8-BIT QUANTIZATION PROVIDES NO PRIVACY BENEFIT AGAINST TRAINING-FREE EMBEDDING INVERSION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Text embeddings enable efficient retrieval but leak private information through inversion attacks that reconstruct original text. Quantization is widely deployed for storage efficiency—does it also provide privacy protection? We present the first utility-matched evaluation of 8-bit quantization against ZSInvert, a training-free inversion attack. By calibrating Gaussian noise to match quantization's retrieval utility (nDCG@10 $\approx$ 0.544), we isolate privacy effects from utility differences. Our key finding is surprising: quantization provides negligible privacy benefit, achieving only 6% relative reduction in attribute recovery (Canary-EM: 6.0% vs 6.4%) compared to 70% reduction for noise (1.9%). Geometric analysis reveals the mechanism: quantization preserves the cosine similarity structure that ZSInvert exploits (0.9° angular deviation, pairwise ρ=0.9999), while noise disrupts it (41.5° deviation, ρ=0.656). Practitioners should not rely on quantization for privacy; efficiency and privacy require separate, explicit mechanisms.

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, powering retrieval-augmented generation (RAG), semantic search, and recommendation engines. These embeddings are routinely stored in third-party vector databases, often containing representations of sensitive documents—medical records, legal contracts, corporate communications. While embeddings are commonly treated as safer artifacts than raw text, recent work demonstrates that they leak substantial private information through inversion attacks that reconstruct original text from vector representations (Morris et al., 2023; Li et al., 2023).

Training-free inversion attacks pose a particularly concerning threat. Unlike trained attacks such as Vec2Text (Morris et al., 2023) that require fitting a decoder to the target encoder, training-free methods like ZSInvert (Zhang et al., 2025) use large language models with cosine-similarity-guided beam search to invert embeddings without any target-specific training. This makes them accessible to adversaries without computational resources for model training and applicable to any encoder without prior preparation.

Embedding quantization (e.g., 8-bit integer representation) is widely deployed for storage and computational efficiency. A natural question arises: does the information loss from quantization also degrade inversion attacks, providing "free" privacy as a side effect of efficiency optimization? Prior work suggests quantization reduces lexical reconstruction scores against trained attacks (Seputis et al., 2025; Zhuang et al., 2024), but these evaluations do not control for utility differences and do not test against training-free attacks. The privacy effect of quantization against modern training-free inversion remains unknown.

We address this gap with the first utility-matched evaluation of quantization against training-free embedding inversion. By calibrating a Gaussian noise baseline to match quantization's retrieval

[1] https://gitlab.com/fars-a/quantized-embeddings-trainingfree-inversion

utility, we isolate the privacy effect from confounding utility differences. Our key finding is surprising: **8-bit quantization provides no meaningful privacy benefit**. At matched utility, quantization achieves only a 6% relative reduction in attribute recovery compared to 70% for noise. Geometric analysis reveals the mechanism: quantization preserves the cosine similarity structure (0.9° angular deviation) that ZSInvert exploits, while noise disrupts it (41.5° deviation).

Our contributions are:

- The first utility-matched evaluation of embedding quantization against training-free inversion attacks, enabling fair comparison of privacy benefits.
- A surprising negative result: 8-bit quantization provides negligible privacy protection (6% reduction) while utility-matched noise achieves 70% reduction.
- A geometric explanation: quantization preserves cosine similarity structure (ρ=0.9999) that attacks exploit, while noise disrupts it (ρ=0.656).
- A practical implication: efficiency optimizations and privacy defenses require separate, explicit mechanisms—practitioners should not rely on quantization for privacy.

