isar innovations

isar innovations Technical Report

Keyed Phase Transform for Private Vector Retrieval

Nino Wagensonner

info@isar-innovations.dev

April 2026

Abstract

Traditional encryption protects stored data but destroys the ability to search by meaning. Distance-preserving schemes maintain searchability but leak pairwise structure. We introduce the *keyed phase transform* (KPT), which makes semantic search itself key-dependent: with the correct key, retrieval works normally; with any other key, the score space collapses to a noise zone where no query produces usable signal. Across three datasets (AG News, 20 Newsgroups, MS MARCO) with sentence-transformer embeddings (384-dim), KPT achieves Recall@10 ≥ 0.906 with public-layer pair correlation ≤ 0.029, while DCPE/SAP achieves Recall@10 ≤ 0.393 with pair correlation ≥ 0.714 at its best tested operating point. At 100,000 documents (AG News), pair correlation tightens to 0.002 under controlled sequential sampling. Through 8-component ablation we identify phase coherence as the sole isolation mechanism and prove $\text{sinc}^2(\delta)$ score decay (Theorem 1) and 623-bit geometric key depth (Theorem 2). Per-document scrambling ensures all scrambled wave components have pair correlation < 0.01 at rest; unscrambled routing components (`mode_weight`, `mode_energy`) show residual signal at pair_corr ≈ 0.07. Five adversarial attacks fail empirically, including chosen-plaintext and two-user collusion. Combined with AES-256-GCM on text payloads, KPT provides defense in depth: content remains AES-protected regardless of KPT status; even a broken KPT layer would expose only pairwise similarity between opaque ciphertext, not content. The method adds 0.05 ms overhead per document and requires only `numpy`.

1 Introduction

Encrypting a file with AES makes it unreadable without the key, but also unsearchable. Finding "documents about cats" in an encrypted vector database requires decrypting everything first. Encryption and searchability are at odds: protecting content destroys the ability to search by meaning.

Vector databases store embeddings, numerical representations of text that capture semantic similarity. These embeddings are powerful because similar documents have similar vectors. But that same property makes them dangerous: Morris et al. [2023] showed 92% text reconstruction from embeddings; OWASP [OWASP, 2025] added "Vector and Embedding Weaknesses" as LLM08 to their Top 10; PoisonedRAG [Zou et al., 2025] demonstrated that 5 crafted documents can manipulate 90% of RAG responses.

Existing defenses face a trilemma. Standard encryption (AES-GCM) makes embeddings unsearchable. Distance-preserving schemes (DCPE/SAP) maintain searchability but intentionally preserve pairwise distances, leaking exactly the similarity structure an attacker needs. ORAM-based approaches (Compass) provide formal security but at 10–100× latency.

We propose a different approach: instead of encrypting the bytes and losing searchability, we make *the search itself* key-dependent. The keyed phase transform (KPT) decomposes each embedding into a superposition of key-dependent carrier waves with controlled phase shifts. The score function measures the coherence of the resulting interference pattern. With the correct key, constructive interference produces a strong retrieval signal. With any other key, the phases are pseudorandom and destructive interference collapses the score to noise, just as a diffraction grating produces no coherent pattern when the slit spacing is unknown.

AES protects the payload. KPT protects the retrieval meaning. The score decay follows a $\text{sinc}^2(\delta)$ law (Theorem 1), identical to single-slit diffraction. A system using both (AES on stored text, KPT on stored vectors) protects content at rest and makes unauthorized semantic search impossible, without decryption at query time.

Scientific claims.

1. KPT exhibits hard key sensitivity with preserved semantic retrieval quality: correct-key scores cluster at $\sim$0.557, wrong-key scores collapse to $\sim$0.012.
2. Wrong-key access collapses to a noise zone, not degraded results. The contribution is noise induction in the search space, not byte-level encryption.
3. The method functions as a single-embedding access layer for at-rest-protected semantic search, requiring no secondary plaintext index.

Scope and non-claims. This work does not yet provide a formal security reduction to a known hard problem. We use SHA-256 for key derivation and provide two analytical theorems plus empirical adversarial

evaluation, but the formal bridge from "empirically secure" to "provably secure" remains open work. The mathematical building blocks are compatible with such a proof (phase retrieval hardness [Maillard et al., 2020], sinc^2 decay, and nonlinear pipeline non-invertibility), but the reduction itself is future work. We do not claim "no leakage," "proven secure," or "better on all datasets." We claim key-gated semantic retrieval with empirically measured isolation and analytical bounds.

Contributions.

