STAMP: Selective Task-Aware Mechanism for Text Privacy

Anonymous ACL submission

Abstract

We present STAMP (Selective Task-Aware Mechanism for Text Privacy), a new framework for task-aware text privatization that achieves an improved privacy–utility trade-off. STAMP selectively allocates privacy budgets across tokens by jointly considering (i) each token's importance to the downstream task (as measured via a task- or query-specific representation), and (ii) its privacy sensitivity (e.g., names, dates, identifiers). This token-level partitioning enables fine-grained, group-wise control over the level of noise applied to different parts of the input, balancing privacy protection with task relevance. To privatize individual token embeddings, we introduce the polar mechanism, which perturbs only the direction of embeddings on the unit sphere while preserving their magnitude. Decoding is performed via cosine nearest-neighbor search, aligning the perturbation geometry with the decoding geometry. Unlike isotropic noise mechanisms, the polar mechanism maintains semantic neighborhoods in the embedding space and better preserves downstream utility. Experimental evaluations on SQuAD, Yelp, and AG News datasets demonstrate that STAMP, when combined with the normalized polar mechanism, consistently achieves superior privacy–utility trade-offs across varying per-token privacy budgets.

1 Introduction

Modern large language models (LLMs) routinely operate on user-supplied text that may contain identifying or otherwise privacy-sensitive content. Practical deployments therefore require client-side protection mechanisms that preserve task utility while preventing the unintended disclosure of sensitive text within the input (Yan et al., 2024) (Pan et al., 2020). This setting encompasses both *inference-time context privacy*—where the user's text is privatized immediately before being sent to a remote model—and *privacy-preserving rewriting*, where the text is locally transformed prior to transmission, storage, sharing, or other downstream use. In this work, we adopt the framework of local differential privacy (LDP) (Arachchige et al., 2019), wherein randomization occurs locally at the user side, ensuring that the server, model owner, or any downstream observer only sees a privatized version of the text. Informally, the released text should not allow an observer to reliably determine whether any specific token was present in the original text.

The Need for Selective, Task-Aware Privacy. Prior approaches to local text privatization face several fundamental limitations. Classical randomized response (Warner, 1965) achieves privacy by randomly replacing each input token with another from the same domain. In natural language, such random substitutions often produce unnatural or incoherent text, severely degrading utility. Adding coordinate-wise Laplace noise (or isotropic Gaussian noise) (Feyisetan and Kasiviswanathan, 2021) (Feyisetan et al., 2020) to embedding vectors is similarly problematic: semantic embeddings are not uniformly sensitive—small perturbations in some directions can flip meanings, while large perturbations in others have negligible effect. Restricting noise to direction-only perturbations helps preserve magnitude (Weggenmann and Kerschbaum, 2021), but when applied uniformly and decoded using mismatched rules, it still distorts fine-grained semantic relations. Applying uniform privacy budgets across all tokens compounds the problem, as it perturbs both innocuous tokens and semantically crucial ones with equal intensity.

A principled mechanism must instead be selective, tailoring privacy to each *token's sensitivity* and *importance to the task*. Prior work has attempted selectivity using intrinsic linguistic heuristics—such as frequency, part of speech, or type-level information content (Meisenbacher and Matthes, 2024b)

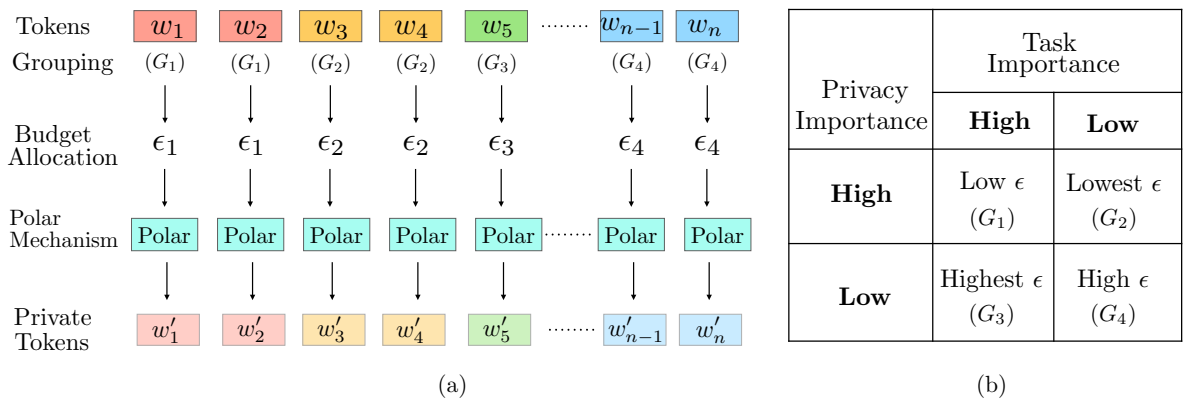

Figure 1: **Overview of the STAMP framework**: Tokens are categorized according to their task and privacy relevance, and then perturbed using adaptively assigned privacy budgets via the polar mechanism. Figure (a) (left) illustrates the overall token perturbation pipeline with group-wise privacy budget allocation; (b) (right) details the grouping and budget assignment process based on task and privacy importance.

(Meisenbacher et al., 2024) —but these ignore the downstream task context. A token may be essential for one query yet irrelevant for another, so its importance varies dynamically across contexts. Such static partitions misallocate privacy, leading to unnecessary noise or insufficient protection. For example, in a question-answering task, the token "`Einstein`" is essential when answering "`Who developed the theory of relativity?`" but irrelevant when the query is "`When was the Nobel Prize established?`". In contrast, in a customer-support context, a token such as "`Alice Johnson`" or "`ID: 123-45-6789`" may be highly sensitive yet irrelevant to the task objective, warranting stronger perturbation despite its low task importance.

Summary of Contributions. To address the above challenges, this work introduces **STAMP** (*Selective Task-Aware Mechanism for Privacy*), a framework for *task-aware text privatization* that provides fine-grained control over privacy–utility trade-offs. Our key contributions are as follows:

1. **Selective, task-aware privacy allocation.** STAMP partitions tokens based on two complementary dimensions: (i) their *importance to the downstream task*, derived from task- or query-specific representations, and (ii) their *privacy sensitivity*, such as the presence of identifiers, names, or dates. This joint assessment enables *group-wise privacy budgeting* that balances utility preservation and privacy protection.

