Transform-Domain Semantic Databases: Representation-Level Confidentiality and Completeness-Gated Search

Jonathan H. West
Independent Researcher
jw@multiverse1.com

Version 1.1 — 2026

Abstract

Large-scale semantic databases store high-dimensional vector embeddings that enable similarity search across biomedical, legal, and industrial information systems. While encryption protects data at rest and in transit, embeddings are typically processed in plaintext during similarity computation, creating confidentiality risks under runtime compromise or partial exfiltration.

We introduce a transform-domain semantic storage architecture in which embeddings are stored as quantised coefficients of a global approximately orthogonal transform. Similarity search is performed directly in this domain, and reconstruction occurs only for selected candidates. Because semantic information is distributed across the full coefficient set, partial coefficient exposure yields limited semantic utility.

We formalise this model, define a threat framework focused on partial runtime compromise, analyse similarity preservation properties, and introduce the concept of *representation-level confidentiality* together with *completeness-gated exfiltration resistance*. Unlike cryptographic approaches such as fully homomorphic encryption or secure enclaves, this method does not aim to provide cryptographic secrecy, but instead reduces semantic interpretability under partial data exposure while preserving computational efficiency.

We argue that transform-domain semantic storage provides a practical confidentiality layer complementary to encryption in large-scale semantic search systems, particularly in sensitive multi-user environments such as biomedical and medico-legal databases.

Keywords: embeddings; vector databases; similarity search; transform-domain storage; confidentiality; governance; sharding.

1 Introduction

Vector embeddings underpin modern semantic information systems, including biomedical similarity search, genomic clustering, legal document retrieval, recommendation systems, and large-scale AI-assisted analytics. Such systems store embeddings in vector databases and perform similarity search via inner products or distance metrics [3, 5].

Although encryption protects data at rest and in transit, embeddings must typically be decrypted during runtime similarity computation. Consequently, compromise of runtime access—via credential misuse, insider activity, or misconfiguration—may expose semantically interpretable vectors at scale.

Cryptographic solutions such as fully homomorphic encryption (FHE) and trusted execution environments (TEEs) provide stronger protection but often impose substantial computational or operational overhead [1, 2].

This paper explores a complementary architectural approach: *transform-domain semantic storage with completeness-gated reconstruction.* Rather than storing embeddings directly, we:

1. apply a global approximately orthogonal transform,
2. quantise the transform coefficients,
3. perform similarity search entirely in transform space, and
4. reconstruct only selected results.

Because semantic structure is dispersed across the full coefficient space, partial exposure yields limited interpretability. We formalise this property as *representation-level confidentiality.*

Contributions.

1. Formal definition of transform-domain semantic storage.
2. Similarity preservation analysis under approximate orthogonality.
3. Definition of representation-level confidentiality.
4. Introduction of completeness-gated exfiltration resistance.
5. Governance model for multi-authority reconstruction control.

2 Related Work

2.1 Vector search and ANN systems

Approximate nearest neighbour (ANN) search systems such as FAISS and HNSW enable efficient similarity search in high-dimensional spaces [3, 5]. These systems optimise performance but do not inherently address semantic confidentiality under runtime compromise.

2.2 Cryptographic protection of vector computation

Fully homomorphic encryption allows computation over encrypted data, but remains computationally intensive for high-dimensional similarity search at production scale [2]. Trusted execution environments (e.g., Intel SGX) offer enclave-based runtime protection, but introduce hardware dependencies and potential side-channel concerns [1].

2.3 Transform coding and orthogonal representations

Orthogonal transforms such as the discrete cosine transform (DCT), Hadamard transforms, and random orthogonal matrices are widely used in signal compression and dimensionality manipulation. The Johnson–Lindenstrauss lemma establishes that certain random linear maps approximately preserve pairwise distances under dimensional projection [4]. Our approach differs in that the transform is not used primarily for compression or projection, but for *information dispersion across coefficient space while preserving similarity structure.*

2.4 Threshold and secret sharing

Threshold cryptographic schemes such as Shamir secret sharing require multiple shares for reconstruction [6]. Our model is not cryptographic secret sharing; instead, it introduces an architectural analogue in which semantic utility depends on near-complete coefficient coverage.

3 Threat Model

We assume:

- encryption at rest and in transit is present,
- runtime similarity search operates on decrypted data,
- an adversary may obtain partial database contents via:
  - compromised credentials,
  - insider export,
  - misconfigured storage, or
  - partial shard access.

We do *not* assume:

- cryptographic secrecy of the transform matrix T, nor
- resistance to full database compromise with reconstruction access.

Our objective is not cryptographic secrecy but *reduction of semantic utility under partial exposure.* This addresses a gap in conventional threat models: runtime access to decrypted vectors is typically treated as an endpoint rather than a surface requiring further architectural mitigation.

