCIL, which integrates contextual and semantic information to enhance privacy preservation while maintaining the utility of the NLP models.

relaxation provides a weaker but practically useful privacy notion: rather than guaranteeing text-level indistinguishability, it bounds the information leaked about individual tokens proportionally to their semantic distance, offering a meaningful layer of protection against token reconstruction attacks (Igamberdiev and Habernal, 2023).

A key limitation of existing token-level metric DP methods (Feyisetan et al., 2020; Yue et al., 2021; Chen et al., 2023) is that they treat each token independently, discarding the contextual relationships between neighboring tokens. Since LLMs derive much of their power from contextual dependencies, sanitized text that preserves semantic similarity at the token level but destroys inter-token coherence suffers utility degradation beyond what the noise alone would cause.

In this work, we introduce d_χ-STENCIL, a novel privacy-preserving mechanism that integrates contextual and semantic information while providing formal privacy guarantees under the d_χ-privacy framework. For each token, d_χ-STENCIL computes a weighted combination of the embedding vectors of its neighbors using a Gaussian kernel, then adds calibrated Laplacian noise to the resulting vector before selecting the nearest token in the embedding space. We prove that this mechanism satisfies 2ϵ-d_χ-privacy for both Euclidean and cosine distance functions, with a tighter ϵ-d_χ-privacy bound holding for all tokens except those within L positions of the input boundaries. By preserving inter-token contextual coherence—an aspect discarded by prior metric DP methods—d_χ-STENCIL achieves a stronger utility–privacy tradeoff within the token-level metric DP setting.

We evaluate d_χ-STENCIL on QNLI (Wang et al., 2018) and IMDB (Maas et al., 2011) by fine-tuning RoBERTa-base (Liu et al., 2019) on clean data and evaluating on sanitized test sets. We compare against CUSTEXT$^+$ (Chen et al., 2023) and the d_χ-noise mechanism (NOISE; Feyisetan et al., 2020), assessing privacy via embedding inversion (EMI@20) and a Bayesian attack, extending the reconstruction attack of Tong et al. (2025) to continuous-noise mechanisms (such as d_χ-STENCIL and NOISE) via Monte Carlo estimation of transition probabilities. At matched accuracy levels, d_χ-STENCIL with even-numbered window size consistently achieves lower embedding inversion rates than both CUSTEXT$^+$ and NOISE, while maintaining comparable Bayesian rank-1 and exhibiting substantially lower rank-5 reconstruction match due to its larger effective output space. These results demonstrate that incorporating contextual information into token-level metric DP yields a stronger utility–privacy tradeoff than both exponential-mechanism and noise-only approaches.

2 Preserving Privacy in Models

There have been many attempts to preserve privacy which primarily focus on anonymization techniques (Liu et al., 2017; Friedrich et al., 2019) and methods that introduce noise into embedding vectors (Zhou et al., 2022, 2023). Although these approaches provide some degree of privacy protection, they often lack rigorous mechanisms to quantify the guarantees they offer. By contrast, the LDP mechanism offers a robust framework with formal guarantees, making it a reliable approach for safeguarding privacy in NLP applications (Cummings and Desai, 2018).

Since the primary goal is to maintain the usefulness of the NLP model, the privacy-preserving technique must achieve a good balance between privacy and utility. The standard LDP mechanism requires that any two samples produce indistinguishable output distributions (Feyisetan et al., 2020), which often results in outputs that lack sufficient information for downstream tasks. A common relaxation is d_χ-privacy (Yiwen et al., 2018; Feyisetan et al., 2020; Qu et al., 2021), which requires output

distributions to be close only in proportion to the distance between inputs, thereby preserving more semantic information.

Qu et al. (2021) proposed a d_χ-LDP mechanism that adds calibrated noise directly to input embeddings. While straightforward, the mechanism requires a high privacy parameter ϵ for effective protection, which may facilitate token reconstruction even by simple adversaries (Harel et al., 2024). SanText (Yue et al., 2021) takes an alternative approach, employing an exponential sampling mechanism (McSherry and Talwar, 2007) where replacement tokens are selected with probability proportional to $e^{-\epsilon \cdot \text{dist}/2}$. However, since replacements are sampled from the entire vocabulary, there remains a non-negligible probability of selecting semantically dissimilar tokens. CusText$^+$ (Chen et al., 2023) addresses this by first restricting the sampling space to the K most semantically similar tokens before applying the exponential mechanism, achieving superior utility at lower ϵ values.

All these methods treat each token independently, overlooking the potential benefits of incorporating contextual information—a critical component of LLM performance. Harel et al. (2024) introduced the Stencil mechanism, which encodes each token's embedding vector alongside its neighboring tokens' vectors, effectively preserving contextual nuances. However, the absence of a random component means Stencil does not satisfy any formal DP guarantee, leaving it vulnerable to deterministic reconstruction.

3 Context- and Semantic-based Privacy Preservation

One method commonly employed for ensuring local differential privacy involves injecting a controlled amount of *noise* into various model components, thereby obfuscating the original input. These components may include sequence embeddings, token embeddings, or the tokens themselves (Mosallanezhad et al., 2019; Feyisetan et al., 2020; Lyu et al., 2020; Qu et al., 2021; Zhou et al., 2022). The success of NLP models in most tasks, however, is primarily attributed to their ability to leverage contextual information. Thus, privacy-preserving mechanisms that incorporate contextual information have the potential to achieve better downstream task performance compared to those that do not consider context.

We introduce d_χ-Stencil, a privacy preservation technique based on the Stencil mechanism and the d_χ noise mechanism. Like Stencil, d_χ-Stencil encodes contextual information from neighboring tokens. Incorporating the d_χ noise component enhances privacy protection and provides formal guarantees under the d_χ-privacy framework.