1. **Phase coherence** identified as sole key-isolation mechanism through 8-component ablation.
2. **Analytical bounds:** $\text{sinc}^2(\delta)$ score decay (Theorem 1) and 623-bit geometric key depth (Theorem 2).
3. **Three-dataset evaluation** (AG News, 20 Newsgroups, MS MARCO at 6k; AG News at 100k) with DCPE/SAP comparison.
4. **Per-document scrambling** that reduces `wave_real` leakage from 0.904 to -0.004 and defeats collusion attacks.
5. **Adversarial evaluation:** five attacks reproduced and defeated (score-oracle, known-plaintext, statistical leakage, chosen-plaintext, collusion).

2 Related Work

2.1 Threat Model

Passive attacker (data-at-rest). Access to stored embeddings (public layer), no query capability. Goal: reconstruct pairwise similarity or original text.

Active attacker (score-oracle). Can issue queries and observe scores without the secret key. Goal: reconstruct document similarity graph.

We do *not* defend against authorized users, traffic analysis, or key compromise.

2.2 Why Not Standard Encryption?

Encrypting embeddings with AES-GCM produces opaque ciphertext. To answer a single semantic query, the system must decrypt all N documents, compute similarities in plaintext, and re-encrypt, turning $O(1)$ vector lookup into $O(N)$ decrypt-compare-encrypt. At 100k documents this is prohibitive and defeats the purpose of a vector index.

KPT makes the encrypted representation itself searchable. The score operator works directly on stored keyed vectors without decryption. The key gates retrieval semantics, not storage bytes.

2.3 Distance-Preserving Encryption

DCPE/SAP [Fuchsbauer et al., 2022]: Scale-And-Perturb with formal RoR indistinguishability. Intentionally preserves approximate distances; our experiments show pair correlation 0.71 to 0.82 (Section 4). DCE [Liu et al., 2025] achieves exact distance comparison but at $8\times$ storage. IronCore Cloaked AI (2025) productionizes SAP but acknowledges CPA weakness.

2.4 Other Approaches

- **Compass** [Zhu et al., 2024]: ORAM-based, formally secure, 10–100× latency.
- **STEER** [He et al., 2025]: Learned alignment transform for query privacy.
- **Keyed Chaotic Dynamics** [Fagan, 2025]: Additive tensor masking ($\tilde{X} = X + S$) via chaotic maps. Purely conceptual: no experiments, no retrieval evaluation, additive masking is trivially invertible.
- **OSNIP** [Cao et al., 2026]: Null-space injection for LLM inference, not retrieval.

2.5 Wave-Based Embeddings

Concurrent works (2025) model embeddings as $\psi(x) = A(x) \cdot \exp(i\varphi(x))$ using interference as similarity. Wave-Based Semantic Memory and PRISM validate our mathematical framework but provide no security analysis.

2.6 Theoretical Foundations

Our sinc^2 decay connects to the Fourier transform of the rectangular function (Bochner's theorem; Rahimi and Recht 2007). Phase retrieval hardness [Maillard et al., 2020] provides information-theoretic support. We distinguish from DRPE (optical encryption), which is vulnerable to known-plaintext attacks via iterative Fourier reconstruction; our nonlinear pipeline (`tanh`, routing, chunking) prevents this.

3 Method

3.1 Architecture

Given $x \in \mathbb{R}^d$ and key K:

1. **Carrier wave:** Project x through key-derived orthogonal matrices A_K, B_K. Amplitude via `tanh` modulation, phase from carrier projection.
2. **Mode decomposition:** Route carrier through M=8 modes via key-dependent sparse top-k softmax routing.
3. **Superposition:** Weight modes with key-dependent phase shifts $\varphi_K^{(m)}$, normalize.
4. **Per-document scramble:** Permute + sign-flip wave vectors with seed from SHA-256(`key` ‖ `mode_weight`). Each document stored in its own coordinate system (Section 3.3).
5. **Public layer:** Destructured projection of mode weights (intentionally lossy).
6. **Score:** Unscramble both doc and query, then compute

$$\alpha\big(0.80 \cdot \text{coherence} + 0.10 \cdot \text{energy} + 0.10 \cdot \text{mode\_support}\big) + (1 - \alpha) \cdot \text{base\_overlap}^2,$$

where $\alpha = 0.46$. The coherence term drives key isolation; energy and mode_support contribute a small key-independent offset $\varepsilon \approx 0.012$.

All parameters derived from key string via SHA-256-seeded PRNG.