4 Transform-Domain Representation

Let $x \in \mathbb{R}^n$ denote an embedding. Let $T \in \mathbb{R}^{n \times n}$ be an approximately orthogonal transform satisfying

$$T^\top T \approx \mathbf{I}. \tag{1}$$

Candidate transforms include structured Hadamard transforms, random orthogonal matrices, and fast structured transforms used in signal processing. The choice of T may be system-specific; crucially, cryptographic secrecy of T is not required for the confidentiality properties described below.

Let $Q(\cdot)$ denote a quantisation operator mapping real-valued coefficients to a reduced-precision representation (e.g., INT8). We define the stored representation:

$$S(x) = Q(Tx). \tag{2}$$

Reconstruction is defined as:

$$\hat{x} = T^{-1}\Big(Q^{-1}(S(x))\Big). \tag{3}$$

Because T is approximately orthogonal, $T^{-1} \approx T^{\top}$, making reconstruction computationally efficient.

5 Similarity Preservation

For approximately orthogonal T, inner products are approximately preserved:

$$\langle x, y \rangle \approx \langle Tx, Ty \rangle . \tag{4}$$

This follows directly from (1): for unit vectors, $\langle Tx, Ty \rangle = x^{\top}T^{\top}Ty \approx x^{\top}y = \langle x, y \rangle$.

Similarity ranking can therefore be performed directly in transform space without reconstruction:

$$\langle Tq, Tx_i \rangle , \tag{5}$$

preserving computational efficiency comparable to conventional dot-product search. This property is foundational: it allows the system to operate entirely in the transform domain during search, limiting the contexts in which full reconstruction is required.

6 Representation-Level Confidentiality

6.1 Information dispersion

Under a global approximately orthogonal transform, each original coordinate contributes to many—typically all—transform coefficients. Formally, for a dense T, the (j, k)-th entry T_{jk} encodes the contribution of original dimension k to coefficient j. Because no column of T is sparse, no small subset of coefficients carries interpretable information about any individual original dimension.

Consequently, semantic information is *distributed rather than localised* across coefficient space. This contrasts with direct storage, in which individual dimensions may correspond to interpretable semantic features or be directly invertible from small subsets.

6.2 Partial reconstruction and completeness

Let $C \subset \{1, \ldots, n\}$ denote an observed subset of coefficient indices. Define partial reconstruction:

$$\tilde{x}_C = T^{-1}\Big(Q^{-1}(S(x)_C)\Big), \tag{6}$$

where $S(x)_C$ denotes a coefficient vector in which only indices in C are retained (others zeroed or omitted per implementation).

Define the *completeness ratio*:

$$\rho = \frac{|C|}{n}. \tag{7}$$

We hypothesise that semantic utility $U(\rho)$ exhibits threshold-like behaviour: limited interpretability for small ρ, with rapid increase only as $\rho \to 1$. This behaviour follows from global information dispersion: reconstructing meaningful semantic structure requires near-complete coefficient coverage. Empirical characterisation of $U(\rho)$ across representative embedding families is a direction for future work.

Definition (Representation-Level Confidentiality). *A stored representation $S(x)$ provides representation-level confidentiality if $U(\rho)$ remains below a utility threshold U^* for all $\rho < \rho^*$, where ρ^* is a system-defined completeness threshold.*

7 Completeness-Gated Exfiltration Resistance

Because semantic utility depends on high completeness, partial exfiltration below threshold ρ^* yields limited interpretability.

Definition (Completeness-Gated Exfiltration Resistance). *A system exhibits completeness-gated exfiltration resistance if an adversary obtaining $\rho < \rho^*$ of coefficient coverage cannot reconstruct semantically useful representations.*

Practical implications include:

- opportunistic partial database dumps yield negligible semantic utility,
- meaningful exfiltration requires coordinated compromise of multiple storage shards,
- the time and data volume required before utility threshold is reached increases substantially.

This does not eliminate risk, but alters the *economics* of attack: partial access is no longer sufficient for semantic interpretation.

8 Governance-Enforced Completeness Control

The completeness-gating property enables architectural enforcement through governance mechanisms.

8.1 Coefficient sharding

Partition the coefficient index set $\{1, \ldots, n\}$ across k independent storage domains. Each domain holds n/k coefficients. Reconstruction of a single embedding requires access to threshold $t \leq k$ domains, enforced architecturally.

8.2 Reconstruction API control

Direct bulk coefficient export is disallowed. Reconstruction is mediated through a controlled service implementing:

- comprehensive request logging,
- rate limiting per identity and time window,
- per-identity caps on total reconstructions,
- multi-authority approval requirements for bulk extraction.

8.3 Multi-authority thresholding

Bulk extraction—any request that would aggregate sufficient coefficients to approach ρ^*—requires digital approval from multiple independent authorities. This introduces institutional oversight before the completeness threshold can be reached, analogous in structure (though not in mechanism) to threshold cryptographic schemes [6].