3.1 The Stencil Mechanism

The Stencil (Harel et al., 2024) privacy mechanism operates by replacing each token $t_i, \forall i \in \{0 \dots N\}$ (where N is the number of tokens in the input) with a new token, $t_i \rightarrow \tilde{t}_i$, that aims to preserve both semantics and context. To derive the new token $\tilde{t}_i$, information from its L neighboring tokens $t_k, \forall k \in \{i - \frac{L}{2}, \dots, i + \frac{L}{2}\}$ is incorporated. This is achieved by obtaining the embedding vector representations, ϕ_k, of each of the $L+1$ tokens using an embedding lookup table, $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times ||\phi||}$, where $|\mathcal{V}|$ is the size of the vocabulary and $||\phi||$ is the dimension of the vector. Next, the embedding vectors are combined using a normalized weighted function f_k (satisfying $\sum_{k=i-\frac{L}{2}}^{i+\frac{L}{2}} f_k = 1$) to generate a new embedded quasi-vector $\tilde{\phi}_i$. The new perturbed token $\tilde{t}_i$ is then selected as the token from the vocabulary $\mathcal{V}$ with the smallest distance to $\tilde{\phi}_i$ that is different than the original one, i.e., $\arg\min_{t_j \in \mathcal{V} \setminus \{t_i\}} \text{dist}\left(\mathbf{E}[t_j], \tilde{\phi}_i\right)$, where dist is a distance scoring function.

Although the Stencil mechanism obscures the original token and offers a certain level of privacy, it does not adhere to the principles of differential privacy due to the absence of a randomized component. This deterministic nature not only limits its compliance with formal privacy guarantees, but also renders it highly vulnerable to reconstruction attacks, as adversaries can exploit the lack of randomness to reverse-engineer the original data.

3.2 The d_χ Mechanism

The relaxed d_χ mechanism, or the Noise mechanism, proposed by Feyisetan et al. (2020) and Qu et al. (2021), requires that for any two inputs $x, x' \in X$ with a randomized function $M : X \rightarrow X$, the following holds:

$$\frac{Pr[M(x) = y]}{Pr[M(x') = y]} = e^{\eta \, \text{dist}\,(x, x')}, y \in X, \quad (1)$$

where $\text{dist}\,(x, x')$ is a distance function, such as cosine distance or euclidean distance, and η is a privacy parameter. This condition guarantees that

Algorithm 1 The d_χ-STENCIL mechanism.

Input: Embeddings table $\mathbf{E}$, N tokens, σ, L
for $i = 0 \ldots N$ **do**
  $\tilde{\phi}_i \leftarrow 0$
  for $j = \min\left(0, i - \frac{L}{2}\right) \ldots \max\left(i + \frac{L}{2}, N\right)$ **do**
    $\phi_j \leftarrow \mathbf{E}[t_j]$
    $f(j, \sigma) \leftarrow e^{-j^2/2\sigma^2}$
    $\tilde{\phi}_i \leftarrow \tilde{\phi}_i + f(j, \sigma) \cdot \phi_j$
  end for
  $\tilde{\phi}_i \leftarrow \tilde{\phi}_i / \sum_j f(j, \sigma) + \mathbf{p}$
  $\tilde{t}_i \leftarrow \arg\min_{t_j \in \mathcal{V} \setminus \{t_i\}} \text{dist}\left(\mathbf{E}[t_j], \tilde{\phi}_i\right)$
end for

the probability distributions over outputs for similar inputs are close. For instance, if the distance function is defined over the embedding space, smaller distances can correlate to semantic similarity.

To apply the d_χ mechanism in LLMs, Qu et al. (2021) proposed that the randomized function be defined as $M(x) = x + \mathbf{p}$, where $x \in \mathbb{R}^{||\phi||}$ represents the sequence embedding vector representation, and $\mathbf{p} \in \mathbb{R}^{||\phi||}$ is a noise vector sampled from a distribution whose density is proportionate to η. Specifically, we generate $\mathbf{p}$ by multiplying a sample from a Gamma distribution $\Gamma(||\phi||, 1/\eta) = \frac{r^{||\phi||-1}\eta^{||\phi||}e^{-r\eta}}{(||\phi||-1)!}$ and a uniform sample from a unit hypersphere $\frac{\Gamma(||\phi||/2)}{||\phi||\pi^{||\phi||/2}}$.

For the token-level implementation of this mechanism, we introduce pre-processing and post-processing stages. The pre-processing step involves looking up the embedding vector representation of the token t in an embedding table $\mathbf{E}[t]$. Then, the d_χ mechanism is applied on $\mathbf{E}[t]$ as described. Finally, the post-processing step returns the token t' that is closest (according to the distance function d) to the output of the d_χ mechanism, effectively yielding a noised but semantically similar token.

3.3 The d_χ-STENCIL Mechanism

Our proposed method, d_χ-STENCIL, integrates the d_χ mechanism with the STENCIL approach, thereby maintaining privacy guarantees while leveraging the contextual and semantic information of the text to better preserve utility. The d_χ-STENCIL mechanism follows the same procedure as the STENCIL mechanism, with one key difference. After computing the quasi-embedding vector $\tilde{\phi}_i$, a calibrated noise $\mathbf{p}$, derived as explained in §3.2, is added. The process is described in Algorithm 1.

The privacy proof is supplied in Appendix A.

The privacy-utility trade-off of the d_χ-STENCIL method can be controlled by adjusting the window size L, the properties of the weighted function f, and η. In our study, we employ a Gaussian smoothing function as the weighting function, where the standard deviation σ plays a pivotal role in balancing privacy and downstream task performance. For odd-numbered window size, the Gaussian weights concentrate on the original token, which may enhance accuracy but also lower privacy. In contrast, an even-numbered window size, the highest weights are distributed between the original token and an adjacent token, resulting in a more balanced allocation.

In section 5, we analyze the impact of these parameters on accuracy and privacy. We note that L and σ do not appear in the formal privacy bound: by Appendix A, the mechanism is 2ϵ-d_χ-private regardless of these choices. Their effect is purely on utility and on the empirical difficulty of reconstruction, not on the worst-case privacy guarantee.