3.2 Physical Intuition: Why Phases Gate Meaning

KPT applies wave mechanics to embedding vectors. The physics is literal, not metaphorical.

Each document embedding is decomposed into a carrier wave with amplitude and phase, then split across M=8 modes with key-dependent phase shifts, analogous to a prism splitting light into frequencies. The key determines the phase shifts; the score function measures coherence of the resulting interference pattern.

Constructive interference (correct key). Shared phase structure means wave components align. Overlaps add coherently, producing a strong score (∼0.557).

Destructive interference (wrong key). Mismatched phases point in pseudorandom directions. Contributions cancel on average. The score (∼0.012) reflects the absence of signal, not wrong information.

sinc^2 decay. Theorem 1 follows the same mathematics as single-slit diffraction: the Fourier transform of a rectangular phase window. Small phase errors collapse the signal rapidly.

Score compression (Section 5.6) follows directly: all correct-key documents produce near-maximal constructive interference, so scores cluster in a narrow band. Wrong-key scores are query-independent because destructive interference yields the same noise floor regardless of the query.

Controlled ablation verifies this is the *sole* isolation mechanism:

- Orthogonal projection alone: ratio = 1.0× (zero isolation, no phase structure).
- Phase rotation alone: ratio scales super-linearly with d (3,753× at d=1024).

3.3 Per-Document Scrambling

A naive implementation stores wave vectors with a single key-dependent rotation shared across all documents. Since orthogonal rotations preserve inner products, an attacker can compute cosine similarity on the stored wave vectors and recover the original similarity structure, even without the key.

We address this with *per-document scrambling*: for each document i, a key-dependent permutation P_i and sign-flip vector S_i are derived from SHA-256(`key` $\|$ `mode_weight`$_i$`.bytes`). The `mode_weight` vector is unique per document (determined by content-dependent routing) and is not itself scrambled, making it a stable seed across encode and decode. The stored wave vector is:

$$\texttt{stored\_wave}[i] = S_i \odot \texttt{wave}[i][P_i].$$

Since $P_i \neq P_j$ for $i \neq j$, cosine similarity between stored vectors no longer corresponds to the original similarity. The score function knows the key, reconstructs P_i and S_i, and unscrambles both doc and query vectors before computing the overlap.

- **Before scrambling:** `wave_real` pair_corr $= 0.904$ (critical leak).
- **After scrambling:** `wave_real` pair_corr $= -0.004$ (noise level).

All scrambled wave components now have pair_corr < 0.01. The unscrambled routing components (`mode_weight`, `mode_energy`) retain weak residual signal (pair_corr ≈ 0.07, ARI ≈ 0.09). The scramble adds $O(d)$ overhead per document (same order as the dot product itself) and does not affect Recall@10, correct-key scores, or wrong-key isolation.

This defense also defeats the *sorted-value attack* (sorting values to bypass permutation) because the sign flips destroy the value distribution. It defeats *collusion attacks* (two users combining their encrypted views of the same documents) because each key produces different per-document scrambles: concatenating differently scrambled vectors yields noise (pair_corr 0.001, down from 0.944 without scrambling).

3.4 Why Post-Hoc Transformations Cannot Recover the Signal

An attacker with access to the stored vectors might attempt mathematical transformations (exponential rescaling, PCA, neural network inversion) to extract the original structure. This fails because:

1. **Per-document scrambling** means each stored vector lives in a different coordinate system. Pairwise operations (cosine, distance) across documents are meaningless.
2. **Eight nonlinear, key-dependent stages** (orthogonal projection $\rightarrow$ `tanh` $\rightarrow$ mode routing $\rightarrow$ softmax $\rightarrow$ phase rotation $\rightarrow$ per-doc scramble $\rightarrow$ chunk pooling $\rightarrow$ normalization) form a composition that is not invertible without the key. Our CPA attack (Section 5.4) confirms this: even with 500 chosen plaintext pairs, an MLP achieves only cosine 0.005 on scrambled wave vectors.
3. **Score compression is a feature.** Wrong-key scores have variance $\sim 10^{-12}$ (Section 5.2). No monotonic transformation can extract signal from a flat distribution.

3.5 Key Derivation

HKDF (RFC 5869) with SHA-512, producing 512-bit key material per derivation context. Each pipeline stage (projection, routing, phase shifts, scrambling) uses a separate HKDF info parameter for domain separation. This provides 256-bit post-quantum security under Grover's algorithm and supports arbitrary key lengths without entropy truncation.

4 Experiments

4.1 Setup

Baseline: DCPE/SAP at $\sigma \in \{0.05, 0.10, 0.20\}$.