2 Related Work

Embedding Inversion Attacks. Text embedding inversion attacks aim to reconstruct original text from dense vector representations. Morris et al. (2023) introduced Vec2Text, which trains a decoder on the target encoder's embedding space to iteratively reconstruct input text, demonstrating that embeddings reveal almost as much information as the original text. Li et al. (2023) proposed GEIA, a generative approach that leverages language models to recover sentences from embeddings. More recently, training-free attacks have emerged as practical threats: Zhang et al. (2025) developed ZSInvert, which combines large language model generation with cosine-similarity-guided beam search to invert embeddings without any target-specific training. Kim et al. (2026) extended this paradigm to cross-domain settings with Zero2Text. Huang et al. (2024) demonstrated transferable attacks that work across different embedding models. Training-free attacks are particularly concerning because they require no access to training data or computational resources for model fitting, making them accessible to a broader range of adversaries.

Privacy Defenses. Several defense mechanisms have been proposed to protect embeddings from inversion attacks. Zhuang et al. (2024) analyzed the threat of Vec2Text to dense retrieval systems and evaluated noise addition as a defense. Differential privacy approaches have been explored by Mai et al. (2023), who proposed Split-and-Denoise for protecting LLM inference, and Zein & Henderson (2026), who developed variational information bottleneck methods for private embeddings. Learned obfuscation techniques include the Stained Glass Transform (Roberts et al., 2025) and BeamClean (Kale et al., 2025). Tsai et al. (2026) proposed concept-aware privacy mechanisms, while Liu et al. (2024) introduced Eguard based on mutual information optimization. Cao et al. (2026) addressed the privacy-utility-efficiency trilemma through obfuscated semantic null space projection. However, most existing defenses have been evaluated primarily against trained attacks like Vec2Text, leaving their effectiveness against training-free attacks underexplored.

Quantization and Privacy. Quantization is widely deployed for reducing storage and computational costs of embedding systems. While prior work has examined quantization's effect on retrieval utility, its impact on privacy against inversion attacks remains understudied. Existing evaluations either focus on trained attacks or do not control for utility differences between defense conditions. Our work addresses this gap by providing the first utility-matched evaluation of quantization against training-free inversion attacks, enabling fair comparison of privacy benefits across defense mechanisms.

3 Method

We design a controlled experiment to evaluate whether 8-bit quantization provides privacy protection against training-free embedding inversion. Figure 1 illustrates our experimental framework.

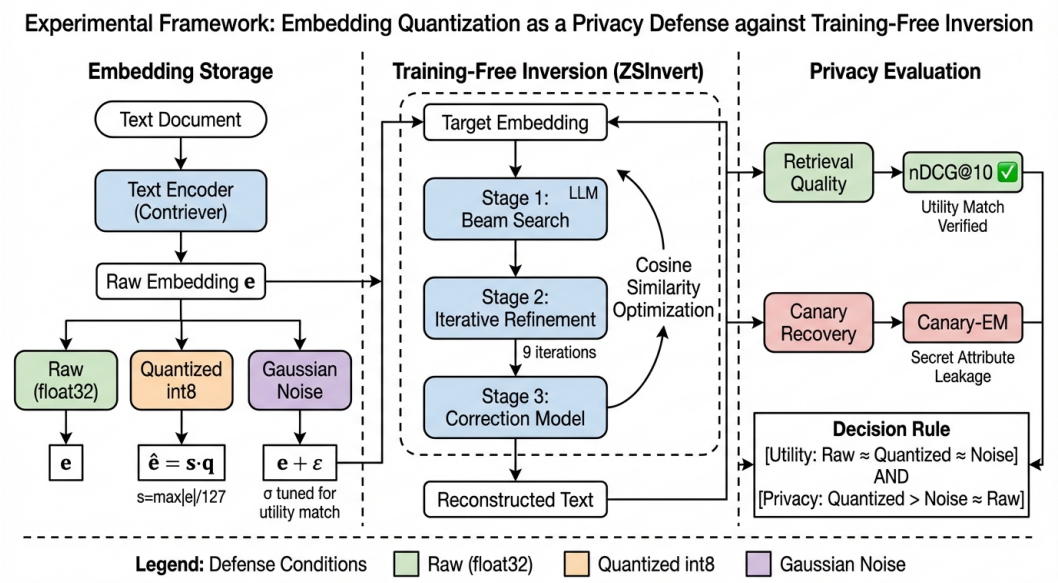

Figure 1: Experimental framework for evaluating quantization as a privacy defense. Three embedding conditions (raw, 8-bit quantized, utility-matched noise) are evaluated against ZSInvert, a training-free inversion attack. Privacy is measured via Canary-EM on synthetic attribute documents; utility via nDCG@10 on SciFact retrieval.