2. **Geometry-aligned perturbation via the Polar Mechanism.** We introduce the *polar mechanism*, which privatizes token embeddings by perturbing only their *direction* on the unit sphere while preserving magnitude. Decoding is performed using *cosine nearest-neighbor search*, aligning the perturbation geometry with the decoding geometry. Unlike isotropic Gaussian or Laplace noise, this approach preserves *semantic neighborhoods* and minimizes distortion of meaning.

3. **Comprehensive empirical evaluation.** Experiments on *SQuAD*, *Yelp*, and *AG News* demonstrate that STAMP, combined with the normalized polar mechanism, consistently achieves *superior privacy–utility trade-offs* across a range of per-token privacy budgets.

2 Preliminaries and Related Work

2.1 Preliminaries on LDP for Text Inputs

We consider the problem of *inference-time context privacy*, where a user wishes to privatize their text input before sending it to a remote model (e.g., an LLM) for inference. Given a user-supplied context $c = (w_1, \ldots, w_n)$ consisting of tokens from a finite vocabulary $\mathcal{V}$, a *local privatization mechanism* $\mathcal{M}$ operates to produce a privatized version $\tilde{c}$. The server or model owner observes only $\tilde{c}$, not the original context. The objective is to preserve task utility for the user while limiting what can be inferred about the presence or absence of sensitive tokens or spans in c. This formulation also naturally extends to *privacy-preserving text rewriting*, where the same mechanism is applied before text storage, sharing, or other downstream use.

Each token $w \in \mathcal{V}$ has an embedding $e(w) \in \mathbb{R}^d$, and we denote its unit-normalized direction by

$\hat{e}(w) = e(w)/\|e(w)\|_2$. A token-level mechanism is a mapping $\mathcal{M} : \mathcal{V} \to \mathcal{Y}$ that outputs a randomized token before transmission. For a context $c = (w_1, \ldots, w_n)$, we write

$$\mathcal{M}^{(n)}(c) = \big(\mathcal{M}(w_1), \ldots, \mathcal{M}(w_n)\big),$$

indicating that the mechanism $\mathcal{M}$ is applied independently to each token.

Definition 1 ((ϵ, δ)-LDP (token level))**.** Let $\mathcal{V}$ be the token vocabulary and $\mathcal{Y}$ the output space (e.g., privatized tokens). A randomized mechanism $\mathcal{M} : \mathcal{V} \to \mathcal{Y}$ satisfies (ϵ, δ)*-local differential privacy (LDP)* if, for every pair of tokens $w, w' \in \mathcal{V}$ and every measurable set $S \subseteq \mathcal{Y}$,

$$\Pr[\mathcal{M}(w) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(w') \in S] + \delta,$$

where $\epsilon \geq 0$ is the privacy budget and δ is a small failure probability.

This definition ensures that observing the privatized token does not allow an adversary to reliably infer whether the original input was w or w'; smaller ϵ (and small δ) imply stronger privacy guarantees.

To better capture linguistic similarity, we also consider a distance-aware relaxation—*metric LDP*—which scales indistinguishability according to a metric d between tokens:

Definition 2 ((ϵ, δ)-metric LDP (token level))**.** Let $\mathcal{V}$ be the token vocabulary and $d : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ a metric on tokens. A randomized mechanism $\mathcal{M} : \mathcal{V} \to \mathcal{Y}$ satisfies (ϵ, δ)*-metric LDP* with respect to d if, for all $w, w' \in \mathcal{V}$ and all measurable $S \subseteq \mathcal{Y}$,

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

Metric LDP is particularly well suited for text, as it allows stronger protection for pairs of semantically similar tokens (small d) while relaxing constraints for distant pairs. In the context of text, the metric $d(\cdot, \cdot)$ can be instantiated using *semantic or geometric similarity* between embeddings—such as cosine distance, Euclidean distance, or distances derived from contextualized representations (Chatzikokolakis et al., 2013) (Chen et al., 2022). This formulation captures the intuition that semantically similar tokens (e.g., "`doctor`" and "`physician`") should be harder to distinguish than unrelated ones (e.g., "`doctor`" and "`banana`"). By embedding this structure directly into the privacy constraint, metric LDP aligns the notion of privacy with the geometry of language, providing a natural foundation for designing mechanisms that operate in embedding space.

2.2 Related Work on Text Privacy

Previous studies have shown that some words directly reveal sensitive information—personal names, email addresses, company identifiers, locations—while others (e.g., stop words) pose little privacy risk and minimal task relevance (Li et al., 2023) (Meisenbacher and Matthes, 2024a). Existing approaches for privatizing text often ignore this heterogeneity (Utpala et al., 2023) (Igamberdiev and Habernal, 2023): they either distort embedding spaces in ways that erase useful structure, or allocate privacy budgets based on language intrinsic properties without regard to task relevance (Meisenbacher et al., 2025).

Isotropic embedding perturbation. A common approach for perturbing embeddings involves adding isotropic Gaussian or Laplace noise (Igamberdiev and Habernal, 2023). While this leverages continuous representations, it implicitly assumes isotropy; yet empirical work shows strong *anisotropy* in embedding spaces—some directions encode fine semantic distinctions while others capture frequency or style (Mu et al., 2017) (Ethayarajh, 2019). Uniform noise disregards this structure and can collapse distinctions embeddings are meant to preserve.

Directional privatization. There exists methods on perturbing directions on the unit sphere, such as von Mises–Fisher (vMF)–based mechanisms and spherical-cap schemes, which offer DP or LDP guarantees under angular metrics (Weggenmann and Kerschbaum, 2021). Directional noise has also been explored for privatizing gradient directions in deep learning models (Faustini et al., 2022). In NLP, word-level metric-DP methods often inject noise in embedding space and then *decode* by cosine nearest-neighbor back to tokens (Arnold et al., 2023; Carvalho et al., 2023). However, when directional noise is applied *uniformly*—without task regard to the current task—and decoded in a mismatched geometry, it can still disrupt fine-grained relations needed downstream.