4 Experiments

We evaluate the utility and privacy of d_χ-STENCIL against CUSTEXT$^+$ and NOISE on two NLP tasks: natural language inference (QNLI; Wang et al., 2018)) and sentiment analysis (IMDB; Maas et al., 2011)). For each task, we fine-tune RoBERTa-base (Liu et al., 2019) on the clean training set and evaluate on the sanitized test set. All methods operate over the GloVe 840B-300d embedding space (Pennington et al., 2014). For CUSTEXT$^+$ we evaluate privacy budget of $\epsilon \in \{1, 3, 5, 7, 8\}$ and NOISE over $\eta \in \{80, 85, 90, 95\}$. For d_χ-STENCIL we evaluate with window size $L = 4, \sigma = 0.75$ over $\eta = \{140, 150, 160, 170, 180, 200, 240\}$, and with window size $L = 5, \sigma = 0.75$ over $\eta \in \{130, 140, 150, 160, 170\}$. We use GloVe rather than model-native embeddings because the LDP setting requires sanitization to occur *before* the text reaches the model; a static, publicly available embedding table allows the user to perform sanitization locally without access to the service provider's model weights. We report results for $\sigma = 0.75$, which consistently achieved the best utility–privacy tradeoff; a detailed parameter study is provided in section 5. All experiments were conducted on a server with two AMD EPYC 7763 64-Core processors and an NVIDIA RTX 6000.

Stopword exclusion. Our mechanism treats all content words as sensitive, since reliably distinguishing sensitive from non-sensitive tokens is it-

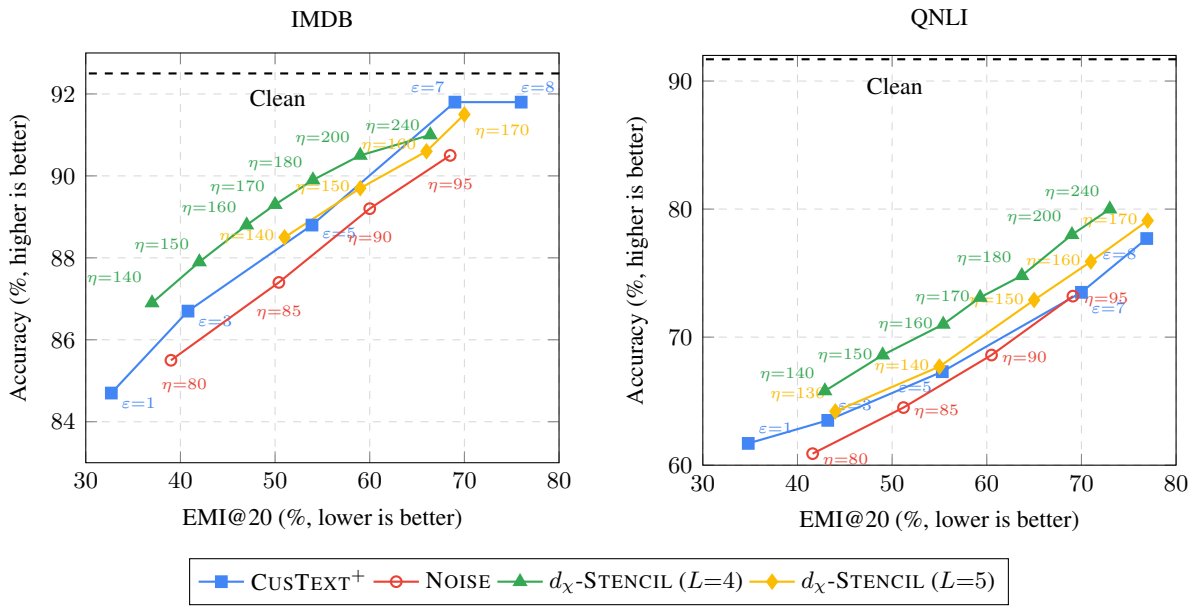

Figure 2: Accuracy versus embedding inversion (EMI@20) trade-offs for CusText$^+$, d_χ-Stencil with $L \in \{4, 5\}$ and $\sigma = 0.75$, and Noise across different privacy budgets (ϵ and η), evaluated on IMDB (left) and QNLI (right).

self a challenging task (Liu et al., 2017; Cohn et al., 2019; Poostchi et al., 2018; Friedrich et al., 2019). Following prior work (Chen et al., 2023) we do not sanitize stopwords, as they pose a low privacy risk.

Out-of-vocabulary tokens. During the experiments we observed that some tokens were not included in the GloVe vocabulary. For these OOV tokens, their original forms were retained, which improved model accuracy but also harmed privacy. The OOV rate across all datasets was below 5%, and all privacy-preserving techniques were similarly affected, ensuring fair comparisons.

Reconstruction attacks. We evaluate privacy using two reconstruction attacks of increasing sophistication. The first is embedding inversion (EMI@20; Qu et al., 2021): given a sanitized token $\tilde{t}$, the attacker extracts its embedding $\mathbf{E}[\tilde{t}]$, ranks all vocabulary tokens by cosine similarity, and selects the top 20 candidates. We report the fraction of content words for which the original token appears among these twenty candidates. The second is a Bayesian attack, extending the context-free reconstruction attack of Tong et al. (2025) to continuous-noise mechanisms. Given $\tilde{t}$, the attacker recovers the original token by computing $\hat{t} = \arg\max_t P(\tilde{t} \mid t) \cdot P(t)$, where $P(t)$ is a token frequency prior estimated from a shadow corpus. For exponential-mechanism methods such as CusText$^+$, the transition probabilities $P(\tilde{t} \mid t)$ are available in closed form. For continuous-noise mechanisms, computing $P(\tilde{t} \mid t)$ requires integrating the noise density over the region of $\tilde{t}$ in embedding space, which is intractable in high dimensions. We address this by estimating $P(\tilde{t} \mid t)$ via Monte Carlo sampling: for each token t, we repeatedly apply the sanitization mechanism and record the empirical frequency of each output. For Noise, the transition probabilities are context-free and can be precomputed globally ($N = 10{,}000$ samples per token). For d_χ-Stencil, the quasi-embedding depends on neighboring tokens, so we aggregate across corpus occurrences of t to obtain a context-averaged estimate ($N = 10{,}000$ samples per occurrence). To the best of our knowledge, this is the first application of Bayesian reconstruction attacks to continuous-noise token-level DP mechanisms. We report the attack success rate at rank 1 (Bayes@1) and rank 5 (Bayes@5), defined as the fraction of content words for which the original token appears as the top-ranked candidate or among the top five candidates, respectively. The Monte Carlo estimation of transition probabilities for continuous-noise mechanisms is computationally intensive (approximately 12 hours per configuration), compared to the closed-form computation available for CusText$^+$.