3.1 Experimental Framework

We evaluate three embedding storage conditions: (1) **Raw (float32)**: undefended baseline embeddings; (2) **Quantized (int8)**: 8-bit absmax scalar quantization, where each embedding e is stored as $q = \text{round}(e/s)$ with scale $s = \max_i |e_i|/127$, and dequantized as $\hat{e} = sq$ for similarity computation; (3) **Noise (σ=0.05)**: Gaussian perturbation $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ added to embeddings, with σ calibrated to match the quantized condition's retrieval utility.

The utility-matching design is critical: by ensuring all conditions achieve equivalent retrieval performance, we isolate the privacy effect of each defense from confounding utility differences.

3.2 Attack: ZSInvert

We use ZSInvert (Zhang et al., 2025), a training-free inversion attack that reconstructs text from embeddings without requiring target-specific training. ZSInvert operates in three stages: (1) an LLM generates candidate texts using beam search with a generic prompt; (2) iterative refinement selects candidates maximizing cosine similarity to the target embedding; (3) a correction model post-processes outputs. We use Qwen2.5-7B-Instruct (Yang et al., 2024) as the generator with beam width 30, 9 iterations, and maximum 32 tokens per candidate.

The key insight is that ZSInvert's optimization objective is cosine similarity: any effective defense must disrupt this similarity structure. Quantization, designed to preserve similarity for retrieval, may therefore fail to degrade the attack.

3.3 Evaluation Metrics

Privacy: Canary-EM. We measure privacy leakage using a controlled secret-attribute recovery task. For 500 corpus documents, we prepend a canary sentence: "After years of work, she moved back to the <ADJ> <NOUN> harbor," where <ADJ> and <NOUN> are sampled from disjoint 512-word vocabularies. Canary-EM (exact match) measures the fraction of inversions that recover both attribute words in order. This metric is paraphrase-robust and automatically verifiable, avoiding the ambiguity of lexical overlap metrics like BLEU.

Table 1: Privacy-utility comparison across embedding conditions. Canary-EM measures exact attribute recovery (lower = more private). All conditions maintain matched retrieval utility (nDCG@10 $\approx$ 0.544). Best privacy (lowest Canary-EM) in **bold**. Quantization provides negligible privacy benefit (6% relative reduction) while noise achieves 70% reduction.

Condition	nDCG@10	Canary-EM ↓	Token-F1 ↓	CosSim	Δ vs Raw
Raw (float32)	0.5444	0.064 ± 0.010	0.223 ± 0.003	0.731	—
Quantized (int8)	0.5434	0.060 ± 0.004	0.218 ± 0.002	0.730	-6%
Noise (σ=0.05)	0.5441	$\mathbf{0.019 \pm 0.003}$	$\mathbf{0.160 \pm 0.004}$	0.584	**−70%**

Utility: nDCG@10. We measure retrieval quality using normalized discounted cumulative gain at rank 10 on the SciFact (Wadden et al., 2020) benchmark from BEIR (Thakur et al., 2021). Higher nDCG@10 indicates better retrieval performance.

3.4 Utility Matching

To enable fair privacy comparison, we calibrate the noise level σ to match the quantized condition's retrieval utility. We perform grid search over $\sigma \in [0.001, 0.1]$ and select $\sigma = 0.05$, which achieves nDCG@10 = 0.5441 compared to 0.5434 for quantization (difference: 0.0007). This ensures that any privacy difference between conditions cannot be attributed to utility degradation.