Uniform budget allocation. In prior works, a common simplification has been to allocate privacy budgets evenly across tokens (Carvalho et al., 2023). This simplification ignores the fact that some tokens—such as named entities or specialized domain terms—are simultaneously highly sensitive and critical for model accuracy, while others contribute little to either privacy or utility (Feyisetan et al., 2020) (Feyisetan and Kasiviswanathan, 2021)

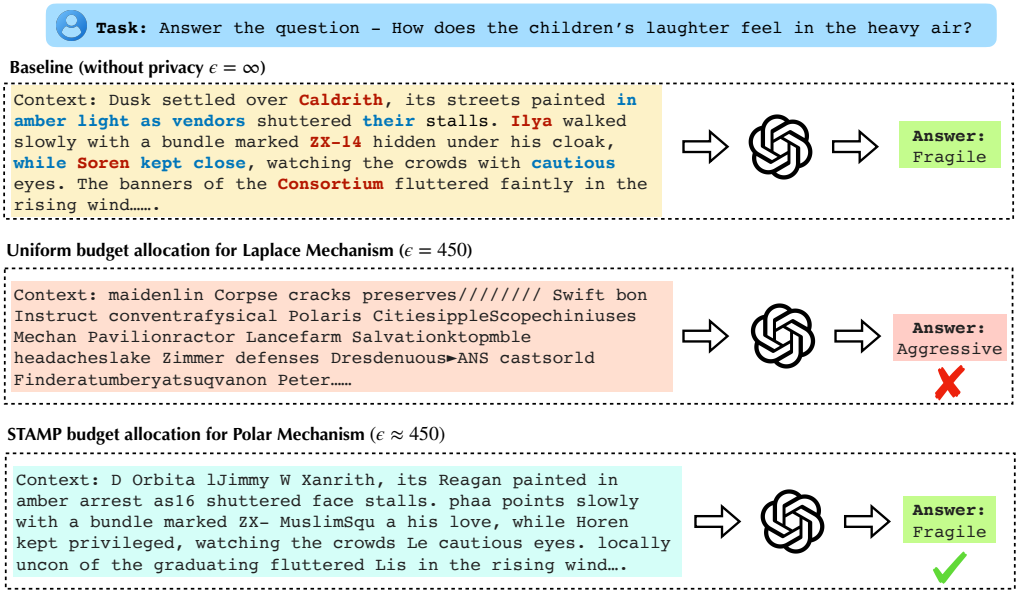

Figure 2: **Prompt-based question answering example** (in the baseline context, **red** bold text indicates privacy-sensitive tokens; **blue** bold text denotes task-relevant text): Uniform perturbation treats all tokens equally, wasting privacy on uninformative words and under-protecting sensitive but predictive ones. STAMP addresses this by stratifying tokens by privacy sensitivity and task relevance, and by aligning the privatization geometry with the decoding geometry.

(Xu et al., 2020). Recent work (Meisenbacher et al., 2025) redistributes the privacy budget ϵ using scores from intrinsic text properties (information content, POS tags, semantic deviation within context C), but remains largely *task-agnostic*.

In contrast, our framework, STAMP , as described next explicitly models both a token's *privacy sensitivity* and its *task relevance*, enabling selective and geometry-aligned privatization that better preserves downstream utility.

3 The STAMP Framework

Unlike mechanisms that allocate a uniform privacy budget, STAMP partitions tokens along two binary axes: *explicit sensitivity* (e.g., PII or NER labels) and *task-dependent importance*. While sensitivity can often be identified from metadata or entity type, importance depends on the specific downstream task—the same token or phrase may be crucial for one query yet irrelevant for another. STAMP captures this by grouping tokens according to their similarity to a task or query representation.

This task-aware selective design yields four token categories (illustrated in Table 1) governed by two principles: (i) sensitive tokens receive the strongest protection, and (ii) among non-sensitive tokens, those with higher task importance are perturbed the least to preserve utility. STAMP assigns each token w to a group $c = g_T(w) \in \{1, 2, 3, 4\}$ via a public, task-specific mapping g_T, and allocates privacy budgets at the *group* level. This selective allocation protects consequential attributes without erasing essential task information. Formally, STAMP instantiates a task-aware metric-LDP mechanism in which budgets are assigned and perturbations applied within each group.

Definition 3 (Task-Aware Metric LDP)**.** Let T be a downstream task with a vocabulary partition $\{\mathcal{V}_1^T, \mathcal{V}_2^T, \mathcal{V}_3^T, \mathcal{V}_4^T\}$. A randomized mechanism $\mathcal{M} : \mathcal{V} \to \mathcal{Y}$ satisfies *task-aware metric LDP* with budget vector $\boldsymbol{\epsilon}^T = (\epsilon_T^{(1)}, \epsilon_T^{(2)}, \epsilon_T^{(3)}, \epsilon_T^{(4)})$ if, for every group $c \in \{1, 2, 3, 4\}$, all tokens $w, w' \in \mathcal{V}_c^T$, and all measurable sets $S \subseteq \mathcal{Y}$,

$$\Pr[\mathcal{M}(w) \in S] \leq e^{\epsilon_T^{(c)} d_u(w,w')} \Pr[\mathcal{M}(w') \in S],$$

where d_u is a distance metric.

Theorem 1 (Privacy Guarantee of STAMP)**.** *Fix a task T and assume that T and its associated grouping map g_T are public. Then* STAMP *satisfies task-aware metric LDP with budget vector $\boldsymbol{\epsilon}^T = (\epsilon_T^{(1)}, \epsilon_T^{(2)}, \epsilon_T^{(3)}, \epsilon_T^{(4)})$.*

Grouping tokens and perturbing them with task-specific budgets yields $(\epsilon_T^{(c)}, 0)$-metric LDP for each group. The complete proof appears in Appendix B.1. In the limiting case where a token's

privacy budget $\epsilon = 0$ (for highly sensitive, low-utility tokens), the mechanism reveals no information about that token, effectively replacing it with a fully masked or placeholder symbol.