Table 1: Dataset overview.

	AG News	**20 Newsgroups**	**MS MARCO**
Type	Classification (4 cls)	Classification (20 cls)	Passage retrieval
Docs / Queries	6,000 / 1,200	6,000 / 1,200	6,000 / 1,200
Vectorizer	sentence-transformers (384-dim)		
Seeds	3	3	3

Table 2: Ablation study: isolation ratio when individual components are disabled.

Component disabled	d=384	d=1024	**Interpretation**
None (baseline)	47×	226×	
Decoy floor	35×	**15,905**×	Primary limiter
Top-k routing	**92**×	95×	Routing reduces isolation
Base squaring	27×	82×	Squaring *helps* (amplifies gap)
All nonlinearities	36×	54×	
Pure phase only	789×	3,753×	Phase is the mechanism

4.2 Ablation: Source of Key Isolation

8 nonlinearities disabled individually across $d \in \{64, 128, 256, 384, 512, 768, 1024\}$:

Finding: The decoy floor (mode minimum weight 0.03) is the primary limiter. Without it, isolation scales super-linearly because non-matching modes contribute zero. Trade-off: decoy improves retrieval robustness.

4.3 Comparison with DCPE/SAP

Table 3: KPT vs. DCPE/SAP across three datasets. Bold indicates best result per column.

	AG News		**20NG**		**MS MARCO**	
Method	Recall	pair_corr	Recall	pair_corr	Recall	pair_corr
KPT	**0.991**	**0.029**	**0.906**	**−0.005**	**0.990**	**−0.018**
DCPE σ=0.05	0.393	0.763	0.340	0.911	0.379	0.714
DCPE σ=0.10	0.076	0.397	0.095	0.702	0.063	0.345
DCPE σ=0.20	0.006	0.135	0.013	0.326	0.006	0.111

KPT achieves Recall ≥ 0.906 with pair_corr ≤ 0.029 across all datasets. Under controlled conditions (fixed seed, sequential sampling), pair_corr drops to 0.001–0.005; the 0.029 value reflects variance across 3-seed balanced sampling in the evaluation pipeline.

At the three tested noise levels, DCPE does not achieve Recall > 0.4 with pair_corr < 0.7. We note that DCPE's noise parameter is continuously tunable; intermediate values may exist, though the fundamental design (intentional distance preservation) creates an inherent leakage floor.

Caveat: DCPE provides formal RoR guarantees. KPT does not.

4.4 Key Sensitivity (Avalanche)

12 key variants at Levenshtein distances 0–27: distance 0 yields full retrieval; distance ≥ 1 yields complete collapse (ratio 48–228×). No correlation between edit distance and attack success. This is not gradual degradation. It is a qualitative phase transition from "fully usable" to "completely useless."

4.5 Key-Space Depth

Theorem 1 (sinc^2 Phasor Decay)**.** *For the coherence component of the score under uniform phase perturbation $\delta \in [-\delta, \delta]$:*

$$\frac{E[\text{coherence}(K')]}{E[\text{coherence}(K)]} = \text{sinc}^2(\delta) = \left(\frac{\sin \delta}{\delta}\right)^2.$$

Proof. Expected phasor overlap $E[\exp(i\eta)] = \sin(\delta)/\delta$ for $\eta \sim \text{Uniform}(-\delta, \delta)$. The coherence term is quadratic in phasor overlaps, so the ratio is $\text{sinc}^2(\delta)$. Verified with SymPy. □

Scope of Theorem 1. The theorem describes the *coherence component* (weight α=0.46 in the score), not the full score. The full score includes `mode_support` and energy terms that contribute a positive offset $\varepsilon \approx 0.012$ independent of key correctness. Empirically: correct-key score = 0.557, wrong-key score = 0.012 (of which ∼0.012 is the constant offset from `mode_support`/energy, and ∼0 is the coherence contribution). The sinc^2 decay correctly predicts the collapse of the coherence term; the residual offset is a known architectural artifact of the decoy floor.

Theorem 2 (Geometric Key Depth)**.** *The phase search space has geometric depth $B = D \times \log_2(\pi/\delta_c)$ bits, where $D = M \times d$ phase parameters and δ_c satisfies $\text{sinc}^2(\delta_c) =$ wrong-key coherence ratio.*

Proof. Phase space $\mathbb{T}^D = [-\pi, \pi]^D$. $P(\text{random phase vector in } \delta_c\text{-ball}) = (\delta_c/\pi)^D$. □

Table 4: Geometric key depth for different configurations.

