Keyed Phase Transform: Authorized Semantic Search over Protected Vector Data

Nino Wagensonner
info@isar-innovations.dev

May 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), a key-bound retrieval method that enables authorized semantic search over protected vector data while preventing unauthorized observers from reconstructing a usable semantic readout either from the stored representation or through the wrong-key search path.

KPT is not byte-level encryption for embeddings. It is a retrieval layer: with the correct key, the score operator reconstructs a usable ranking signal; with a wrong key, the score path becomes residual interference rather than a usable semantic ranking engine. Throughout the paper, a *readout* means this authorized score computation, not the key itself. The method combines key-dependent phase carriers, row-specific scrambling, authorized unbinding, and score aggregation. Payload bytes remain protected by standard encryption such as AES-256-GCM.

Across AG News, 20 Newsgroups, and MS MARCO with 384-dimensional sentence-transformer embeddings, KPT achieves Recall@10 at least 0.906 with at-rest observable pair correlation at most 0.029. At 100,000 AG News documents, Recall@10 reaches 0.993 and pair correlation tightens to 0.002 under controlled sequential sampling. The distance-comparison-preserving DCPE/SAP baseline reaches Recall@10 at most 0.393 with pair correlation at least 0.714 at its best tested operating point. Here pair correlation measures how well plaintext pairwise similarity can be recovered from the tested observable. Per-document scrambling reduces at-rest wave-vector leakage from 0.904 to -0.004, and five reproduced attacks did not recover useful semantic geometry in the tested regimes: known-plaintext, score-oracle, statistical leakage, chosen-plaintext, and collusion.

We separate the proof status sharply. This paper states the algebraic recovery, wrong-key residual, shared-transform failure, and monotone readout facts used by the implementation. The companion KPT-HGA and wrong-key manuscripts develop the larger margin and threshold proof program [7, 8]. The benchmark layer validates concrete operating points; it does not replace the theorem chain or imply a full reduction to a standard cryptographic hard problem.

1 Introduction

Encrypting a file with AES makes it unreadable without the key, but also unsearchable by semantic content. A query such as "documents about cats" over an encrypted vector database typically requires decrypting all candidate embeddings, computing similarities in plaintext, and returning the nearest neighbors. Standard encryption and semantic search therefore pull in opposite directions.

Vector databases store embeddings: numerical representations of text whose geometry encodes semantic similarity. This geometry is useful because nearby vectors tend to represent related documents. It is also sensitive. Prior work has shown text reconstruction risks from embeddings

[5]; OWASP lists vector and embedding weaknesses in its LLM Top 10 [6]; and poisoning attacks against retrieval systems show how small crafted corpora can manipulate downstream generation [10].

Existing approaches face a trilemma. Conventional encryption such as AES-GCM protects bytes but destroys searchability. Distance-preserving schemes such as DCPE/SAP and DCE preserve approximate comparison but intentionally preserve the distance structure that an attacker may want [2, 4]. ORAM-based systems such as Compass provide stronger formal protection but at substantial latency and systems cost [9].

KPT takes a different route. Instead of preserving plaintext geometry or encrypting it into an unsearchable blob, it makes the retrieval score itself key-dependent. A document embedding is transformed into key-dependent wave channels. With the correct key, the query-side transform restores coherent geometry along the authorized score path and yields meaningful rankings. It does not reconstruct a plaintext embedding as the security object. With a wrong key, phase mismatch and scrambling turn the score path into low-utility residual interference. In a combined deployment, AES protects text payloads and KPT protects retrieval semantics.

The phrase *retrieval semantics* is intentionally narrower than cryptographic confidentiality. KPT tries to make the ranking relation unusable without the authorized readout path. It does not claim that every observable of the stored record is indistinguishable, and it does not replace encryption of payloads, private metadata, or key-bearing files.

Claims. The paper makes three claims.

1. KPT provides key-gated semantic retrieval: authorized queries preserve ranking utility, while wrong-key queries do not produce a usable semantic readout.
2. The isolation mechanism is phase coherence plus row-specific scrambling, not ordinary byte-level encryption.
3. The current implementation occupies a measured practical point on the retrieval-leakage frontier: high recall, low measured pair correlation, low encoding overhead, and empirical resistance to the reproduced attack suite.

Scope. KPT is not a replacement for payload encryption, and this paper does not claim a full end-to-end cryptographic reduction to a standard hard problem. The formal claim is the search-core claim: under the stated keyed score model, the correct key preserves the retrieval readout while wrong-key readouts are bounded as residual interference and do not provide a usable semantic ranking signal. The proof stack covers exact recovery, wrong-key decomposition, deterministic ranking preservation under explicit budgets, threshold decision rules, quantized-decision adapters, and structural failure of shared scrambling.