Extending Token-Level Guarantees to Contexts. While the single-token guarantee secures individual embeddings, real applications release entire sentences or documents. Privacy loss compounds when multiple tokens are observed together, so we extend the guarantee to sequences via standard composition in metric LDP. Applying the mechanism independently to each position extends per-token guarantees to full contexts (Wang et al., 2020) (Kairouz et al., 2015). Consider a sequence $w^{(n)} = (w_1, \ldots, w_n)$ with group labels $c_i = g_T(w_i)$. The STAMP mechanism applies independently to each token:

$$\mathcal{M}_T^{(n)}(w^{(n)}) = \big(\mathcal{M}_T(w_1), \ldots, \mathcal{M}_T(w_n)\big).$$

Theorem 2 (Context-Level Privacy Guarantee of STAMP)**.** *For any task T, $\mathcal{M}_T^{(n)}$ satisfies task-aware metric LDP with budget vector $(\epsilon_{T,1}^{(c_1)}, \ldots, \epsilon_{T,n}^{(c_n)})$, where $\epsilon_{T,i}^{(c_i)} = \epsilon_T^{(c_i)}$. In particular, for any pair of sequences $w^{(n)}, w'^{(n)}$ with the same group sequence $(c_1, \ldots, c_n)$ and any measurable set S,*

$$\begin{aligned}&\Pr[\mathcal{M}_T^{(n)}(w^{(n)}) \in S] \\ \leq &e^{\sum_{i=1}^n \epsilon_T^{(c_i)} d_u(w_i, w_i')} \Pr[\mathcal{M}_T^{(n)}(w'^{(n)}) \in S].\end{aligned}$$

Since STAMP privatizes each token independently, the sequence-level guarantee follows directly by composition; see Appendix B.2.

3.1 Group Maps & Privacy Budget Allocation

Designing the Group Map. The grouping map g_T partitions tokens along two dimensions: explicit sensitivity and task relevance. Sensitivity can be detected using named-entity or PII recognizers (e.g., for names, organizations, or identifiers), while task relevance can be estimated from gradient-based saliency, attention scores, or cosine similarity to a task- or query-specific embedding. Each token w is thus assigned to one of four groups based on its sensitivity and importance for task T. Making g_T public simplifies analysis and reproducibility, as the grouping policy itself contains no private information and enables well-defined, composable privacy guarantees.

Allocating Privacy Budgets. Once groups are defined, privacy budgets $\boldsymbol{\epsilon}^T$ are assigned according to both sensitivity and relevance. Sensitive tokens (e.g., names or identifiers) receive the strongest protection through smaller budgets, while non-sensitive yet task-critical tokens are granted larger budgets to preserve downstream utility. Tokens that are both non-sensitive and task-irrelevant tolerate stronger perturbations without degrading performance. A practical rule of thumb is to assign privacy budgets that increase with task importance and decrease with privacy sensitivity. In our experiments, we adopt the following proportional allocation:

$$\epsilon_T^{(1)} : \epsilon_T^{(2)} : \epsilon_T^{(3)} : \epsilon_T^{(4)} = 2 : 1 : 4 : 3,$$

where group 1 contains tokens that are both highly privacy-sensitive and task-critical (thus receiving a moderate budget to balance the trade-off); group 2 includes tokens that are highly sensitive but less important to the task (hence the smallest budget); group 3 corresponds to tokens with lower privacy sensitivity but high task importance (receiving the largest budget); and group 4 covers tokens with low sensitivity and moderate task importance. This allocation provides a principled balance between protecting sensitive content and maintaining utility for downstream tasks.

4 Polar Mechanism under Metric LDP

In this Section, we present a geometry-aware pipeline for privatizing token embeddings that aligns the *perturbation geometry* with the *decoding geometry*. Rather than injecting isotropic noise in $\mathbb{R}^d$, we factor each embedding into **radial** and **angular** components and privatize each in its natural metric. For an embedding $\mathbf{e} \in \mathbb{R}^d$ of token w,

$$\mathbf{e} = r \cdot \mathbf{u}, \qquad r = \|\mathbf{e}\|_2, \qquad \mathbf{u} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2} \in \mathbb{S}^{d-1}.$$

(a) Magnitude/Radial Perturbation via Laplace. We can privatize the radius with Laplace noise

$$r' \sim \mathrm{Laplace}(r, b_r), \qquad b_r = \Delta_r / \epsilon_r,$$

which gives $(\epsilon_r, 0)$-metric LDP with respect to $d_r(r, r') = |r - r'|$.

(b) Direction Perturbation via vMF noise. Independently, the direction is privatized on the unit sphere with a von Mises–Fisher (vMF) noise

$$\mathbf{u}' \sim \mathrm{vMF}(\mu = \mathbf{u}, \kappa),$$

where $\kappa \geq 0$ controls concentration. Let $d_c(\mathbf{u}, \mathbf{u}') = \|\mathbf{u} - \mathbf{u}'\|_2$ denote the chordal metric on $\mathbb{S}^{d-1}$. We formalize the privacy guarantee by the vMF noise based mechanism below.

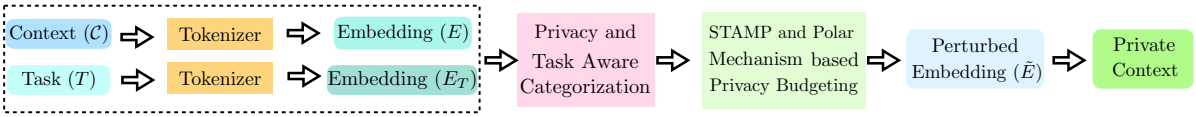

Figure 3: **STAMP privatization pipeline.** Embeddings are decomposed into radial and angular components, perturbed under metric LDP, and decoded by angular proximity into privatized tokens suitable for downstream tasks.