Config	D	Geometric Depth
dim = 96, m=8	768	156 bits
dim = 384, m=8	**3072**	**623 bits**
dim = 768, m=8	6144	1245 bits

Critical qualification: geometric depth ≠ key security. The 623 bits measure the volume of the continuous phase torus, but all phase parameters are derived from a single SHA-256 seed (256 bits). The effective security is therefore:

$$\text{effective\_bits} = \min\big(H(\text{key}),\ 256,\ 623\big) = \min\big(H(\text{key}),\ 256\big),$$

where $H(\text{key})$ is the entropy of the key string. For a random 32-byte key: 256 bits. For a 20-character alphanumeric password: ∼119 bits. The geometric depth (623 bits) confirms that the *phase geometry does not limit security*; the bottleneck is the key/seed entropy, not the transform.

Validation. 10,000 random keys: μ=0.0065, σ=0.0044, max =0.0196 vs. correct = 0.5577. Gap: 124σ.

4.6 Multi-User Demonstration

Alice encrypts 500 documents. Three queries with Alice's key, Bob's key, and public-only:

Table 5: Multi-user query results for "Stock market crash."

Searcher	Top result	Score
Alice (correct)	Business: markets, Greenspan	0.558
Bob (wrong key)	World: Arafat, Sharon	0.012
Attacker (public)	Random: Parliament, Microsoft	0.060

4.7 Scaling to 100,000 Documents

To address the concern that small-scale experiments may not transfer, we evaluate KPT and DCPE/SAP at production scale: 100,000 AG News documents encoded with sentence-transformers (384-dim) on an RTX 3060 GPU (sbert) + CPU (KPT/DCPE scoring).

Table 6: Scaling behavior from 1,000 to 100,000 documents.

	KPT			DCPE	
Scale	Recall@10	pair_corr	encode/doc	Recall@10	pair_corr
1,000	0.990	0.001	0.04 ms	0.403	0.755
5,000	0.990	0.002	0.04 ms	0.328	0.763
10,000	0.990	0.002	0.05 ms	0.340	0.763
50,000	0.990	0.002	0.07 ms	0.317	0.763
100,000	**0.993**	**0.002**	0.05 ms	0.352	0.763

KPT recall *improves* slightly at scale (0.993 at 100k vs. 0.990 at 6k). Under controlled sequential sampling (single seed, no balanced resampling), pair_corr tightens to 0.001–0.002 across all scales, consistent with the ≤ 0.029 measured in the 3-seed balanced evaluation (Section 4.3), where sampling variance accounts for the difference. Encode overhead is 0.05 ms/doc.

4.8 Public Layer Dimensionality

Observation 1 (Public Layer Compression)**.** *PCA on the public layer of 10,000 encoded documents at 100k scale:*

Table 7: Public layer dimensionality analysis.

Measure	**Value**
Public layer dimensions	14
Effective dimensions (95% variance)	12
Plaintext dimensions	384
Dimensionality ratio	**12/384 = 3.1%**

The nonlinear pipeline (orthogonal projection → `tanh` → mode routing → `tanh` → chunk pooling → normalize) collapses 384 input dimensions to 14 public dimensions, of which only 12 carry 95% of the variance.

Qualification. PCA effective rank measures linear variance, not mutual information. A formal bound on $I(X; \text{Public}(X))$ via the Data Processing Inequality is in preparation. The observation is consistent with empirical pair_corr ≤ 0.029.

4.9 Vectorizer Independence

KPT operates on any embedding model without retraining or parameter adjustment. We test three models spanning different architectures and dimensions:

Scores, isolation, and leakage are consistent across all three. KPT adds 0.05 ms/doc regardless of embedding dimension because the pipeline is pure matrix arithmetic (no model inference). The embedding model dominates wall-clock time; KPT overhead is negligible.

Table 8: Vectorizer independence: KPT performance across embedding models.

Model	Dim	Correct Score	Wrong Score	Recall@3	wave pair_corr
all-MiniLM-L6-v2	384	0.557	0.012	1.000	+0.002
mxbai-embed-large-v1	1024	0.558	0.011	1.000	+0.002
bge-m3	1024	0.558	0.011	1.000	+0.010

4.10 Quantization Sensitivity

Production vector databases routinely quantize embeddings to reduce storage. We test whether KPT tolerates reduced precision by quantizing only the wave vectors (`wave_real`, `wave_imag`, `base_wave_real`, `base_wave_imag`) while keeping `mode_weight` at float32 (required as scramble seed).

Table 9: Quantization sensitivity for wave vector storage.