This boundary is intentional. Payload confidentiality, traffic patterns, routing-observable side channels, magnitude-multiset observables, adaptive oracle composition outside the modeled score process, and full implementation-level equivalence remain storage, deployment, or follow-on proof surfaces. They do not weaken the search-core theorem, but they must not be silently promoted into a full-system indistinguishability claim. The companion KPT-HGA and wrong-key manuscripts provide theorem-bearing details for the broader proof program [7, 8]; this paper uses them as support for the retrieval-layer claim and marks where the benchmark layer is empirical.

Contributions.

1. A complete system description for key-gated semantic retrieval over protected vector data.
2. A proof-aligned explanation of authorized recovery, wrong-key residuals, row-specific scrambling, and monotone readout, with budgeted decision certificates delegated to the companion proof manuscripts.

3. A three-dataset evaluation with DCPE/SAP comparison and a 100,000 document scaling study.

4. A per-document scrambling mechanism that reduces at-rest wave leakage from 0.904 to -0.004 and removes the collusion failure mode of a shared scramble.

5. Five reproduced adversarial analyses: known-plaintext, score-oracle, statistical leakage, chosen-plaintext, and collusion.

2 Related Work and Threat Model

2.1 Searchable protection

Standard encryption provides strong payload confidentiality, but a semantic query over encrypted embeddings requires decrypting all candidates. At 100,000 documents with $d = 384$, this means touching roughly 147 MB of vector payload per query before similarity can even be computed.

Distance-preserving encryption has the opposite tradeoff. DCPE/SAP provides formal indistinguishability in its own model, but its purpose is to preserve approximate distances [2]. In our experiments this produces high pair correlation at the operating points where retrieval remains useful. DCE supports distance comparison but increases storage [4]. Compass uses an ORAM-based approach with stronger formal security but higher latency [9]. STEER targets query-side exposure through learned transforms [3]; KPT instead measures at-rest observable pair geometry and wrong-key score paths. OSNIP and related null-space methods target LLM inference rather than vector retrieval [1].

Wave-based semantic memory and related phase-style embedding models motivate the terminology of interference and coherence, but they do not by themselves give a security model for protected retrieval. KPT uses this mathematical language for access control on semantic rankings.

2.2 Adversaries

We consider two primary adversary views. A passive at-rest adversary can inspect stored vector-side artifacts but cannot issue authorized queries. Its goal is to recover pairwise similarity, clusters, or text. An active score-oracle adversary can submit or replay queries through an unauthorized or wrong-key readout and observe returned scores, but does not obtain the retrieval key or plaintext embeddings. Its goal is to reconstruct document rankings or a similarity graph.

We do not defend against authorized users, key compromise, unrestricted traffic analysis, or all metadata leakage. Those belong to the operational layer around KPT. In the intended deployment, AES-256-GCM protects payload text, KPT protects semantic retrieval structure, and access controls protect keys, query rate, and metadata.

3 Method

3.1 Architecture

Given an embedding $x \in \mathbb{R}^d$ and key K, KPT derives all transform parameters from domain-separated key material. A stored row is indexed by i, a wave channel or mode by $r \in \{1, \ldots, M\}$, and the reported implementation uses $M = 8$. Routing observables are the mode weights and mode energies used to organize candidate scoring; they are metadata, not payload text. Only the positive envelope below is constrained to be positive; the projection matrices are not assumed positive definite. The implementation used in the experiments applies the following pipeline.

1. **Carrier construction.** Project x through key-derived matrices, with amplitude and phase modulation.

2. **Mode decomposition.** Route the carrier through $M = 8$ modes using key-dependent sparse top-k softmax routing.

3. **Superposition.** Combine modes with key-dependent phase shifts and normalization.

4. **Per-document scramble.** For each document, derive a scramble nonce from domain-separated key material and the document's base-wave state, then derive label-specific wave and base permutations/sign masks from that nonce. Store each wave channel in its own coordinate system.

5. **Routing side information.** Use or reconstruct non-secret routing observables needed to evaluate candidate scoring. In the reported v2 path these observables are not the scramble seed and need not be persisted as plaintext fields.

6. **Score.** With the correct key, regenerate the scramble, unbind the wave channels, and compute a retrieval score dominated by phase coherence with smaller energy, mode-support, and base-overlap terms.

The security-relevant algorithm is the triple

$$\mathrm{DocEncode}_K, \qquad \mathrm{QueryEncode}_K, \qquad \mathrm{Score}_K.$$

Payload encryption is compositional and external to this triple. If the document payload is stored in plaintext, KPT can still gate semantic retrieval but does not protect payload content. If the payload is protected by AES-GCM or another authenticated cipher, KPT supplies the searchable keyed representation while the cipher protects bytes.