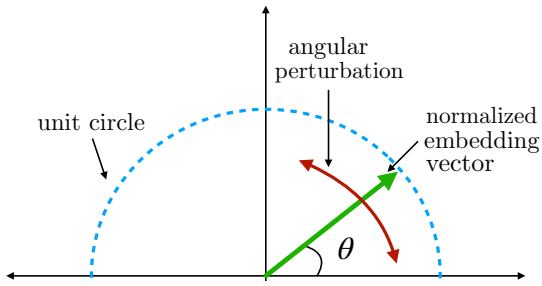

Figure 4: Directional privatization using the Polar mechanism, which perturbs the angle of the normalized embedding vector while normalizing its magnitude.

Theorem 3 (vMF mechanism satisfies metric LDP)**.** *Let $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ and consider the angular mechanism*

$$\mathcal{M}(u) \sim \mathrm{vMF}(\mu = u, \kappa),$$
$$\textit{with density } f(y \mid u) = C_d(\kappa) \exp\left(\kappa\, u^\top y\right),$$

where $C_d(\kappa)$ is the vMF normalizing constant. Then $\mathcal{M}$ satisfies metric local differential privacy with respect to the metric $d(u, u') = \|u - u'\|_2$, in the sense that for all $u, u' \in \mathbb{S}^{d-1}$ and all measurable $S \subseteq \mathbb{S}^{d-1}$,

$$\Pr[\mathcal{M}(u) \in S] \le e^{\left(\kappa\, \|u - u'\|_2\right)} \Pr[\mathcal{M}(u') \in S].$$

Equivalently, the mechanism is ϵ-metric LDP with parameter $\epsilon = \kappa$ under metric $d(u, u') = \|u - u'\|_2$.

The full proof is deferred to Appendix B.3. Combining the two independent perturbations yields a *two-budget* guarantee (ϵ_r, ϵ_u).

Semantic Decoding The polar mechanism outputs a privatized embedding $\mathbf{e}' = r'\,\mathbf{u}' \in \mathbb{R}^d$ which may not correspond to a valid token. Therefore, a downstream *decoding step* identifies the token whose embedding is most similar in cosine space:

$$\mathrm{Decode}(\mathbf{e}') = \arg\max_{\mathbf{v} \in E} \frac{\mathbf{e}'^\top \mathbf{v}}{\|\mathbf{e}'\|_2\, \|\mathbf{v}\|_2},$$

Because the above decoding rule depends only on direction, perturbations of $\|\mathbf{e}'\|_2$ cannot affect the argmax—a property we call *radial invariance*. This motivates discarding the radial (magnitude) perturbation entirely and privatizing only the directional component.

Normalized Polar Mechanism. Motivated by this radial invariance, we project each embedding onto the unit sphere—discarding the original radius entirely—and privatize only its direction:

$$\bar{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}, \qquad \mathbf{u}' \sim \mathcal{M}_u(\bar{\mathbf{e}})\ (\text{vMF}).$$

This yields $(0, \epsilon_u)$-metric LDP under the angular metric and reduces Polar to a single-budget, direction-only mechanism in practice. In particular, the *magnitude is perfectly private*: the released vector has fixed norm ($\|\mathbf{u}'\|_2 = 1$), so the output distribution is independent of r, i.e., 0-metric LDP in the radial component. As illustrated in Figure 4, the normalized Polar mechanism perturbs only the angular component while keeping the magnitude constant.

Computational Complexity. Decoding then reduces to nearest-neighbor search over random directions. The computational cost is that of standard nearest-neighbor search on $\mathbb{S}^{d-1}$: $\mathcal{O}(|\mathcal{V}|\, d)$ for exact search, and sublinear in $|\mathcal{V}|$ using modern approximate nearest-neighbor (ANN) methods such as HNSW or FAISS (Johnson et al., 2019) (Malkov and Yashunin, 2018). Formally, on the unit sphere, ranking by cosine similarity is equivalent to nearest-neighbor search under both geodesic (angular) and chordal (Euclidean) distances, aligning the decoding geometry with our privacy metric. We formalize this property in the next Proposition.

Proposition 1 (Equivalence of Semantic Decoding and Nearest-Neighbor Search on the Sphere)**.** *Let $E = \{v_1, \ldots, v_{|\mathcal{V}|}\} \subset \mathbb{R}^d$ denote the set of token embeddings, and let $\hat{e} = e'/\|e'\|_2$ be the direction of a privatized vector $e' \neq 0$. Then the following*

Group Privacy Budgets $\{G_1, G_2, G_3, G_4\}$	SQuAD (Cosine similarity)		Non-Private Baseline	Yelp (Acc %)		Non-Private Baseline	AG News (Acc %)		Non-Private Baseline
	Polar (vMF)	Laplace		Polar (vMF)	Laplace		Polar (vMF)	Laplace	
{150, 50, 450, 350}	0.393	0.325	0.839	0.360	0.220	0.580	0.540	0.560	0.920
{200, 100, 500, 400}	0.470	0.334	0.839	0.380	0.140	0.580	0.640	0.580	0.920
{250, 150, 550, 450}	0.587	0.335	0.839	0.480	0.180	0.580	0.680	0.560	0.920
{300, 200, 600, 500}	0.654	0.341	0.839	0.560	0.200	0.580	0.760	0.520	0.920
{350, 250, 650, 550}	0.833	0.343	0.839	0.560	0.140	0.580	0.800	0.520	0.920

Table 1: Polar vs. Laplace vs. Baseline at matched protocol (cosine decoding for all), embedding dimension at d=768. Rows are privacy points; columns group mechanisms by dataset. *SQuAD (Cosine)* cells show cosine similarity. *Yelp/AG News* cells show accuracy.

decoding rules are equivalent:

$$\begin{aligned}\mathrm{Decode}(e') &= \arg\max_{v\in E} \frac{e'^{\top} v}{\|e'\|_2\,\|v\|_2} && (1)\\ &= \arg\max_{v\in E} \hat{e}^{\top}\hat{v} && (2)\\ &= \arg\min_{v\in E} \arccos\big(\hat{e}^{\top}\hat{v}\big) && (3)\\ &= \arg\min_{v\in E} \|\hat{e}-\hat{v}\|_2, && (4)\end{aligned}$$

where $\hat{v} = v/\|v\|_2$ for $v \neq 0$, and any v with $\|v\|_2 = 0$ can be excluded. In particular, decoding by cosine similarity is identical to nearest-neighbor search on the unit sphere under both geodesic (angular) and Euclidean (chordal) distances.