Precision	Recall@10	Score Mean	Storage/doc
float64	0.980	0.5566	54.1 KB
float32 (baseline)	0.980	0.5566	27.1 KB
int16 fixed-point	**0.982**	**0.5566**	**13.6 KB**
float16	0.906	0.5566	13.6 KB
12-bit quantized	0.912	0.5566	13.6 KB
8-bit quantized	0.214	0.5566	6.9 KB

int16 fixed-point (scaling each tensor to $[-32767, 32767]$) achieves identical recall at half the storage of float32. float64 provides no benefit over float32.

Below 12 bits, recall degrades rapidly. The root cause is score compression: correct-key scores cluster within ± 0.001, so quantization errors of comparable magnitude flip rankings. The score mean remains stable (the coherence signal is preserved) but the fine-grained ordering is lost.

Critical constraint. `mode_weight` must remain at full float32 precision. It serves as the per-document scramble seed via SHA-256(`key` $\|$ `mode_weight`$_i$`.bytes`). Even single-bit rounding changes the hash, producing a wrong permutation and destroying the score. At 8 floats per document (32 bytes), this is negligible storage overhead.

4.11 Beyond Search: Key-Gated Clustering

KPT gates not only retrieval but any operation that depends on pairwise similarity. We test clustering: KMeans on score-matrix rows (each document's similarity profile as a feature vector) with 10 clusters, 1,000 documents.

Table 10: Key-gated clustering: cluster recovery under different access levels.

Space	ARI	NMI
Plaintext (baseline)	1.000	1.000
Authorized (correct key)	**1.000**	**1.000**
Wrong-key scores	0.040	0.125
Public layer	0.001	0.021
Scrambled wave (at rest)	−0.000	0.017

With the correct key, cluster recovery is perfect (ARI = 1.0). With a wrong key, it collapses to noise. The kNN neighbor overlap tells the same story: 89.7% with correct key, 0.9% with wrong key (random baseline at k=10).

This extends KPT beyond retrieval to any downstream task that operates on pairwise similarity: clustering, deduplication, recommendation, anomaly detection, and topic tracking can all be key-gated through the same stored representation.

5 Adversarial Analysis

Two independent red-team analyses (cryptographic + ML), then empirical reproduction on 2,000 documents.

5.1 Known-Plaintext Attack

Claim: Neural net achieves cosine 0.4–0.55 from 500+ pairs.

Result: MLP ($256 \to 512 \to 384$) achieves cosine **0.228** (N=800). Similarity correlation 0.048. The nonlinear pipeline destroys enough information for inversion to fail.

5.2 Score-Oracle Attack

Claim: Cross-key score profiles reconstruct the Gram matrix.

Result: Gram-matrix correlation **0.024** (200 queries $\times$ 2,000 docs). kNN overlap 0.006 vs. random 0.005. Cross-key scores are too flat (variance $\sim 10^{-12}$) for any signal.

5.3 Statistical Leakage

Claim: pair_corr = 0.029 becomes significant at $N \geq 1{,}200$.

Result: Controlled sweep N=50–2,000 on the public layer: pair_corr oscillates around $\mathbf{0.00 \pm 0.01}$ at all N. Public-layer homophily equals random chance. Note: this tests only the public layer; `mode_weight` and `mode_energy` show weak residual signal (pair_corr ≈ 0.07, ARI ≈ 0.09), which is documented in Section 5.6.

5.4 Chosen Plaintext Attack (CPA)

Claim: Attacker chooses documents with known structure (orthogonal basis, known clusters) and learns the transformation.

Three CPA strategies tested with 500 chosen pairs:

- **Linear reconstruction** (least-squares on `wave_real`): cosine **0.001**. Linear model cannot approximate the nonlinear scrambled pipeline.
- **Cluster injection** (10 known clusters): ARI $-\mathbf{0.002}$ on `wave_real`, $-\mathbf{0.002}$ on public. No cluster recovery in any stored representation.
- **MLP inversion** ($512 \to 512$ network): cosine **0.005** on `wave_real`. The per-document scramble makes the mapping doc-dependent, defeating generalization.

CPA is strictly stronger than the known-plaintext attack (Section 5.1) because the attacker controls the input distribution. It still fails.

5.5 Collusion Attack

Claim: Two users with different keys combine their encrypted views of the same documents.

Four collusion strategies tested (500 documents, 2 keys):

Table 11: Collusion attack strategies and results.

