ANISOTROPIC NOISE FINGERPRINTS REVEAL CONCEPT CHOICE IN CONCEPT-AWARE EMBEDDING PRIVACY

FARS
Analemma
fars@analemma.ai

ABSTRACT

Concept-aware embedding sanitization applies learned, concept-specific anisotropic noise to protect privacy while preserving utility. We show that under multi-release access—where an attacker observes multiple sanitized versions of the same document—the anisotropic noise structure leaks which privacy concept was selected. Our variance fingerprint attack computes per-dimension variance profiles from embedding differences and matches them to concept templates, achieving 100% concept-identification accuracy on SPARSE-style sanitization, compared to 18.9% for isotropic noise (near the 20% chance level for $K = 5$ concepts). Covariance smoothing fails to mitigate the attack: even at $\lambda = 0.99$ (mixing 99% identity covariance), accuracy remains 63.3%. Importantly, this is metadata leakage about the privacy mode, not content leakage—token-level privacy is preserved across all conditions (AUC $\sim$0.52 vs 0.87 for clean embeddings). Our findings reveal a fundamental limitation of concept-aware anisotropic noise under realistic multi-release threat models.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Dense text embeddings are fundamental to modern information retrieval systems, including vector databases, retrieval-augmented generation (RAG), and semantic search. However, these embeddings encode rich semantic information that can be exploited to recover sensitive content. Embedding inversion attacks (Morris et al., 2023; Li et al., 2023; Song & Raghunathan, 2020) demonstrate that text can be reconstructed from embeddings with high fidelity, raising significant privacy concerns for systems that store or share embeddings.

Differential privacy mechanisms offer a principled approach to mitigate these risks by adding calibrated noise to embeddings before release. Recent work has proposed *concept-aware* sanitization, where users specify which sensitive concept to protect (e.g., medical conditions, locations, demographics), and the mechanism applies stronger perturbations along embedding dimensions that encode that concept. SPARSE (Tsai et al., 2026) exemplifies this approach: it learns concept-specific sensitivity masks and applies anisotropic Mahalanobis noise calibrated by dimension sensitivity, achieving superior utility-privacy tradeoffs compared to isotropic noise.

However, concept-aware sanitization introduces a new attack surface: the privacy configuration itself. In many deployments, the choice of privacy concept is sensitive—selecting "protect diseases" may imply medical context, while "protect politics" may indicate political affiliation risk. If an attacker can infer which concept was selected, they can profile users even without recovering document content.

We identify a fundamental vulnerability in concept-aware anisotropic noise under *multi-release access*. In practice, embeddings may be released multiple times through periodic re-indexing, service

[1] https://gitlab.com/fars-a/sparse-concept-choice-leakage

retries, or version snapshots. When the sanitizer re-samples noise per release, an attacker who observes multiple releases of the same document can estimate noise statistics. For anisotropic mechanisms, the per-dimension variance profile acts as a unique fingerprint that reveals the concept choice.

We propose a simple variance fingerprint attack that achieves **100% concept-identification accuracy** on SPARSE-style sanitization with just 10 releases per document, compared to 18.9% for isotropic noise (near the 20% chance level for $K = 5$ concepts). Covariance smoothing—mixing concept-specific covariance with identity—fails to mitigate the attack: even at $\lambda = 0.99$ (99% identity mixing), accuracy remains 63.3%, well above chance. Importantly, we show this is *metadata leakage* about the privacy mode, not *content leakage*: token-level privacy is preserved across all conditions.

Our contributions are fourfold. First, we propose a variance fingerprint attack that achieves 100% concept-identification accuracy under multi-release access, demonstrating complete concept-choice leakage for SPARSE-style anisotropic noise. Second, we evaluate covariance smoothing as a mitigation strategy and show it fails even at extreme smoothing ($\lambda = 0.99$), where accuracy remains 63.3%. Third, we provide mechanistic analysis confirming that leakage requires the specific dimension-concept alignment of learned masks, not merely anisotropic noise structure—random permutation of covariance diagonals eliminates the attack. Fourth, we characterize the scope of the vulnerability, showing that concept-choice leakage is metadata inference about the privacy mode, not content recovery—token-level privacy remains intact across all conditions.

2 Related Work

Embedding Privacy Attacks. Text embeddings encode rich semantic information that can be exploited to recover sensitive content. Song & Raghunathan (2020) systematically studied information leakage in embedding models, demonstrating embedding inversion attacks that recover 50–70% of input words, attribute inference attacks that extract authorship, and membership inference attacks on training data. Vec2Text (Morris et al., 2023) advanced embedding inversion by framing it as controlled generation, achieving 92% exact recovery of 32-token inputs through iterative correction and re-embedding. Li et al. (2023) proposed generative embedding inversion attacks (GEIA) that train decoder models to reconstruct full sentences from embeddings. These attacks focus on *content recovery*—extracting the original text or its attributes. Our work identifies a distinct vulnerability: *metadata inference* about the privacy mechanism itself.