4.1 Embedding Inversion

Figure 2 presents the accuracy–privacy trade-off under embedding inversion (EMI@20) on IMDB and QNLI for various privacy budgets.

Mechanism	ϵ / η	IMDB				QNLI			
		Acc (↑)	EMI (↓)	Bayes@1 (↓)	Bayes@5 (↓)	Acc (↑)	EMI (↓)	Bayes@1 (↓)	Bayes@5 (↓)
CusText$^+$	3.0	86.7	40.8	46.2	78.4	63.5	43.2	33.5	70.1
	5.0	88.8	53.9	52.5	82.4	67.3	55.3	42.1	75.3
Noise	80	85.5	39.0	71.2	77.5	60.9	41.6	74.9	79.9
	85	87.4	50.4	76.6	81.4	64.5	51.2	79.1	84.3
d_χ-Stencil($L=4$)	140	86.9	37.2	48.5	55.1	65.8	42.9	37.2	42.5
	150	87.9	42.9	50.9	58.4	68.6	49.2	42.1	47.5
d_χ-Stencil($L=5$)	130	87.1	41.1	52.7	61.4	64.2	44.5	53.4	59.5
	140	88.5	51.9	60.2	69.4	67.6	55.3	61.7	67.3

Table 1: Accuracy and privacy evaluation of CusText$^+$, Noise and d_χ-Stencil on IMDB and QNLI with EMI@20 and Bayes reconstruction.

d_χ-Stencil consistently achieves better utility and privacy than Noise, supporting the hypothesis that incorporating contextual information preserves utility while improving privacy. Moreover, across both datasets, d_χ-Stencil with even-numbered window size (L=4, σ=0.75) achieves the strongest trade-off: at matched accuracy levels, it yields lower EMI@20 scores than both CusText$^+$ and Noise. For example, on IMDB at approximately 88.8% accuracy, d_χ-Stencil (L=4) achieves an EMI@20 score of around 47%, notably lower than both baselines at comparable accuracy. This advantage stems from the even-window Gaussian weighting, which distributes the highest weights between the original token and its adjacent token rather than concentrating on the original alone, reducing direct leakage of the original embedding while preserving local semantic coherence.

At high values of the privacy parameter (large η or ε), where noise is negligible, Noise, CusText$^+$, and the odd-window d_χ-Stencil converge to the unsanitized ("clean") accuracy, as the sanitization process becomes highly likely to preserve the original token. The even-window configuration, however, does not converge to that accuracy because the Gaussian weighting ensures that, even with negligible noise, the quasi-embedding remains a blend of the original token and its adjacent, producing a different output token. This results in an inherent accuracy ceiling—a trade-off for its superior privacy at lower accuracy levels.

The odd-window configuration (L=5) performs comparably to CusText$^+$ on both datasets. On QNLI, the even-window d_χ-Stencil (L=4) substantially outperforms CusText$^+$ at matched EMI@20 levels, reflecting the benefit of contextual aggregation on a task that depends heavily on inter-token relationships. Noise exhibits the weakest trade-off across all settings, as it perturbs each token independently without preserving contextual information, leading to rapid accuracy degradation as noise increases.

4.2 Bayesian Attack

Table 1 presents reconstruction rates under the Bayesian attack alongside embedding inversion and accuracy, at operating points selected for comparable accuracy across mechanisms.

One key observation is the substantial gap between embedding inversion and Bayesian reconstruction for continuous-noise mechanisms. For Noise, Bayes@1 exceeds EMI@20 across both datasets, while CusText$^+$ shows a much smaller gap between the two attacks. This disparity arises because continuous-noise mechanisms produce peaked transition distributions: a token perturbed by small noise almost always maps to one of a few nearest neighbors, making the reverse mapping nearly deterministic for a mechanism-aware attacker. CusText$^+$'s exponential sampling over K=20 candidates distributes probability more uniformly, leaving greater uncertainty even with full knowledge of the mechanism. These results demonstrate that embedding inversion substantially underestimates the privacy risk of continuous-noise mechanisms, and that mechanism-aware attacks are essential for meaningful privacy evaluation.

Mirroring the EMI@20 findings, Noise exhibits the weakest accuracy–privacy trade-off across all settings on Bayes@1 as well. It underperforms d_χ-Stencil for the same reason as before, namely that contextual aggregation simultaneously improves utility and privacy, and underperforms CusText$^+$ due to the peaked transition distribu-

tions described above. Among d_χ-STENCIL configurations, the even-window variant (L=4) again outperforms the odd-window variant (L=5), reinforcing the embedding inversion results and confirming that the even-window Gaussian weighting provides privacy benefits that extend beyond geometric embedding distance to mechanism-aware reconstruction.

In contrast to the EMI@20 results, however, CUSTEXT$^+$ is more competitive on Bayes@1. It outperforms the odd-window d_χ-STENCIL on both datasets and achieves comparable results to the even-window variant, indicating that contextual mixing can match the Bayesian resistance of the exponential mechanism on Bayes@1.

Examining Bayes@5 reveals a structural difference between the mechanisms, and a reversal: d_χ-STENCIL substantially outperforms CUSTEXT$^+$ (as well as NOISE). CUSTEXT$^+$ exhibits a large gap between Bayes@1 and Bayes@5. For example, on QNLI at ε=3, Bayes@1 is 33.5% but Bayes@5 rises to 70.1%. This reflects CUSTEXT$^+$'s restricted candidate pool of K=20 tokens: the attacker cannot identify the exact original token, but can narrow it to a small set with high confidence. d_χ-STENCIL (L=4) shows a much smaller gap between Bayes@1 and Bayes@5 across both datasets. The larger effective output space of continuous-noise mechanisms means that expanding the candidate set to five does not substantially improve the attacker's success, offering a qualitatively different privacy profile than exponential-mechanism methods.

