Transform-Domain Semantic Storage: Dual Utility Regimes, Proportional Confidentiality, and Completeness-Gated Governance

Jonathan H. West
Independent Researcher
jw@multiverse1.com

Version 2.1 – 2026

Abstract

Large-scale semantic systems store high-dimensional embeddings for similarity search across biomedical, legal, and industrial domains. While encryption protects data at rest and in transit, embeddings are typically processed in plaintext during runtime similarity computation, exposing semantic content under partial compromise or insider misuse.

We introduce a transform-domain semantic storage architecture in which embeddings are stored as quantised coefficients of an approximately orthogonal transform. Similarity search is performed entirely in the transform domain, and reconstruction is required only for selected results.

We analytically derive and empirically characterise the relationship between coefficient completeness ratio ρ and semantic utility $U(\rho)$ across two standard benchmarks (GloVe-100d and SIFT1M), demonstrating approximately linear proportional scaling $U(\rho) \approx \rho$ for similarity tasks, consistent across embedding families and dropout modes. This enables quantitative, proportionate access control: restricting access to ρ fraction of coefficients limits retrieval utility to approximately ρ.

Separately, we analyse reconstruction behaviour of transform-domain compressed neural network weights and observe nonlinear emergence behaviour in generative inference. We identify two distinct utility regimes: a linear geometric regime governing similarity preservation, and a nonlinear functional regime governing solvability of deep inference. These phenomena arise from different mathematical properties of their respective tasks.

We argue that transform-domain storage provides a practical confidentiality gradient complementary to encryption, raising the cost of partial exfiltration without homomorphic overhead, and enabling governance-enforced completeness controls in sensitive domains.

1 Introduction

Vector embeddings underpin modern semantic systems including biomedical similarity search, genomic clustering, legal discovery, recommendation engines, and large-scale AI analytics [1, 2].

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) [3] and trusted execution environments (TEEs) [4] provide stronger guarantees but often impose substantial computational or operational overhead.

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,
4. reconstruct only selected outputs.

We show that transform-domain storage induces two distinct utility regimes:

- **Geometric regime:** similarity utility degrades approximately linearly with coefficient completeness, $U(\rho) \approx \rho$.
- **Functional regime:** deep neural inference on transform-domain compressed weights exhibits nonlinear emergence behaviour under incomplete reconstruction.

Recognising this distinction clarifies both the confidentiality properties and their limits.

2 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,
  - partial shard access.

We do not assume:

- Cryptographic secrecy of the transform matrix T.
- Resistance to full compromise with reconstruction access.

Our objective is reduction of semantic utility under partial exposure, not cryptographic secrecy.

3 Transform-Domain Representation

Let $x \in \mathbb{R}^n$ denote an embedding.

Let $T \in \mathbb{R}^{n \times n}$ satisfy:

$$T^\top T \approx I.$$

Stored representation:

$$S(x) = Q(Tx),$$

where Q denotes quantisation (e.g., INT8).

Reconstruction:

$$\hat{x} = T^{-1}(Q^{-1}(S(x))) \approx T^\top Q^{-1}(S(x)).$$

Similarity search is performed as:

$$\langle Tq, Tx_i \rangle,$$

avoiding reconstruction during retrieval.

4 Similarity Preservation and the Geometric Regime

For approximately orthogonal T:

$$\langle x, y \rangle \approx \langle Tx, Ty \rangle.$$

4.1 Analytical Scaling

If a fraction $1 - \rho$ of coefficients is masked via diagonal matrix M_ρ:

$$\langle M_\rho Tx, M_\rho Ty \rangle = (Tx)^\top M_\rho^2 (Ty).$$

For random masking with retention probability ρ:

$$\mathbb{E}[M_\rho^2] = \rho I,$$

thus:

$$\mathbb{E}[\langle M_\rho Tx, M_\rho Ty \rangle] = \rho \langle Tx, Ty \rangle \approx \rho \langle x, y \rangle.$$

For block masking, independence does not strictly hold; however, empirical results demonstrate negligible deviation from the random-mask expectation, indicating that approximate isotropy of transform coefficients yields similar proportional scaling.