Strategy	Correlation
Concatenate public layers	-0.002
Concatenate wave vectors	**0.001**
Wave product (element-wise)	0.000
CCA alignment	-0.001

Without per-document scrambling, concatenating wave vectors yielded 0.944 correlation, a critical vulnerability. The scramble reduces this to noise because each key produces different per-document permutations: combining differently scrambled vectors is like combining unrelated random vectors.

5.6 Remaining Limitations

- **Score compression.** Correct-key scores cluster at 0.557 ± 0.001. Ranking is preserved and retrieval quality is unaffected (Recall@10 = 0.993), but absolute score differences are small. A natural first impulse is to stretch scores post-hoc (e.g., via exponential rescaling), but this would amplify the signal available to score-oracle attackers. The compression is a direct consequence of the isolation mechanism: the same phase coherence that collapses wrong-key scores to $\sim$0.01 also compresses correct-key scores into a narrow band. Any monotonic rescaling applied server-side would equally benefit an attacker. Client-side display normalization after authenticated retrieval is safe and recommended.

- **Cross-key score bias.** Wrong-key results are query-independent (same documents always rank highest). This does not affect security but reveals wrong-key status.

- **No formal security reduction (yet).** We provide analytical bounds (2 theorems) and empirical evidence (5 defeated attacks), but no reduction to a known hard problem. The building blocks for such a proof exist (phase retrieval hardness, sinc^2 decay structure, nonlinear non-invertibility) and this is an open research direction.

- **Mode routing leakage.** `mode_weight` and `mode_energy` are not scrambled (`mode_weight` serves as the per-document scramble seed via SHA-256(`key` $\|$ `mode_weight`$_i$`.bytes`)). They show weak cluster signal (ARI ≈ 0.09, pair_corr ≈ 0.07) because similar documents route through similar modes. This is not sufficient for practical cluster recovery but is the weakest remaining component.

- **Precision sensitivity.** KPT requires at least 16-bit precision for wave vectors. int16 fixed-point achieves identical Recall@10 (0.982) at half the storage of float32. Below 12 bits, recall degrades rapidly due to score compression: all correct-key scores cluster within ± 0.001, so even small quantization errors flip rankings. `mode_weight` must remain at float32 because it serves as the scramble seed; any rounding changes the SHA-256 hash and produces a wrong permutation. int8 quantization destroys retrieval entirely (Recall@10 = 0.214).

6 Discussion

6.1 What KPT Is

KPT is access control on meaning. The same stored object supports authorized search (correct key produces ranked results) and denies unauthorized access (wrong key produces noise). No secondary plaintext index, no decryption at query time, no trust in the storage layer.

In multi-tenant vector databases, different users search the same physical index with different keys. Each sees only their own semantic space.

6.2 Why Not Just Encrypt Everything?

Standard symmetric encryption (AES-GCM) on embeddings provides stronger cryptographic guarantees but eliminates searchability entirely. To answer "find documents about cats," the system must:

1. Decrypt all N embeddings ($O(N)$ AES operations).
2. Compute query similarity in plaintext ($O(N \cdot d)$ dot products).
3. Return results, re-encrypt.

At N=100,000, d=384, this requires decrypting 147 MB of vectors per query. With KPT, the score operator works directly on the stored keyed vectors: no decryption, $O(N \cdot d)$ dot products on the encrypted representation, same asymptotic cost as plaintext retrieval.

KPT occupies the gap between "perfectly secure but unsearchable" and "searchable but leaky."

Table 12: Property comparison: AES-GCM vs. KPT vs. DCPE/SAP.

Property	**AES-GCM**	**KPT**	**DCPE/SAP**
Search without decryption	No	Yes	Yes
Leakage (pair_corr)	0	$\leq$0.029	$\geq$0.714
Formal proof	Yes	No	Yes (RoR)
Query cost	$O(N)$ decrypt + $O(Nd)$	$O(Nd)$	$O(Nd)$

6.3 Security–Utility Trade-off

The decoy floor controls the trade-off: 0.24 (default) gives ratio 47× with robust retrieval; 0.0 gives 15,905× at d=1024 but retrieval is more brittle. Analogous to DCPE's noise parameter but on a different axis.

Score compression is inherent, not fixable. The same phase coherence that provides key isolation also compresses correct-key scores into a narrow band (0.557±0.001). This is not an implementation artifact; it follows directly from the sinc^2 decay structure: correct-key coherence is near-maximal for all documents, with fine-grained ranking encoded in the residual base-overlap term. Attempts to widen the score distribution (e.g., exponential stretching) would proportionally increase information available to score-oracle attackers. We recommend treating KPT scores as ordinal (ranking) rather than cardinal (absolute similarity), which aligns with standard top-k retrieval practice.