5 Parameter Study

To investigate the impact of window size L and the standard deviation σ of the Gaussian weights f_i on the accuracy and resilience against reconstruction attacks, we conducted tuning experiments for these parameters.

5.1 Impact of σ

The standard deviation σ of the Gaussian weighting function controls how much contextual information is incorporated into the quasi-embedding. Lower σ concentrates weight on the original token, making d_χ-STENCIL behave more similarly to NOISE; higher σ distributes weight across neighboring tokens, increasing context mixing at the cost of shifting the quasi-embedding further from the original.

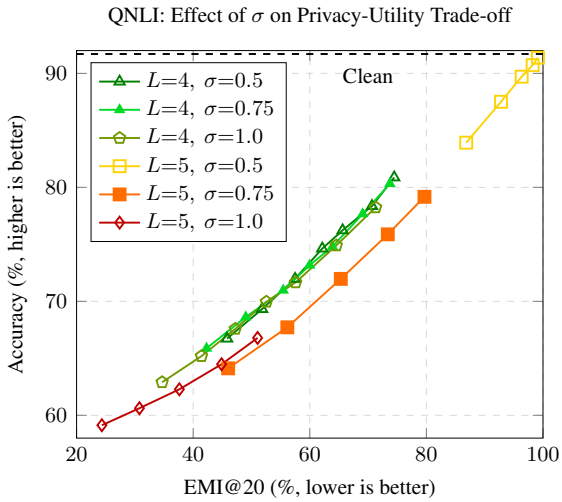

Figure 3: Effect of standard deviation σ on accuracy versus embedding inversion (EMI@20) trade-offs for d_χ-STENCIL with $L \in \{4, 9\}$ and $\sigma \in \{0.5, 0.75, 1.0\}$, evaluated on QNLI.

Figure 3 presents accuracy versus EMI@20 on QNLI for $\sigma \in \{0.5, 0.75, 1.0\}$ with both L=4 and L=5. For L=5, the effect of σ is particularly pronounced. At σ=0.5, the Gaussian weight at offset zero dominates, and the quasi-embedding remains nearly identical to the original token embedding; consequently, EMI@20 exceeds 86% even at the lowest η values tested, offering minimal privacy despite high accuracy. At σ=1.0, context mixing is aggressive, reducing EMI@20 below 51% but also degrading accuracy substantially. The intermediate value σ=0.75 achieves the best utility–privacy tradeoff, balancing sufficient context mixing for privacy with enough original-token signal for downstream accuracy.

For L=4, the three σ values produce more closely spaced curves, since the even-window weighting already distributes weight away from the original token. Although σ=1.0 achieves lower EMI@20 at matched accuracy, it limits the achievable accuracy and requires substantially larger η values, weakening the formal privacy guarantee. We therefore fix σ=0.75 for all main experiments.

5.2 Impact of Window Size L

Figure 4 presents the effect of window size $L \in \{4, 5, 6, 7, 8, 9\}$ with fixed σ=0.75 on QNLI. The most prominent pattern is the separation between even- and odd-numbered window sizes, which dominates the effect of window size itself.

For odd-numbered L (5, 7, 9), the tradeoff curves are nearly indistinguishable: increasing the win-

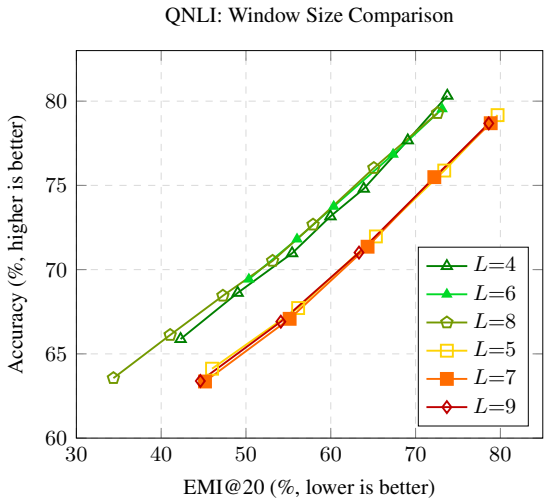

Figure 4: Effect of window size $L \in \{4, 5, 6, 7, 8, 9\}$ with $\sigma = 0.75$ on the utility–privacy trade-off on QNLI.

dow size has minimal impact on either accuracy or EMI@20. This is because odd windows place the highest Gaussian weight on the original token at offset zero, and the additional neighbors contribute diminishing weight as L grows. The quasi-embedding remains anchored to the original token regardless of how many neighbors are included.

For even-numbered L $(4, 6, 8)$, window size has a more visible effect. Larger even windows incorporate more context, shifting the quasi-embedding further from the original token and reducing EMI@20 at matched η values. However, this comes at the cost of lower accuracy: $L{=}8$ achieves the lowest reconstruction rates but also the lowest accuracy, while $L{=}4$ reaches the highest accuracy among even configurations. The pattern mirrors the effect of increasing σ—both widen the effective context contribution, with similar consequences for the utility–privacy tradeoff.

Based on these results, we select $L{=}4$ for the even-window configuration and $L{=}5$ for the odd-window configuration in our main experiments, as these provide the widest useful accuracy range while demonstrating the characteristic behavior of each window parity.

6 Conclusion

We presented two contributions to token-level metric DP for text sanitization. First, we proposed d_χ-STENCIL, which integrates contextual information from neighboring tokens via Gaussian-weighted embedding aggregation before adding calibrated Laplacian noise, satisfying 2ε-d_χ-privacy. The even-window configuration achieves a stronger utility–privacy tradeoff than both CUSTEXT$^+$ and the d_χ-noise mechanism under embedding inversion, while maintaining comparable Bayesian reconstruction resistance. The benefit of contextual aggregation is task-dependent: on QNLI, where inter-token relationships drive task performance, both d_χ-STENCIL variants outperform CUSTEXT$^+$, whereas on IMDB the odd-window configuration performs comparably to CUSTEXT$^+$. The even-window configuration introduces an inherent accuracy ceiling due to its weight distribution away from the original token, a controlled cost of its privacy advantage.