Differential Privacy for Embeddings. Differential privacy provides formal guarantees for protecting sensitive information in embeddings. Feyisetan et al. (2019) applied calibrated multivariate perturbations to word embeddings using d_χ-privacy, achieving practical utility with less than 2% loss on binary classification. Local differential privacy (LDP) has been widely deployed for privacy-preserving data collection (Yang et al., 2020), with mechanisms like RAPPOR (Erlingsson et al., 2014) enabling randomized responses. Recently, Tsai et al. (2026) proposed SPARSE, a concept-aware framework that learns dimension-specific masks to identify privacy-sensitive embedding dimensions and applies anisotropic Mahalanobis noise calibrated by dimension sensitivity. While SPARSE achieves superior utility-privacy tradeoffs compared to isotropic noise, we show that its concept-specific covariance structure creates exploitable fingerprints under multi-release access.

Multi-Release Privacy. Privacy guarantees degrade under repeated data releases. Composition theorems (Mironov, 2017) quantify this degradation for differential privacy. Pool inference attacks (Hiesgen et al., 2023) demonstrate that aggregating multiple LDP responses can compromise individual privacy guarantees. Lokna et al. (2023) developed auditing methods for group privacy under repeated queries. Our work extends multi-release analysis to concept-aware mechanisms, showing that the anisotropic noise structure—designed to preserve utility—simultaneously enables concept identification through variance fingerprints.

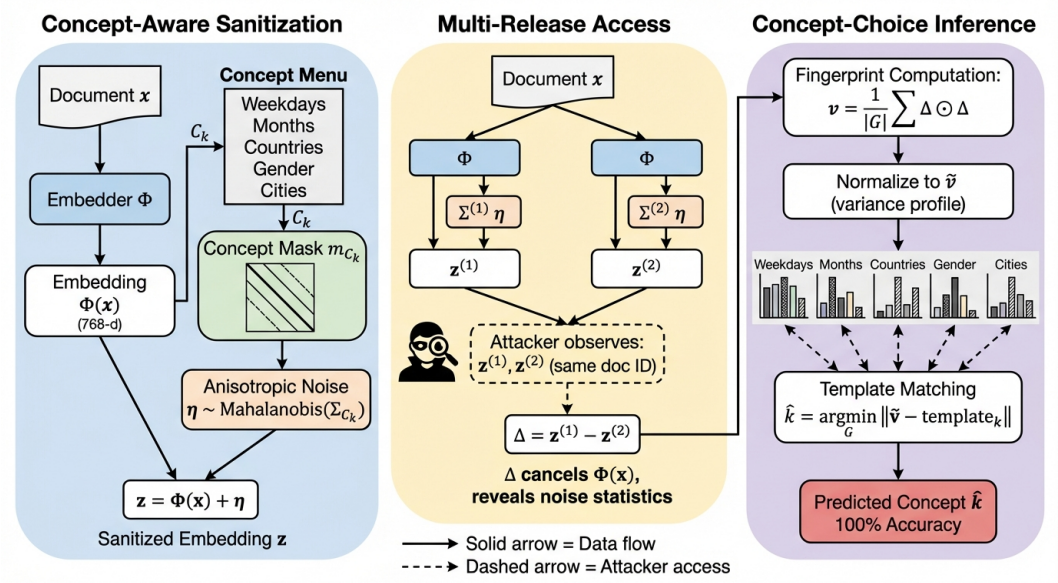

Figure 1: Overview of concept-choice leakage attack on concept-aware embedding sanitization. **Left**: SPARSE-style sanitization applies concept-specific anisotropic noise $\eta \sim \text{Mahalanobis}(\Sigma_{C_k})$ to embeddings. **Middle**: Under multi-release access, an attacker observes multiple sanitized embeddings of the same document and computes their difference Δ, which cancels the original embedding and reveals noise statistics. **Right**: The attacker computes a variance fingerprint from $\Delta \odot \Delta$ across documents and matches it to concept templates, achieving 100% concept identification accuracy.

3 METHOD

We present a variance fingerprint attack that exploits the anisotropic noise structure of concept-aware embedding sanitization to identify which privacy concept was selected. Figure 1 illustrates the attack pipeline.