The proof is provided in Appendix B.4. Angular decoding preserves semantic neighborhoods while discarding norm information, which often correlates with token frequency or salience. Together with Polar perturbation, this alignment yields privatized text consistent with cosine semantics without introducing additional computational overhead.

Algorithm 1 STAMP–Polar Mechanism

Require: Context $c = (w_1, \ldots, w_n)$; task T; grouping map $g_T : \mathcal{V} \to \{1,2,3,4\}$; group privacy budgets $\{\epsilon_u^{(c)}\}_{c=1}^4$; vocabulary embeddings $V = \{\mathbf{v}\}$

Ensure: Privatized sequence $\tilde{c} = (w'_1, \ldots, w'_n)$

for $i = 1$ **to** n **do**
  $c_i \leftarrow g_T(w_i)$
  $\mathbf{e}_i \leftarrow \mathbf{e}(w_i)$; $\bar{\mathbf{e}}_i \leftarrow \mathbf{e}_i/\|\mathbf{e}_i\|_2$
  $\kappa_i \leftarrow \epsilon_u^{(c_i)}$
  Sample $\mathbf{u}'_i \sim \mathrm{vMF}(\mu = \bar{\mathbf{e}}_i,\ \kappa_i)$
  $w'_i \leftarrow \arg\max_{\mathbf{v}\in V} \frac{\mathbf{u}_i'^{\top}\mathbf{v}}{\|\mathbf{v}\|_2}$ *(decoding)*
end for
return $\tilde{c} = (w'_1, \ldots, w'_n)$

Figure 3 illustrates the overall STAMP workflow: tokens are grouped using the public task map g_T, group-level privacy budgets are assigned, embeddings are privatized, and the resulting representations are decoded back to tokens. The privatization step is instantiated with the *normalized Polar* mechanism. Algorithm 1 summarizes the context-level procedure: for each token position i, determine its group label $c_i = g_T(w_i)$, normalize its embedding, sample a vMF-perturbed direction with concentration parameter $\kappa_i = \epsilon_u^{(c_i)}$, and decode by cosine similarity to obtain the privatized token w'_i.

5 Experiments

In This Section, we present the experimental results on three commonly used datasets: SQuAD (Rajpurkar et al., 2016), Yelp (Yelp, Inc., 2025), and AG News (Zhang et al., 2015), and analyzed the privacy utility trade-off for Polar (vMF), Laplace Mechanism and non-private baseline Mask and Fill. Specifically, we tried answering two questions: (i) at matched privacy budget, how do the Polar (angular vMF) and isotropic Laplace mechanisms compare in utility; and (ii) at matched privacy budget, how does STAMP framework compare to uniform framework.

Evaluators, rewrite, and metrics. For QA we use GPT-4 prompted to answer questions from the privatized context; for classification we use existing models (Kmack, 2024) (TextAttack, 2020) trained on the original (non-privatized) training splits of Yelp and AG News, with privatization applied only at inference on test. Unless stated otherwise, results are reported without rewriting. For SQuAD, we additionally include an optional post-privatization rewrite variant that performs coherence repair without introducing external facts. QA metrics are EM (exact match) accuracy and cosine similarity between predicted and gold answers (Sentence-BERT); classification tasks use accuracy. We report dataset means averaged over three runs.

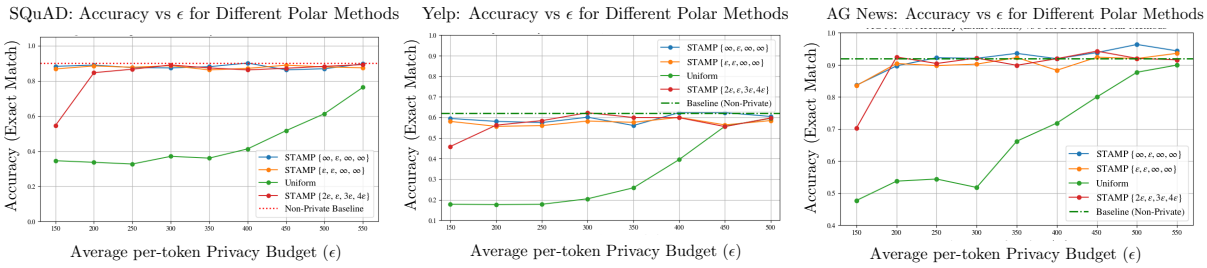

Figure 5: STAMP vs. Uniform (single angular budget) under the Polar mechanism across three tasks: SQuAD (QA), Yelp (sentiment), and AG News (topic classification). Curves show task performance versus the *base* angular budget ϵ_0; both framework use normalized polar mechanism and cosine decoding. STAMP variants (e.g., private-only, private-low, heuristic) retain higher utility by concentrating noise on sensitive, low-importance spans, while Uniform converges as noise vanishes at large ϵ_0.

5.1 Experimental Setup

We evaluate our proposed approach on three datasets for different tasks—SQuAD (question answering), Yelp (sentiment rating), and AG News (topic classification)—using task-specific metrics: answer cosine similarity for QA and accuracy for classification. All comparisons are conducted at matched per-token ϵ values. Tokens are partitioned into a 2×2 grid by importance $\times$ privacy: importance is $\cos(\text{token}, \text{query}) \geq \tau$ with $\tau = 0.5$, and privacy follows standard NER/PII rules. We report realized $(\#G_1, \ldots, \#G_4)$ and the resulting $\bar{\epsilon}$.

Our code is available at [1].

5.2 Polar vs Laplace

We compare *Polar mechanism* (normalized, direction-only vMF) to *isotropic Laplace mechanism* under identical privacy budgets and the same embedding dimension (d=768). For this comparison only, we sweep a *per-token* privacy budget specific to the mechanism (directional ϵ_{dir} for Polar, isotropic ϵ_{iso} for Laplace) and report the mean over the sweep. Table 1 summarizes results (SQuAD: answer cosine similarity; Yelp/AG News: accuracy), averaged over three seeds. Across datasets, Polar consistently yields a stronger utility–privacy trade-off: at relatively low budgets, Laplace collapses toward chance-level performance while Polar remains effective; as the budget increases, Polar mechanism's performance improves smoothly and approaches the non-private baseline, whereas Laplace continues to lag across the sweep.