Second, we extended the Bayesian reconstruction attack of Tong et al. (2025) to continuous-noise mechanisms via Monte Carlo estimation of transition probabilities, the first mechanism-aware attack applicable to this class of methods. This revealed that embedding inversion substantially underestimates the privacy risk of continuous-noise mechanisms, with the Bayesian attack recovering tokens at rates far exceeding nearest-neighbor analysis. Examining reconstruction at both rank 1 and rank 5 further exposed a structural difference: CUSTEXT$^+$'s restricted candidate pool yields low Bayes@1 but high Bayes@5, while d_χ-STENCIL's larger output space produces a much smaller gap, reflecting qualitatively different privacy profiles. These findings underscore the need for mechanism-aware, multi-criteria privacy evaluation in token-level differential privacy.

Future work could explore richer contextual aggregation schemes, such as incorporating syntactic structure rather than fixed-window proximity, and develop tighter privacy analyses that account for the context-dependent nature of d_χ-STENCIL's transition probabilities.

Limitations

The d_χ-privacy guarantee operates at the token level: it bounds the information leaked about individual tokens proportionally to their semantic distance, but does not provide text-level indistinguishability. An adversary observing the full sanitized sequence may exploit global patterns—such as syntax, topic coherence, or stylometric features—that token-level perturbation does not obscure. This is a fundamental limitation shared by all token-level metric DP methods (Feyisetan et al., 2020; Yue et al., 2021; Chen et al., 2023), and closing

the gap between token-level and text-level privacy remains an open problem (Brown et al., 2022).

Moreover, the quasi-embedding $\tilde{\phi}_i$ is computed as a Gaussian-weighted average of static embedding vectors within a fixed window, capturing only local co-occurrence patterns. This contrasts with the deep contextualized representations produced by transformer-based models, where each token's representation reflects attention over the entire input and multiple layers of nonlinear transformation. As a result, d_χ-STENCIL cannot capture long-range dependencies, syntactic structure, or contextual disambiguation of polysemous words.

In addition, our experiments focus on two classification benchmarks, in English only, and on an encoder-only model. Evaluating d_χ-STENCIL on generative tasks such as open-ended question answering or summarization would provide a more complete picture of its practical utility, and is left for future work.

References

Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. 2019. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7):5827–5842.

Catherine Barrett. 2019. Are the eu gdpr and the california ccpa becoming the de facto global standards for data privacy and protection? *Scitech Lawyer*, 15(3):24–29.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2280–2292.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

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, Toronto, Canada. Association for Computational Linguistics.

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio de-identification - a new entity recognition task. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 197–204, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachel Cummings and Deven Desai. 2018. The role of differential privacy in gdpr compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, volume 20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Principles of Security and Trust: 8th International Conference, POST 2019, Part of ETAPS 2019, Proceedings 8*, pages 123–148.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5829–5839, Florence, Italy. Association for Computational Linguistics.

Re'em Harel, Yair Elboher, and Yuval Pinter. 2024. Protecting privacy in classifiers by token manipulation. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 29–38, Bangkok, Thailand. 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.

Yang Li, Sandro Schulze, and Gunter Saake. 2017. Reverse engineering variability from natural language documents: A systematic literature review. In *Proceedings of the 21st International Systems and Software Product Line Conference-Volume A*, pages 133–142.

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*.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75:S34–S42.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365, Online. Association for Computational Linguistics.

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: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2360–2369, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2021. Chatgpt. OpenAI Website. Accessed on 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018. BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1488–1497.

Samuel Sousa and Roman Kern. 2023. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*, 56(2):1427–1492.

Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu, and Jie Zhang. 2025. On the vulnerability of text sanitization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5150–5164, Albuquerque, New Mexico. Association for Computational Linguistics.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

NIE Yiwen, Wei Yang, Liusheng Huang, Xike Xie, Zhenhua Zhao, and Shaowei Wang. 2018. A utility-optimized framework for personalized private histogram estimation. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):655–669.

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, Online. Association for Computational Linguistics.

Xin Zhou, Jinzhu Lu, Tao Gui, Ruotian Ma, Zichu Fei, Yuran Wang, Yong Ding, Yibo Cheung, Qi Zhang, and Xuanjing Huang. 2022. TextFusion: Privacy-preserving pre-trained model inference via token fusion. In *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pages 8360–8371, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuanjing Huang. 2023. TextObfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5459–5473, Toronto, Canada. Association for Computational Linguistics.

A Privacy Proof

A.1 Differential Privacy of d_χ-STENCIL

In this section, we prove that d_χ-STENCIL is d_χ-private, with the main result stated in Theorem 1. As preliminary steps preceding the theorem, we define the distance functions for which the theorem applies, state the noise distribution $\mathbf{p_i}$ in more convenient form and prove Lemma 1, which is a bound on the overall contribution of a specific token to the output.

We prove Theorem 1 for two distance functions defined as follows. Consider two vectors $x = (x_1, \ldots, x_N), (x'_1, \ldots, x'_N) \in \left(\mathbb{R}^\ell\right)^N$ and define the first distance function as the sum of the Euclidean distances between the components, i.e.,

$$d_2(x, x') = \sum_{i=1}^{N} \left|\left|x_i - x'_i\right|\right|,$$

where $||z_i|| = \sqrt{z_{i,1}^2 + \ldots + z_{i,\ell}^2}$ is the Euclidean norm for any $z_i = (z_{i,1}, \ldots, z_{i,\ell}) \in \mathbb{R}^\ell$. The second distance function is defined as the sum of the cosine distances between the components, i.e.,

$$d_C(x, x') = \sum_{i=1}^{N} 1 - \frac{\langle x_i, x'_i \rangle}{||x_i|| \cdot ||x'_i||},$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^\ell$ and $\frac{\langle x_i, x'_i \rangle}{||x_i|| \cdot ||x'_i||}$ is cosine similarity.

We define the perturbation $\mathbf{p_i}$ in the algorithm using the Laplacian density function in ℓ dimensions from (Fernandes et al., 2019). The function with parameter $\epsilon > 0$, is defined for every vector $v = (v_1, \ldots, v_\ell) \in \mathbb{R}^\ell$ by $\mathrm{Lap}_{\ell,\epsilon}(v) = ce^{-\epsilon||v||}$, where $||v||$ is the Euclidean norm and $c = c(\ell, \epsilon)$ is chosen so that the associated cumulative distribution function is correct. That is, it is chosen so that $\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} ce^{-\epsilon||v||} dv_1 \ldots dv_\ell = 1$.