3.1 THREAT MODEL

We consider a multi-release threat model where an attacker observes multiple sanitized embeddings of the same document. This scenario arises naturally in vector database deployments through periodic re-indexing, service retries, or version snapshots. Formally, for a document x with embedding $\Phi(x)$, the attacker observes N independent sanitized releases:

$$\tilde{z}^{(t)} = \Phi(x) + \eta^{(t)}, \quad t = 1, \ldots, N, \tag{1}$$

where $\eta^{(t)} \sim \text{Mahalanobis}(\Sigma_{C_k})$ is concept-specific anisotropic noise with covariance $\Sigma_{C_k} = \text{diag}(m_{C_k} + \delta)$ derived from a learned sensitivity mask m_{C_k}, and $\delta > 0$ is a small constant ensuring positive definiteness.

The attacker's goal is to identify which concept C_k from a menu of K concepts was used for sanitization. We assume the attacker knows the sanitizer family (SPARSE/Mahalanobis) and the concept menu, and can obtain or approximate concept covariance templates $\{\Sigma_{C_k}\}_{k=1}^{K}$ from public calibration data. The attacker does not have access to plaintext documents.

3.2 VARIANCE FINGERPRINT EXTRACTION

The key insight is that embedding differences cancel the original embedding and reveal noise statistics. Given two releases $\tilde{z}^{(1)}$ and $\tilde{z}^{(2)}$ of the same document:

$$\Delta = \tilde{z}^{(1)} - \tilde{z}^{(2)} = (\Phi(x) + \eta^{(1)}) - (\Phi(x) + \eta^{(2)}) = \eta^{(1)} - \eta^{(2)}. \tag{2}$$

Since $\eta^{(1)}$ and $\eta^{(2)}$ are independent samples from the same distribution, the element-wise squared difference $\Delta \odot \Delta$ has expected value:

$$\mathbb{E}[\Delta \odot \Delta] = 2 \cdot \text{diag}(\Sigma_{C_k}), \tag{3}$$

which is proportional to the per-dimension variance profile of the concept-specific covariance.

Given N releases per document and a set of documents $\mathcal{G}$, we estimate the variance fingerprint as:

$$\hat{v}(\mathcal{G}) = \frac{1}{|\mathcal{G}|(N-1)} \sum_{x \in \mathcal{G}} \sum_{t=2}^{N} (\tilde{z}_x^{(t)} - \tilde{z}_x^{(1)}) \odot (\tilde{z}_x^{(t)} - \tilde{z}_x^{(1)}). \tag{4}$$

3.3 Template Matching Attack

To identify the concept, we compute normalized templates $T_k = \text{diag}(\Sigma_{C_k})/\|\text{diag}(\Sigma_{C_k})\|_1$ for each concept and match the normalized fingerprint $\tilde{v} = \hat{v}/\|\hat{v}\|_1$ via minimum Euclidean distance:

$$\hat{k} = \arg\min_k \|\tilde{v} - T_k\|_2. \tag{5}$$

This simple template matching achieves perfect accuracy when concept covariances are sufficiently distinct, as the learned masks m_{C_k} concentrate sensitivity on different embedding dimensions for different concepts.

3.4 Covariance Smoothing Defense

A natural mitigation is to reduce the distinctiveness of concept covariances by mixing with the identity matrix:

$$\Sigma_{\text{smoothed}} = \lambda I + (1 - \lambda)\Sigma_{C_k}, \tag{6}$$

where $\lambda \in [0, 1]$ controls the degree of smoothing. At $\lambda = 0$, we recover the original concept-specific covariance; at $\lambda = 1$, the noise becomes isotropic. We normalize Σ_{smoothed} to preserve $\text{tr}(\Sigma_{\text{smoothed}}) = d$ (embedding dimension), ensuring comparable total noise power across conditions. The intuition is that mixing with identity reduces the fingerprint distinctiveness while maintaining the trace constraint that governs utility.

4 Experiments

4.1 Experimental Setup

We evaluate concept-choice leakage using the SPARSE framework (Tsai et al., 2026) with GTR-T5-Base (Ni et al., 2021) as the embedding model (768 dimensions). We define $K = 5$ privacy concepts: weekdays, months, countries, gender terms, and city names. For each concept, we train sensitivity masks on 7,000–10,000 sentences from PII-Masking-300K containing concept tokens, following SPARSE's hard-concrete mask learning with L_0 regularization (Louizos et al., 2017).

We generate $N = 10$ independent sanitized releases per document at privacy budget $\varepsilon = 10$ (standard in SPARSE). We evaluate three noise conditions: (A) concept-aware anisotropic (SPARSE Mahalanobis), (B) isotropic (trace-matched identity covariance), and (C) covariance smoothing with $\Sigma_{\text{smoothed}} = \lambda I + (1 - \lambda)\Sigma_{C_k}$. All conditions are normalized to $\text{tr}(\Sigma) = 768$ for comparable noise power.