6.4 Beyond Search

KPT creates a key-gated semantic layer, not just a search index. Any operation on pairwise similarity becomes key-dependent: clustering, recommendation, deduplication, anomaly detection, topic tracking. In multi-user memory systems, each user's semantic structure (topics, episodes, connections) exists only in their keyed space. The storage layer holds scrambled vectors that support none of these operations without the key. RAG pipelines benefit directly: the retrieval step is key-gated, so the language model receives only authorized context.

6.5 Scaling Beyond 100k: Two-Stage Architecture

Brute-force KPT scoring scales linearly with document count. In multi-tenant deployments, a user-ID filter restricts scoring to each user's documents. At 100k documents per user, query latency is ∼180 ms on CPU.

6.6 When to Use KPT

Appropriate: Data-at-rest protection for vector databases, multi-tenant semantic search, Recall > 0.99 required, no HE/ORAM budget, rate-limited score API.

Not appropriate: Formal cryptographic guarantees required, compliance regimes that mandate proven encryption, or attacker has unrestricted score-oracle access. Our score-oracle test (Section 5.2) used 200 queries × 2,000 docs and found Gram correlation 0.024. As an informal intuition: wrong-key score variance is $\sim 10^{-12}$ per query, suggesting $O(10^{12})$ queries would be needed to accumulate a distinguishable signal. However, this is not a formal bound; the actual threshold depends on the attacker's aggregation strategy and detector design, which we have not formalized. We recommend rate-limiting as a practical deployment measure.

6.7 Method Comparison

7 Conclusion

We present the keyed phase transform, a method that gates semantic retrieval on a secret key. Phase coherence is the sole isolation mechanism (8-component ablation); the score decays as $\text{sinc}^2(\delta)$ under phase perturbation (Theorem 1), yielding 623-bit geometric key depth (Theorem 2). Per-document scrambling

Table 13: Comprehensive method comparison.

Property	KPT	AES-GCM	DCPE/SAP	Compass
Recall@10 (100k docs)	**0.993**	N/A	0.352	∼1.0
pair_corr (public)	$\leq$ **0.029**	0	$\geq$0.714	0
Public dimensionality	12/384 (3.1%)	N/A	full	0
Formal proof	No	Yes	Yes (RoR)	Yes
Search w/o decrypt	**Yes**	No	Yes	Yes
Overhead	0.05 ms/doc	∼0	<1 ms/doc	10–100×
Vectorizer-agnostic	Yes	Yes	Yes	No

reduces at-rest wave vector leakage from 0.904 to −0.004, defeating collusion and chosen-plaintext attacks. Five adversarial evaluations fail empirically.

Across three datasets, KPT achieves Recall@10 $\geq$ 0.906 with public-layer pair_corr $\leq$ 0.029. At 100,000 documents (AG News), Recall@10 reaches 0.993 with pair_corr tightening to 0.002 under controlled sampling. KPT consistently outperforms DCPE/SAP on the retrieval-leakage frontier at all tested operating points.

In a combined deployment, text payloads are AES-256-GCM encrypted and vector representations are KPT-protected. This dual-layer architecture provides defense in depth: AES protects content, KPT protects semantic structure. Even if the KPT layer were broken, an attacker gains only pairwise similarity information between opaque ciphertext blobs, not content. KPT is access control on meaning: a lightweight, mathematically grounded key-isolation layer that makes stored embeddings searchable with the right key and useless without it.

References

Cao et al. OSNIP: Obfuscated semantic null space injection, 2026.

Fagan. Keyed chaotic dynamics for privacy-preserving neural inference, 2025.

Georg Fuchsbauer, Riddhi Ghosal, Adam O'Neill, and Krzysztof Pietrzak. Approximate distance-comparison-preserving symmetric encryption. In *Security and Cryptography for Networks (SCN)*, 2022.

Jiaqi He et al. STEER: Transform before you query, 2025.

Yifan Liu et al. Privacy-preserving approximate nearest neighbor search (DCE), 2025.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

OWASP. LLM Top 10: LLM08 — Vector and Embedding Weaknesses. https://owasp.org/www-project-top-10-for-large-language-model-applications/, 2025.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Yifan Zhu et al. Compass: Encrypted semantic search with approximate nearest neighbors. IACR ePrint 2024/1255, 2024.

Wei Zou, Isack Lee, Yangsibo Huang, Peter Hartigan, and Yunfang Chen. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *USENIX Security Symposium*, 2025.