8.4 Application to sensitive domains

In multi-user environments such as biomedical databases or medico-legal archives, tiered completeness controls can be mapped onto governance access levels. A researcher granted standard access may receive sufficient coefficient completeness to identify thematic clusters and aggregate patterns, while remaining below the threshold for individual case reconstruction. Bulk reconstruction for population-level analysis would require elevated multi-authority approval. This provides architectural support for proportionate access frameworks without relying solely on procedural controls.

9 Query Confidentiality

A property not addressed by conventional privacy-preserving search architectures is *query confidentiality*: the semantic content of a search query may itself be sensitive. In biomedical or legal retrieval, a query encoding a clinical presentation or legal fact pattern could be identifying.

Because similarity search in the transform domain operates on Tq rather than q directly (equation (5)), the query is presented to the search index in its distributed transform representation. As with stored embeddings, partial observation of Tq yields limited interpretability of the underlying query semantics.

This property does not require query encryption and introduces no homomorphic overhead. It is a natural consequence of operating entirely within the transform domain during search.

10 Comparison with Encryption-Based Approaches

Property	Encryption	FHE	Transform-domain
At-rest protection	✓	✓	partial
In-transit protection	✓	✓	partial
Runtime plaintext exposure	yes	no	distributed
Similarity search overhead	low	very high	low
Partial exfiltration resistance	no	no	yes
Query confidentiality	no	yes	partial
Governance integration	limited	limited	native

Table 1: Comparison of confidentiality approaches for semantic databases.

Encryption protects against disk theft, network interception, and offline attacks. However, decrypted embeddings in runtime memory remain fully interpretable.

Transform-domain storage:

- does not replace encryption,
- reduces semantic utility of partial runtime exposure,
- preserves similarity computation efficiency,
- provides native architectural support for completeness controls, and
- offers partial query confidentiality without cryptographic overhead.

It therefore functions as a complementary confidentiality layer rather than a cryptographic substitute.

11 Performance Considerations

Similarity search remains a standard dot-product operation in transform space (equation (5)). No homomorphic overhead is introduced. Transform application at query time costs $O(n \log n)$ for structured transforms (e.g., Hadamard) or $O(n^2)$ for dense random orthogonal matrices, though in practice fast structured transforms are preferred.

Reconstruction cost scales with the number of selected top-k candidates, not database size. Experimental results on large embedding collections (including SIFT1M vectors) confirm that ANN pipelines operating entirely in the transform domain achieve recall and latency comparable to conventional approaches, while enabling the confidentiality properties described above. Detailed benchmarks are reported in the associated experimental repository [7].

12 Limitations

- The method is not cryptographically secure and should not be treated as a substitute for encryption.
- Full coefficient exposure with knowledge of T permits reconstruction.

- The threshold behaviour of $U(\rho)$ is hypothesised on the basis of information dispersion arguments; empirical characterisation across embedding families is required.
- Effective deployment requires governance infrastructure; architectural properties alone do not enforce access controls.
- The choice of transform T and quantisation scheme affects both similarity preservation and the sharpness of the completeness threshold.

13 Conclusion

Transform-domain semantic storage disperses semantic information across coefficient space while preserving similarity structure. This yields two complementary properties: representation-level confidentiality, in which partial coefficient exposure provides limited semantic utility; and completeness-gated exfiltration resistance, in which meaningful attack requires near-complete coefficient access.

A further consequence is partial query confidentiality: because similarity search operates on transform-domain query representations, the semantic content of queries is not exposed to the search index in plaintext form.

When combined with encryption and governance controls, this architecture reduces the semantic damage potential of runtime compromise without incurring the computational cost of fully homomorphic encrypted search. It provides a practical intermediate confidentiality layer for large-scale semantic databases, and is particularly well suited to sensitive multi-user domains such as biomedical information systems and medico-legal archives, where proportionate access control and query privacy are governance requirements as well as technical desiderata.

Availability

Experimental materials, ANN benchmarks, and related work are available at:

https://github.com/UnrealJon/DTDR

DOI: 10.5281zenodo.18760874

References

[1] Victor Costan and Srinivas Devadas. Intel SGX explained. *IACR Cryptology ePrint Archive*, 2016. Report 2016/086.

[2] Craig Gentry. *A Fully Homomorphic Encryption Scheme*. Stanford University, 2009. PhD thesis.

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. FAISS: A library for efficient similarity search and clustering of dense vectors. In *Proceedings of the International Conference on Big Data*, 2017.

[4] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[5] Yu. A. Malkov and D. A. Yashunin. 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, 2018.

[6] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[7] Jonathan H. West. DTDR: Distributed transform-domain representation — experimental repository. https://github.com/UnrealJon/DTDR, 2026. UK Patent Application No. GB2602157.6.