4 Experiments

4.1 Setup

We use Contriever (Izacard et al., 2021) as the embedding encoder (768 dimensions, mean pooling). The retrieval benchmark is SciFact (Wadden et al., 2020) from BEIR (Thakur et al., 2021), containing 5,183 corpus documents and 300 test queries. For privacy evaluation, we sample 500 documents and prepend canary sentences with attributes from disjoint 512-word vocabularies. We run ZSInvert with beam width 30, 9 iterations, and maximum 32 tokens, using three random seeds (42, 123, 456) and report mean $\pm$ standard deviation.

4.2 Main Results

Table 1 presents the privacy-utility comparison across embedding conditions. The key finding is striking: **quantization provides negligible privacy benefit**. At matched retrieval utility (nDCG@10 $\approx$ 0.544), quantized embeddings yield Canary-EM of 6.0% $\pm$ 0.4% compared to 6.4% $\pm$ 1.0% for raw embeddings—a mere 6% relative reduction that is not statistically significant.

In contrast, Gaussian noise at matched utility achieves a 70% relative reduction in Canary-EM (1.9% $\pm$ 0.3% vs 6.4% $\pm$ 1.0%). This demonstrates that meaningful privacy protection is achievable without sacrificing retrieval quality—but quantization does not provide it. The Token-F1 metric shows a similar pattern: noise reduces partial token overlap by 28% while quantization reduces it by only 2%.

4.3 Geometric Analysis

Why does quantization fail as a privacy defense? Table 2 reveals the mechanism: quantization preserves the cosine similarity geometry that ZSInvert exploits, while noise disrupts it.

Quantization causes only 0.9° mean angular deviation from original embeddings, while noise causes 41.5°—a 46$\times$ difference. More critically, quantization preserves pairwise similarity structure almost perfectly (Spearman $\rho = 0.9999$), while noise substantially disrupts it ($\rho = 0.656$). Figure 2 visualizes these geometric differences.

ZSInvert's optimization objective is to maximize cosine similarity between reconstructed text embeddings and target embeddings. When quantization preserves this similarity landscape almost per-

Table 2: Geometric properties of embedding perturbations. Quantization preserves cosine similarity structure almost perfectly (0.9° deviation, ρ=0.9999), while noise substantially disrupts it (41.5° deviation, ρ=0.656). This explains why ZSInvert succeeds against quantization but fails against noise.

Perturbation	L2 Norm	Angular Dev. (°)	Pairwise ρ	Pairwise r
Quantization	0.025 ± 0.002	0.90 ± 0.09	0.9999	0.9999
Noise (σ=0.05)	1.383 ± 0.034	41.5 ± 1.6	0.656	0.676

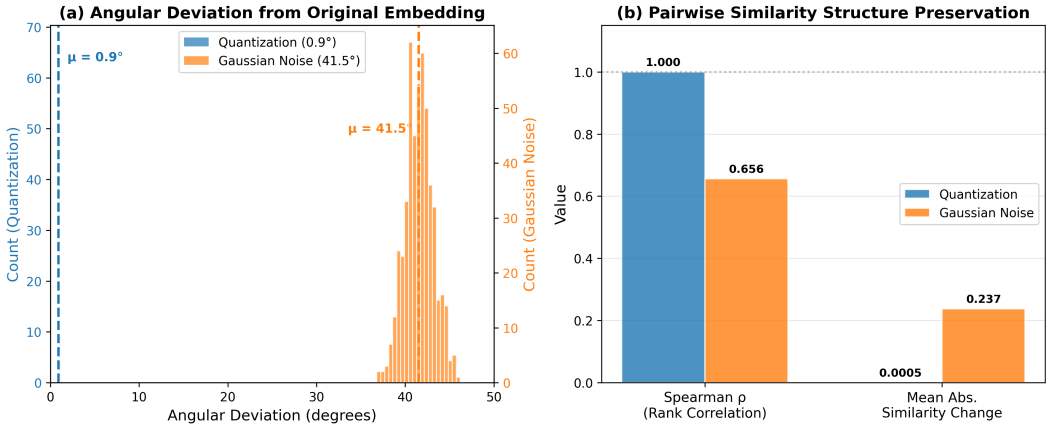

Figure 2: Geometric analysis of perturbation effects. Left: Angular deviation from original embeddings—quantization causes only 0.9° deviation while noise causes 41.5°. Right: Pairwise cosine similarity preservation—quantization maintains near-perfect rank correlation (ρ=0.9999) while noise substantially disrupts similarity structure (ρ=0.656).

fectly, the attack's search succeeds. When noise disrupts the landscape, the attack degrades proportionally.