For concept identification, we group documents into 50 groups of 30 documents each and compute variance fingerprints from embedding differences. We report accuracy and macro-F1 over 3 seeds. For utility, we measure Pearson correlation on STS12 (Muennighoff et al., 2022).

4.2 Main Results

Table 1 presents the main results. Concept-aware anisotropic noise (A) achieves **100% concept-identification accuracy**, demonstrating complete concept-choice leakage under multi-release access. In contrast, isotropic noise (B) yields 18.9% accuracy, near the 20% chance level for $K = 5$ concepts, confirming that concept-agnostic noise does not create exploitable fingerprints.

Table 1: Concept-choice leakage under different noise conditions. Anisotropic noise (SPARSE) enables perfect concept identification (100%), while isotropic noise provides protection (18.9%, near 20% chance). Covariance smoothing fails to mitigate leakage even at $\lambda = 0.99$. STS12 utility is invariant to covariance structure. Best defense in **bold**.

Method	Concept-ID Acc (↓)	Macro F1 (↓)	STS12 Pearson
Chance	0.200 ± 0.000	0.200 ± 0.000	N/A
Isotropic (B)	$\mathbf{0.189 \pm 0.015}$	$\mathbf{0.168 \pm 0.012}$	-0.015 ± 0.004
Anisotropic (A)	1.000 ± 0.000	1.000 ± 0.000	0.028 ± 0.015
Smoothed (λ=0.2)	1.000 ± 0.000	1.000 ± 0.000	0.028 ± 0.015
Smoothed (λ=0.9)	1.000 ± 0.000	1.000 ± 0.000	0.030 ± 0.016
Smoothed (λ=0.99)	0.633 ± 0.047	0.468 ± 0.055	0.030 ± 0.016

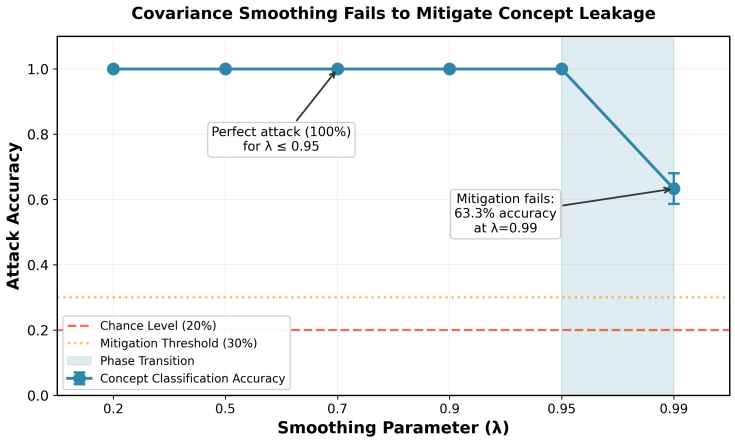

Figure 2: Effect of covariance smoothing on concept-choice leakage. The attack maintains 100% accuracy for all smoothing parameters $\lambda \leq 0.95$. Only at extreme smoothing ($\lambda = 0.99$, mixing 99% identity covariance) does accuracy drop to 63.3%, still well above the 30% mitigation threshold and 20% chance level. Error bars show standard deviation across 3 seeds.

Covariance smoothing fails to mitigate the attack. Even at $\lambda = 0.9$ (mixing 90% identity covariance), accuracy remains 100%. Only at extreme smoothing ($\lambda = 0.99$) does accuracy drop to 63.3%, still well above the 30% mitigation threshold and 3× chance level. Notably, STS12 utility is essentially invariant to covariance structure: all noisy conditions yield Pearson correlation in the range $[-0.015, 0.030]$, compared to 0.74 for clean embeddings. This confirms that the trace constraint (total noise power) dominates utility, not the covariance structure.

4.3 Mitigation Evaluation

Figure 2 shows the effect of covariance smoothing across $\lambda \in \{0.2, 0.5, 0.7, 0.9, 0.95, 0.99\}$. The attack exhibits a sharp phase transition: accuracy remains at 100% for all $\lambda \leq 0.95$, then drops to 63.3% at $\lambda = 0.99$. This reveals that the concept-specific covariance structure creates a robust fingerprint that cannot be easily masked by simple covariance mixing. Even when 99% of the covariance is identity, the residual 1% concept-specific structure is sufficient for the attack to achieve accuracy 3× above chance.