(Fernandes et al., 2019) prove that the noise sampled in Algorithm 1 as a product of a gamma distribution and a uniform distribution on specific domains is distributed with the Laplacian density function in ℓ dimensions if $\mathbf{E}(t_i) \in \mathbb{R}^\ell$ for every token t. In the d_χ-STENCIL mechanism, each token t_i is associated with $\tilde{\phi}_i$, a weighted sum of the embeddings of all tokens t_j within a window of at most L tokens around t_i. Denote the weights used to compute $\tilde{\phi}_i$ by $f'_{i,j}$ before normalization and by $f_{i,j} = f'_{i,j} / \sum_j f'_{i,j}$ after normalization, which implies that $\sum_j f_{i,j} = 1$. The specific

weights $f'_{i,j}$ used in the algorithm have two important properties: *symmetry* around i, meaning that $f'_{i,i-j} = f'_{i,i+j}$ for all i, j, and *decay* with distance from i, i.e. $f'_{i,j'} \leq f'_{i,j}$ for all i, and all j, j' such that $|i-j| \leq |i-j'|$.

Now, instead of considering the contribution of all neighboring tokens t_j to $\tilde{\phi}_i$, we estimate the total contribution of a token t_j to the $\tilde{\phi}_i$ of all its neighbors. Let $C_j = \sum_{i=1}^{N} f_{i,j}$ be the *contribution* of a token t_j. The following lemma bounds C_j.

Lemma 1. *For any non-negative integers N, L the d_χ-*STENCIL *mechanism satisfies the two following properties:*

- $C_j = 1$ *for all j such that $L \leq j \leq N - L$.*
- $C_j \leq 2$ *for all j such that $0 \leq j \leq N$.*

Proof. If $L \leq j \leq N - L$, then t_j contributes to values $\tilde{\phi}_i$ in the range $L/2 \leq j - L/2 \leq i \leq j + L/2 \leq N - L/2$. The window for each of the associated tokens t_i is of length L, and therefore the sequence of weights $(f_{i,i-L/2}, \ldots, f_{i,i+L/2})$ is identical for all these tokens, including t_j. It follows that due to the symmetry property noted above, the contribution of t_i to $\tilde{\phi}_j$ is equal to the contribution of t_j to $\tilde{\phi}_i$, or formally $f_{i,j} = f_{j,i}$. Therefore, $C_j = \sum_{i=0}^{N} f_{i,j} = \sum_{i=0}^{N} f_{j,i} = 1$, which proves the first part of the lemma.

For general j, $0 \leq j \leq N$, let L' (respectively L'') be the number of tokens to the left (resp. right) of t_j that contribute to $\tilde{\phi}_j$, and assume w.l.o.g. that $L' \leq L'' \leq L/2$. Let $F_i = \sum_j f_{i,j}$ be the normalization factor for the i-th token, that is, $f_{i,j} = f'_{i,j}/F_i$ for all i, j. By the symmetry of the weights we have that $f'_{i,j} = f'_{j,i}$. Note that symmetry implies that at least half the weight F_j is concentrated in $f'_{j,j}$ and to its right, that is $f'_{j,j} + \sum_{i=1}^{L''} f'_{j+i,j} \geq F_j/2$. To prove the second part of the lemma's statement, we show that $F_i \geq F_j/2$ for all $j' - L' \leq i \leq j' + L''$.

We start with $t_{j-L'}$, which is the leftmost token that receives contribution from t_j. That token receives contribution from at least L'' tokens to the right of it, since it is not to the right of t_j. Therefore, $F_{j'-L'} \geq f'_{j,j} + \sum_{i=1}^{L''} f'_{j+i,j} \geq F_j/2$. We now move right from $t_{j-L'}$ maintaining for each token a set of $L'' + 1$ tokens that contribute to it. For t_i in the range $j - L' \leq i \leq j - L' + L''$ that set is exactly $\{t_{j-L'}, \ldots, t_{j-L'+L''}\}$, while in the range $j - L' + L'' < i \leq j + L''$ that set is $\{t_{i-L''}, \ldots, t_i\}$. In the second range, due to symmetry, $F_i = F_{j-L'} \geq F_j/2$. In the first range, due to both symmetry and decay with distance from the token, we have that $F_i \geq F_{j-L''} \geq F_j/2$.

Therefore, $C_j = \sum_{i=0}^{N} f_{i,j} = \sum_{i=0}^{N} f'_{i,j}/F_i \leq \sum_{i=0}^{N} 2f'_{j,i}/F_j = 2$, which completes the proof of the lemma.

□

The following theorem states the differential privacy property that the d_χ-STENCIL mechanism satisfies. Its proof depends on the Laplace distribution of the noise.

Theorem 1. *For all non-negative integers ℓ, L, N, for all embedding functions $\mathbf{E}$ mapping tokens to $\mathbb{R}^\ell$, and for every $\epsilon > 0$, the d_χ-*STENCIL *mechanism with parameters $L, N, \mathbf{E}$ and Laplace noise with parameter ϵ is 2ϵ-d_χ-private for both the d_2 and the d_C distance functions.*

Proof. Let $t = (t_1, \ldots, t_N), t' = (t'_1, \ldots, t'_N)$ be two sequences of N input tokens and let $x = \mathbf{E}(t) = (x_1, \ldots, x_N), x' = \mathbf{E}(t') = (x'_1, \ldots, x'_N)$ be the associated sequences of embeddings in $\mathbb{R}^\ell$. Let $M(x)$ denote the output of d_χ-STENCIL on a sequence of embeddings x. If $y = M(x)$, then $y = (y_1, \ldots, y_N)$ is a sequence of N output embeddings.

The proof proceeds in three steps. The first step assumes that d_χ-STENCIL uses Euclidean distance and gives a bound on the probability that the i-th output component is y_i as a function of the weighted distance between the components of x and x' around i. The second uses the independence of the $\mathbf{p_i}$ noise components to prove that d_χ-STENCIL is 2ϵ-d_χ-private for Euclidean distance between the embeddings. The third step proves the theorem for cosine distance.