4.4 Stage 3 Ablation

We ablate ZSInvert's Stage 3 correction model to understand which component drives attribute recovery. The correction model contributes zero additional exact matches across all conditions: Canary-EM is identical whether using the full pipeline or Stage 2 alone. The cosine-similarity-guided beam search (Stage 2) accounts for 100% of attribute recovery. This confirms that the attack's success depends entirely on the preserved similarity structure, which quantization maintains but noise disrupts.

5 Conclusion

We demonstrate that 8-bit quantization provides no meaningful privacy protection against training-free embedding inversion. At matched retrieval utility, quantization achieves only a 6% relative reduction in attribute recovery compared to 70% for Gaussian noise. The mechanism is clear: quantization preserves the cosine similarity geometry (0.9° angular deviation, ρ=0.9999) that ZSInvert exploits, while noise disrupts it (41.5° deviation, ρ=0.656).

The practical implication is direct: practitioners should not rely on quantization for privacy protection. Efficiency optimizations and privacy defenses require separate, explicit mechanisms. Simple noise addition at matched utility provides substantial protection, suggesting that effective defenses need not sacrifice retrieval quality.

Limitations. Our evaluation uses a single encoder (Contriever), dataset (SciFact), quantization scheme (absmax int8), and attack (ZSInvert). Future work should evaluate generalization across encoders, quantization methods, and emerging attacks.

REFERENCES

Zhiyuan Cao, Zeyu Ma, Chenhao Yang, Han Zheng, and Mingang Chen. Osnip: Breaking the privacy-utility-efficiency trilemma in llm inference via obfuscated semantic null space. 2026.

Yu-Hsiang Huang, Yu-Che Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. Transferable embedding inversion attack: Uncovering privacy risks in text embeddings without model queries. pp. 4193–4205, 2024.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022, 2021.

Kaan Kale, K. Mylonakis, Jay Roberts, and Sidhartha Roy. Beamclean: Language aware embedding reconstruction. *ArXiv*, abs/2505.13758, 2025.

Doohyun Kim, Donghwa Kang, Kyungjae Lee, Hyeongboo Baek, and B. Kang. Zero2text: Zero-training cross-domain inversion attacks on textual embeddings. 2026.

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.

Tiantian Liu, Hongwei Yao, Feng Lin, Tong Wu, Zhan Qin, and Kui Ren. Eguard: Defending llm embeddings against inversion attacks via text mutual information optimization. 2024.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-denoise: Protect large language model inference with local differential privacy. pp. 34281–34302, 2023.

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

Jay Roberts, K. Mylonakis, Sidhartha Roy, and Kaan Kale. Learning obfuscations of llm embedding sequences: Stained glass transform. *ArXiv*, abs/2506.09452, 2025.

Dominykas Seputis, Yongkang Li, Karsten Langerak, and Serghei Mihailov. *Rethinking the Privacy of Text Embeddings: A Reproducibility Study of "Text Embeddings Reveal (Almost) As Much As Text"*. 2025.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663, 2021.

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

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *ArXiv*, abs/2004.14974, 2020.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Dina El Zein and James Henderson. Differential privacy for transformer embeddings of text with nonparametric variational information bottleneck. *ArXiv*, abs/2601.02307, 2026.

Collin Zhang, John X. Morris, and Vitaly Shmatikov. Universal zero-shot embedding inversion. *ArXiv*, abs/2504.00147, 2025.

Shengyao Zhuang, B. Koopman, Xiaoran Chu, and G. Zuccon. Understanding and mitigating the threat of vec2text to dense retrieval systems. *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 2024.