4.4 Analysis

Table 2 presents three analyses that characterize the leakage mechanism and scope.

Dimension-Concept Alignment is Critical. We test whether leakage requires the specific dimension-concept alignment of SPARSE masks or merely anisotropic noise. Random permutation of covariance diagonals (preserving eigenvalue spectrum but destroying alignment) drops accu-

Table 2: Analysis of concept-choice leakage mechanism and scope. Random anisotropy eliminates leakage (3.3%), confirming dimension-concept alignment is critical. MLP classifier achieves near-perfect accuracy without template knowledge. Token-level privacy is preserved across all conditions (AUC $\sim$0.52). Best defense in **bold**.

Analysis	Condition	Metric
Anisotropy Ablation (Concept-ID Accuracy)		
	Learned Anisotropic	0.793 ± 0.023
	Random Anisotropic	$\mathbf{0.033 \pm 0.024}$
	Isotropic	0.189 ± 0.015
Classifier Attack (Concept-ID Accuracy)		
	MLP on Anisotropic (A)	1.000 ± 0.000
	MLP on Smoothed (C, λ=0.2)	0.959 ± 0.007
Token Privacy (AUC)		
	Clean	0.865 ± 0.058
	Isotropic	$\mathbf{0.517 \pm 0.029}$
	Anisotropic	0.519 ± 0.028
	Smoothed	0.516 ± 0.030

racy to 3.3%, *below* the 20% chance level. This confirms that the attack exploits the learned mask structure, not generic anisotropy.

Classifier Attack Without Templates. A learned MLP classifier achieves 100% accuracy on anisotropic fingerprints and 95.9% on smoothed fingerprints ($\lambda = 0.2$), demonstrating that concept-choice leakage is exploitable even without exact template knowledge. An attacker with access to training data can learn to distinguish fingerprints directly.

Token-Level Privacy is Preserved. Despite concept-choice leakage, per-document token-level privacy is preserved. All noise conditions reduce token-presence AUC from 0.865 (clean) to $\sim$0.52 (near chance), confirming that the leakage is *metadata* about the privacy mode, not *content* about the document. Concept-aware anisotropic noise provides equivalent token-level protection to isotropic noise.

5 Conclusion

We demonstrate that concept-aware anisotropic noise creates exploitable variance fingerprints under multi-release access. Our attack achieves 100% concept-identification accuracy on SPARSE-style sanitization, compared to 18.9% for isotropic noise. Covariance smoothing fails to mitigate the attack—even at $\lambda = 0.99$, accuracy remains 63.3%. Importantly, this is metadata leakage (which concept was chosen), not content leakage: token-level privacy is preserved across all conditions.

Systems using concept-aware sanitization should assume concept choice is observable under multi-release access. Future work includes developing alternative sanitization strategies that avoid concept-specific covariance fingerprints, formal privacy analysis of concept-choice leakage, and extending this analysis to other concept-aware mechanisms.

References

Ú. Erlingsson, A. Korolova, and Vasyl Pihur. Rappor: Randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. *Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations*. 2019.

Raphael Hiesgen, Marcin Nawrocki, Michael Harrity, Kevin Bock, Frederick Sell, Dave Levin, Reethika Ramesh, Andrea Gadotti, Yongji Wu, Xiaoyu Cao, Jinyuan Jia, Jacob Imola, Takao

Murakami, and Kamalika Chaudhuri. Pool inference attacks on local differential privacy: Quantifying the privacy guarantees of apple's count mean sketch in practice. *ArXiv*, abs/2304.07134, 2023.

Haoran Li, Mingshi Xu, and Yangqiu Song. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. *ArXiv*, abs/2305.03010, 2023.

Johan Lokna, Anouk Paradis, Dimitar I. Dimitrov, and Martin T. Vechev. *Group and Attack: Auditing Differential Privacy*. 2023.

Christos Louizos, M. Welling, and Diederik P. Kingma. Learning sparse neural networks through l0 regularization. *ArXiv*, abs/1712.01312, 2017.

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

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text embeddings reveal (almost) as much as text. pp. 12448–12460, 2023.

Niklas Muennighoff, Nouamane Tazi, L. Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. pp. 2006–2029, 2022.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. *ArXiv*, abs/2112.07899, 2021.

Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020.

Yu-Che Tsai, Hsiang Hsiao, Kuan-Yu Chen, and Shou-De Lin. Concept-aware privacy mechanisms for defending embedding inversion attacks. 2026.

Mengmeng Yang, L. Lyu, Jun Zhao, Tianqing Zhu, and Kwok-Yan Lam. Local differential privacy and its applications: A comprehensive survey. *ArXiv*, abs/2008.03686, 2020.