For every $1 \leq i \leq N$, let $A_{y_i} \subseteq \mathbb{R}^\ell$ be the subset of all vectors that are closer to y_i than to z for any possible output z. d_χ-STENCIL with input t returns output y if and only if for every $i = 1, \ldots, N$, the weighted sum of the embeddings in x together with the i-th noise component $\mathbf{p_i}$ is in A_{y_i}. Let $r_i + \mathbf{p_i} = \sum_j f_{i,j} x_j + \mathbf{p_i}$ be the random variable which measures the probability that on input t the mechanism outputs y_i in its i-th component. Similarly $r'_i + \mathbf{p_i} = \sum_j f_{i,j} x'_j + \mathbf{p_i}$ measures the same probability when the input is t'. The probability that $r_i + \mathbf{p_i} \in A_{y_i}$ is

$$\Pr[r_i \mathbf{p_i} \in A_{y_i}] = \int_{A_{y_i}} c e^{-\epsilon ||z - r_i||} dz.$$

By the reverse triangle inequality of Euclidean norm (the difference between the norms of two vectors is at least the norm of the difference between the vectors):

$$\begin{aligned}-\epsilon\,||z-r_i|| &= -\epsilon\,||z-r_i|| + \epsilon\,\left|\left|z-r'_i\right|\right| - \\ &\quad - \epsilon\,\left|\left|z-r'_i\right|\right| \\ &\leq \epsilon\,\left|\left|r_i-r'_i\right|\right| - \epsilon\,\left|\left|z-r'_i\right|\right|.\end{aligned}$$

Therefore:

$$\Pr[\mathbf{q_i} \in A_{y_i}] \leq e^{\epsilon\left|\left|r_i-r'_i\right|\right|} \cdot \int_{A_{y_i}} c e^{-\epsilon\left|\left|z-r'_i\right|\right|} dz,$$

which is $e^{\epsilon\left|\left|r_i-r'_i\right|\right|} \cdot \Pr[\mathbf{q'_i} \in A_{y_i}]$. This completes the first step of the proof.

In the second step, we prove that in the case of Euclidean distance d_χ-STENCIL is 2ϵ-d_χ-private. Observe that due to the independence of $\mathbf{p_i}$ the probability that the mechanism outputs $y = (y_1, \ldots, y_N)$ is the product of the probabilities that the i-th component of the output is y_i for all i. Therefore,

$$\begin{aligned}\Pr[M(x) = y] &= \prod_{i=1}^{N} \Pr[r_i + \mathbf{p_i} \in A_{y_i}] \\ &\leq \prod_{i=1}^{N} e^{\epsilon\left|\left|r'_i-r_i\right|\right|} \Pr[r'_i + \mathbf{p_i} \in A_{y_i}] \\ &= e^{\epsilon \cdot \sum_{i=1}^{N}\left|\left|r'_i-r_i\right|\right|} \cdot \Pr[M(x') = y].\end{aligned}$$

Observe further that due to the definition of r_i, r'_i and to triangle inequality:

$$\begin{aligned}\left|\left|\sum_{i=1}^{N} r'_i - r_i\right|\right| &\leq \sum_{i=1}^{N}\sum_{j=1}^{N} f_{i,j}\,\left|\left|x'_i - x_i\right|\right| \\ &= \sum_{j=1}^{N}\sum_{i=1}^{N} f_{i,j}\,\left|\left|x'_j - x_j\right|\right|.\end{aligned}$$

But, $\sum_{i=1}^{N} f_{i,j} = C_j$, where C_j is the contribution of the j-th token to all other tokens. By Lemma 1, $C_j \leq 2$ for all j.

Recall that $d_2(x, x') = \sum_{j=1}^{N} \left|\left|x'_j - x_j\right|\right|$, and therefore,

$$\begin{aligned}\Pr[M(x) = y] &\leq e^{\epsilon \cdot \sum_{i=1}^{N}\left|\left|r'_i-r_i\right|\right|} \cdot \Pr[M(x') = y] \\ &\leq e^{2\epsilon \sum_{j=1}^{N}\left|\left|x'_j-x_j\right|\right|} \Pr[M(x') = y] \\ &\leq e^{2\epsilon d_2(x,x')} \cdot \Pr[M(x') = y],\end{aligned}$$

which completes the proof of the second step.

To prove the third step, we note that the privacy for Euclidean distance holds for any sequence of embeddings $\mathbb{E}(t_i)$ in Euclidean space $\mathbb{R}^\ell$. In particular, it holds if all the embeddings lie on the unit sphere, which would be the case if the vectors are normalized by dividing each vector by its Euclidean norm.

It is well-known that the cosine distance CD between two vectors x_i, x'_i on the unit sphere is equal to $\frac{1}{2}\,||x'_i - x_i||^2$, for the Euclidean norm $||\cdot||$. On the unit sphere $||x'_i - x_i|| \leq 2$, which implies that $CD(x_i, x'_i) \leq ||x'_i - x_i||$. Plugging this into the final equations of the previous step we have that

$$\begin{aligned}\Pr[M(x) = y] &\leq e^{2\epsilon \sum_{j=1}^{N}\left|\left|x'_j-x_j\right|\right|} \Pr[M(x') = y] \\ &\leq e^{2\epsilon \sum_{j=1}^{N} CD(x_j,x'_j)} \Pr[M(x') = y] \\ &\leq e^{2\epsilon d_C(x,x')} \cdot \Pr[M(x') = y],\end{aligned}$$

which completes the proof.

□

Remark 1. *The conclusions of Theorem 1 regarding the privacy of d_χ-STENCIL are pessimistic for large N. Lemma 1 proves that the contribution C_j of the j-th token is at most 2 in the first and last L tokens, but is exactly 1 otherwise. Repeating the proof of Theorem 1 with this tighter bound on C_j, instead of the looser $C_j \leq 2$ for all j, results in the mechanism providing 2ϵ-d_χ privacy for the two L token substrings at ends of the input, but ϵ-d_χ privacy for the substring in the middle of the input.*