5.3 STAMP vs Uniform

At matched per-token privacy budget ϵ (Fig. 5), STAMP consistently outperforms a Uniform scheme in utility, especially in the low–to–mid privacy regime. STAMP preserves high-importance content while concentrating noise on spans that are privacy-sensitive yet task-unimportant. Utility is highest when perturbations are confined to such a small subset; as the protected set expands (from one group to two, three, and then all four groups), performance degrades monotonically. Uniform—allocating the same budget to every token—forms the lower envelope across the grid. Although STAMP perturbs only a subset of groups, it effectively protects privacy-sensitive content while leaving high-importance tokens largely intact, thereby avoiding the typical utility penalty. Moreover, question-conditioned importance (e.g., SQuAD) amplifies STAMP's gains, whereas fixed prompts (Yelp/AG News) dampen them by stabilizing group proportions. Beyond utility, STAMP provides a controllable knob over which content to protect, aligning with the paper's claim that privacy preferences are inherently subjective and task-dependent.

6 Conclusion

We introduced STAMP , a task-aware mechanism for privatizing text under local differential privacy. By combining geometry-aware perturbation with task-dependent budget allocation, STAMP aims to balance privacy and utility across diverse NLP tasks. We provide formal guarantees and report empirical results on SQuAD, Yelp, and AG News. STAMP outperforms Uniform at matched per-token budgets and Polar surpasses isotropic Laplace. Future work will address dynamic tasks and sequence-level dependencies, moving toward more robust privacy-preserving NLP systems.

[1] https://anonymous.4open.science/r/STAMP-D6F5

7 Limitations

STAMP represents an initial step toward task-aware LDP for text, but it carries several limitations. Most notably, the framework assumes the availability of a meaningful task description that can be encoded as a fixed representation at inference time; this assumption may not hold in interactive, open-ended, or multi-turn scenarios. Additionally, token-level relevance grouping is based on static embedding similarity, which may miss nuances of functional importance for tasks involving complex syntactic structures, discourse-level reasoning, or external knowledge. Finally, STAMP does not currently account for long-range dependencies or structured interactions between tokens during budget allocation, which may result in over- or under-perturbation in semantically rich contexts. We leave addressing these challenges to future work.

8 Ethical Considerations

STAMP is designed to enhance privacy in NLP by providing formal LDP guarantees, enabling the use of sensitive corpora such as email or financial text without exposing raw content. This represents a positive step toward building systems where privacy is a first-class objective alongside accuracy.

At the same time, several risks remain. Task-aware budgets may protect some attributes more strongly than others, raising the possibility of uneven coverage across demographic or domain-specific categories. Careless or adversarial configuration of privacy budgets could also weaken effective guarantees, giving a false sense of protection. Finally, STAMP focuses on token-level privatization and does not address broader concerns such as fairness, data misuse, or downstream harms that can arise even from privatized text.

We emphasize that STAMP is a methodological tool, not a complete solution, and that responsible deployment requires auditing, fairness checks, and clear communication of privacy parameters.

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.

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Driving context into text-to-text privatization. *arXiv preprint arXiv:2306.01457*.

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)*, pages 883–890. SIAM.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *international symposium on privacy enhancing technologies symposium*, pages 82–102. Springer.

Huimin Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2022. A customized text sanitization mechanism with differential privacy. *arXiv preprint arXiv:2207.01193*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations. *Comparing the geometry of BERT, ELMo, and GPT-2 Embeddings*, 2.

Pedro Faustini, Natasha Fernandes, Shakila Tonni, Annabelle McIver, and Mark Dras. 2022. Directional privacy for deep learning. *arXiv preprint arXiv:2211.04686*.

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.

Oluwaseyi Feyisetan and Shiva Kasiviswanathan. 2021. Private release of text embedding vectors. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 15–27.

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. *arXiv preprint arXiv:2302.07636*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR.

Kmack. 2024. Yelp-review_classifier. Hugging Face model.

Haoran Li, Yulin Chen, Jinglong Luo, Jiecong Wang, Hao Peng, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, Zenglin Xu, and 1 others. 2023. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Stephen Meisenbacher, Chaeeun Joy Lee, and Florian Matthes. 2024. Spend your budget wisely: Towards an intelligent distribution of the privacy budget in differentially private text rewriting. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy*, pages 84–95.

Stephen Meisenbacher, Chaeeun Joy Lee, and Florian Matthes. 2025. Spend your budget wisely: Towards an intelligent distribution of the privacy budget in differentially private text rewriting. *arXiv preprint arXiv:2503.22379*.

Stephen Meisenbacher and Florian Matthes. 2024a. 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*, pages 1–11.

Stephen Meisenbacher and Florian Matthes. 2024b. Thinking outside of the differential privacy box: A case study in text privatization with language model prompting. *arXiv preprint arXiv:2410.00751*.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

TextAttack. 2020. textattack/distilbert-base-uncased-ag-news. Hugging Face model; DistilBERT fine-tuned on AG News.

Saiteja Utpala, Sara Hooker, and Pin Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. *arXiv preprint arXiv:2310.16111*.

Teng Wang, Xuefeng Zhang, Jingyu Feng, and Xinyu Yang. 2020. A comprehensive survey on local differential privacy toward data statistics and analysis. *Sensors*, 20(24):7030.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, 60(309):63–69.

Benjamin Weggenmann and Florian Kerschbaum. 2021. Differential privacy for directional data. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1205–1222.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*.

Yelp, Inc. 2025. Yelp open dataset.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Dataset: AG's News Topic Classification.