Therefore:

$$U(\rho) \approx \rho.$$

This is the geometric regime.

5 Representation-Level Confidentiality

Define completeness ratio:

$$\rho = \frac{|C|}{n},$$

where C is the retained coefficient subset.

Definition (Representation-Level Confidentiality). A representation exhibits representation-level confidentiality if semantic utility is a monotone increasing function of completeness ratio ρ, and proportional restriction of coefficient access results in proportional restriction of retrieval capability.

In the geometric regime, confidentiality operates as a proportional utility gradient rather than a hard secrecy threshold.

6 Completeness-Gated Exfiltration Resistance

Partial exfiltration yields proportional utility loss:

$$U(\rho) \approx \rho.$$

Thus meaningful semantic reconstruction requires high completeness. This raises the cost of attack rather than prohibiting it absolutely.

7 Governance-Enforced Completeness Control

7.1 Coefficient Sharding

Partition coefficients across k domains. Reconstruction requires threshold t shards.

7.2 Reconstruction API Controls

- Logging
- Rate limiting
- Per-identity caps
- Multi-authority approval for bulk extraction

Because $U(\rho)$ is approximately linear, governance bodies can make quantitative access decisions.

8 Empirical Evaluation

Benchmarks:

- GloVe-100d [5]
- SIFT1M [6]

Metrics:

- Cosine similarity

- Recall@10

Results confirm:

$$U(\rho) \approx \rho,$$

independent of embedding family and dropout mode.
There is no sharp collapse threshold in the geometric regime.

9 Functional Utility Regime: Neural Inference

When transform-domain storage is applied to neural network weights, progressive reconstruction experiments reveal nonlinear behaviour:

- Low ρ produces incoherent output.
- Coherent inference emerges beyond a completeness threshold.
- Compatible priors reduce the apparent threshold.

This reflects nonlinear constraint satisfaction across interacting layers.
Important distinction:

Property	Geometric	Functional
Operation	Linear	Nonlinear
Utility type	Similarity	Generative coherence
Degradation	Proportional	Emergence-like

The nonlinear behaviour does not imply cryptographic secrecy; full coefficient access still permits reconstruction. It reflects instability of deep inference under incomplete constraints.

10 Comparison with Encryption

Property	Encryption	FHE	Transform-domain
At-rest protection			Partial
Runtime plaintext exposure	Yes	No	Distributed
Similarity overhead	Low	High	Low
Partial exfiltration resistance	No	Yes	Yes (proportional)
Governance gradient	No	No	Yes

Transform-domain storage complements encryption. It does not replace cryptographic protection.

11 Limitations

- Not cryptographically secure.
- Full exposure permits reconstruction.
- Geometric regime exhibits proportional scaling, not collapse.
- Functional regime is currently characterised qualitatively.
- Governance infrastructure is required.

12 Conclusion

Transform-domain semantic storage induces two distinct utility regimes.

In the geometric regime (vector embeddings), semantic utility scales approximately linearly with completeness ratio:

$$U(\rho) \approx \rho.$$

This enables proportionate, quantitative access control without homomorphic overhead.

In the functional regime (neural weights), inference exhibits nonlinear emergence behaviour under incomplete reconstruction, reflecting constraint satisfaction dynamics rather than linear projection loss.

Recognising the distinction between regimes resolves apparent contradictions between earlier experiments and clarifies the confidentiality profile of transform-domain systems.

When combined with encryption and governance controls, transform-domain storage provides a practical intermediate confidentiality layer for large-scale semantic databases.

References

[1] Johnson J, Douze M, Jégou H. FAISS: A library for efficient similarity search and clustering of dense vectors. Big Data, 2017.

[2] Malkov YA, Yashunin DA. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE TPAMI, 2018.

[3] Gentry C. A Fully Homomorphic Encryption Scheme. Stanford University, 2009.

[4] Costan V, Devadas S. Intel SGX explained. IACR ePrint 2016/086.

[5] Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. EMNLP, 2014.

[6] Jégou H, Douze M, Schmid C. Product quantization for nearest neighbor search. IEEE TPAMI, 2011.