Appendix Index

This appendix provides supplementary material referenced in the main text:

- Threat Model (§A)
- Privacy Guarantees: Detailed Proofs (§B)
  - Proof of Theorem 1
  - Proof of Theorem 2 (Sequence-level composition)
  - Proof of Theorem 3 (vMF is κ-mLDP)
  - Proof of Proposition 1 (Cosine decoding is MLE under vMF)

A Threat Model

Setting. Each user holds a private text input C. Before transmission or use for task T, the user applies the local mechanism M to obtain a perturbed output $\tilde{C} = M(C, T)$ and only $\tilde{C}$ is shared.

Aggregator model (honest-but-curious). The central server follows the protocol (collects $\tilde{C}$) but may attempt to infer information about the original inputs C from $\tilde{C}$.

Adversary's goals. Infer sensitive content or attributes of C, including:

1. identifying specific sensitive tokens/spans (e.g., names, locations, medical terms);
2. inferring user attributes correlated with C (e.g., demographics or preferences);
3. partial or full reconstruction of C.

Adversary's knowledge. We assume knowledge of: (i) the mechanism M and its parameters, (ii) the vocabulary/embeddings, (iii) the public grouping map g_T for task T, and (iv) any auxiliary background knowledge. The adversary observes only privatized outputs (and decoded tokens, if decoding is used), not raw inputs.

User/device assumptions. Randomization occurs locally; per-position channels are sampled independently conditioned on inputs and groups; the RNG is not adversarially controlled.

Public disclosure. The grouping map g_T and mechanism hyper-parameters (e.g., κ or ϵ per group) are treated as public; privacy does not rely on secrecy of g_T.

Out-of-scope threats. We do not address side channels (timing, memory), compromised clients, corrupted RNG, or post-aggregation leakage unrelated to the local mechanism.

Guarantees in scope. Protection follows the metric LDP guarantees for the angular channel (Theorem 3) and their sequence-level composition (Theorem 2). Experiments report the *per-token* budget $\bar{\epsilon}$ and realized group counts $(\#G_1, \#G_2, \#G_3, \#G_4)$; $\bar{\epsilon}$ is computed from the observed mix for comparability across runs.

B Privacy Guarantees: Detailed Proofs

Notation. We work on the unit sphere $\mathbb{S}^{d-1} = \{y \in \mathbb{R}^d : \|y\|_2 = 1\}$. Geodesic (angular) distance is $d_g(u, v) = \arccos(u^\top v)$ and chordal distance is $d_2(u, v) = \|u - v\|_2$. They satisfy $d_2(u, v) = 2\sin\big(\frac{1}{2} d_g(u, v)\big) \le d_g(u, v)$. The vMF density with mean $\mu \in \mathbb{S}^{d-1}$ and concentration $\kappa \ge 0$ is $f(y \mid \mu) = C_d(\kappa) \exp\big(\kappa\, \mu^\top y\big)$, where $C_d(\kappa)$ is independent of μ.

B.1 Proof of Theorem 1 (MLDP for STAMP)

Proof. Let g_T be the public grouping map for task T assigning token w to group $c = g_T(w)$. Condition on any fixed group c. The angular channel is $Y \sim \mathrm{vMF}(\mu(w), \kappa_c)$; by Theorem 3, it satisfies $(\epsilon, 0)$-metric LDP with $\epsilon = \kappa_c$ on $\mathbb{S}^{d-1}$. As g_T is a deterministic (public) function of the input, restricting to the subset with $g_T(w) = c$ preserves the worst-case guarantee. Therefore the mechanism is $(\epsilon_T^{(c)}, 0)$-metric LDP with $\epsilon_T^{(c)} = \kappa_c$ for each group c. □

B.2 Proof of Theorem 2 (Sequence-level composition)

Proof. Let $x = (x_1, \ldots, x_n)$ and independent per-position channels M_t that are ϵ_t-mLDP w.r.t. a single-token metric d. For any measurable $S \subseteq \mathcal{Y}^n$,

$$\frac{\Pr[M(x) \in S]}{\Pr[M(x') \in S]} \le \prod_{t=1}^{n} \exp\big(\epsilon_t\, d(x_t, x'_t)\big)$$

$$= \exp\Big(\sum_{t=1}^{n} \epsilon_t\, d(x_t, x'_t)\Big).$$

Thus the product channel is $(\sum_t \epsilon_t)$-mLDP w.r.t. the sequence metric $d_\Sigma(x, x') = \sum_t d(x_t, x'_t)$. For group-specific budgets, set $\epsilon_t = \epsilon(c_t)$ with $c_t = g_T(x_t)$ to obtain the stated form. □

B.3 Proof of Theorem 3 (vMF is κ-mLDP)

Proof. Fix unit $\mu, \nu \in \mathbb{S}^{d-1}$ and measurable $S \subseteq \mathbb{S}^{d-1}$. Using the vMF density and cancellation of $C_d(\kappa)$,

$$\begin{aligned} \log \frac{f(y \mid \mu)}{f(y \mid \nu)} &= \kappa\,(\mu - \nu)^\top y \\ &\le \kappa\,\|\mu - \nu\|_2\,\|y\|_2 \\ &= \kappa\,\|\mu - \nu\|_2 \end{aligned}$$

By Cauchy–Schwarz and $\|y\|_2 = 1$, exponentiating and taking supremum over S yields

$$\Pr[M(\mu) \in S] \;\le\; \exp\!\big(\kappa\|\mu - \nu\|_2\big)\;\Pr[M(\nu) \in S],$$

i.e., $(\epsilon, 0)$-metric LDP with $\epsilon = \kappa$ under d_2. Since $d_2(\mu, \nu) \le d_g(\mu, \nu)$ the same bound holds under d_g.
Let $\theta = d_g(u, v) \in [0, \pi]$. Then $d_2(u, v) = 2\sin(\theta/2) \le \theta = d_g(u, v)$ because $\sin x \le x$ for $x \ge 0$. □

B.4 Proof of Proposition 1 (Cosine decoding)

Proof. For fixed κ, $\log f(y \mid \mu(w)) = \kappa\,\mu(w)^\top y + \mathrm{const}$, hence

$$\arg\max_w \log f(y \mid \mu(w)) = \arg\max_w \mu(w)^\top y,$$

which is exactly cosine nearest-neighbor on $\mathbb{S}^{d-1}$. □