For a document embedding x_i, the document encoder forms a complex base carrier

$$b_i = \mathrm{Norm}(a_K(x_i) \odot \exp(\mathrm{i}\,\theta_K(x_i))) \in \mathbb{C}^h,$$

routes it through sparse mode weights

$$g_i = \mathrm{TopKSoftmax}_{K,k_d}(\rho_K(b_i)) \in \Delta^{M-1},$$

and stores mode waves

$$w_{i,r} = \sqrt{g_{i,r}}\,\exp(\mathrm{i}\phi_{K,r}) \odot b_i, \qquad r = 1, \ldots, M.$$

The per-document scramble derives a row nonce $\eta_i = F_K(b_i)$ and label-specific permutations and signs from that nonce:

$$\widetilde{w}_{i,r} = \xi_{\eta_i,\mathrm{wave},r} \odot P_{\eta_i,\mathrm{wave},r}(w_{i,r}), \qquad \widetilde{b}_i = \xi_{\eta_i,\mathrm{base}} \odot P_{\eta_i,\mathrm{base}}(b_i).$$

The stored retrieval state is the scrambled wave/base state plus public or derived routing observables needed for candidate scoring; plaintext embeddings and payload bytes are not part of the KPT protected state.

For a query embedding q, $\mathrm{QueryEncode}_K(q)$ applies the same keyed carrier construction with query routing parameters k_q and collapse gain. The authorized score path regenerates the inverse scrambles from K and the stored nonce, obtains clear wave/base states $(w_{i,r}, b_i)$ and $(w_{q,r}, b_q)$, and evaluates

$$\mathrm{Score}_K(q,i) = \lambda_c\, C_K(q,i) + \lambda_e\, E_K(q,i) + \lambda_m\, M_K(q,i) + \lambda_b\, B_K(q,i),$$

where

$$C_K(q,i) = \frac{1}{M}\sum_{r=1}^{M} |\langle w_{q,r}, w_{i,r}\rangle|^2 \, G_K(q,i,r), \qquad M_K(q,i) = \langle g_q, g_i\rangle, \qquad B_K(q,i) = |\langle b_q, b_i\rangle|^2.$$

The factors G_K and E_K are the mode/energy gates induced by the routed wave state. In the reported implementation the weights are fixed explicitly as

$$(\lambda_c, \lambda_e, \lambda_m, \lambda_b) = (0.46, 0.10, 0.10, 0.44).$$

Equivalently, the prose formula below writes $\lambda_c = \alpha_0$ and $\lambda_b = 1 - \alpha_0$ with $\alpha_0 = 0.46$; the two 0.10 auxiliary terms are added routing stabilizers, so the implementation score is not constrained to be a convex mixture summing to one. This score operator is the KPT readout; it is not replaced by reconstructing a plaintext embedding and running ordinary cosine search.

The theorem-bearing part of the paper concerns this readout. Same-key unbinding, wrong-key relative-phase decomposition, threshold separation, margin transport, and quantized decision preservation are formalized for the score path. The stored-observable questions around routing metadata, magnitude multisets, and traffic are measured or scoped separately because they are not the semantic search decision itself.

Steps 1–3 are the keyed carrier map. Step 4 is the storage destructuring step. Step 6 is the readout: for a query q and row i, it evaluates a score $s_K(q,i)$ after regenerating the row-specific inverse path from the correct key. In simplified algebraic form, one stored wave path is

$$T_{K,r}(x) = P_{K,r}\big((\xi_{K,r} \odot \alpha_{K,r}) \odot x\big),$$

where $P_{K,r}$ is a row-specific permutation, $\xi_{K,r} \in \{\pm 1\}^d$ is a sign mask, and $\alpha_{K,r} \in \mathbb{R}^d_{>0}$ is a strictly positive envelope. Positivity is the only condition needed for coordinatewise division; no positive-definite matrix assumption is involved. The authorized inverse exists for the same derived path:

$$\widehat{x} = (\xi_{K,r} \odot \alpha_{K,r})^{-1} \odot P_{K,r}^{-1}(T_{K,r}(x)).$$

The implementation score used in the reported experiments has the form

$$\alpha_0 \cdot \text{coherence} + 0.10 \cdot \text{energy} + 0.10 \cdot \text{mode\_support} + (1-\alpha_0) \cdot \text{base\_overlap}^2,$$

with $\alpha_0 = 0.46$. This is a retrieval scoring rule, not the security definition, and the weights are fixed implementation parameters rather than a probability simplex. The coherence term is the phase-alignment signal; the energy and mode-support terms stabilize routing; the squared base-overlap term is the "base squaring" component used in the ablation. These auxiliary terms create a small key-independent offset, empirically around 0.012. The value 0.46 is not a theorem constant. It is the fixed implementation weight used for the reported benchmark suite after tuning the recall/isolation tradeoff on development runs.

3.2 Why phases gate meaning

The core mechanism is phase alignment. With the correct key, query and document waves share phase structure and add coherently. With a wrong key, phase differences behave like pseudorandom directions, so the coherence term cancels. The wrong-key result is not a weaker version of the right answer; it is a score path with no usable semantic ranking signal under the tested operating points.

Mathematically, a wrong readout key $K' \neq K$ inserts relative phase and scramble mismatch into the coherence term:

$$C_{K'}(q,i) = \frac{1}{M}\sum_{r=1}^{M} \left| \sum_j A_{q,i,r,j} \exp(\mathrm{i}\Delta_{K,K'}(q,i,r,j)) \right|^2 .$$

Under the modeled wrong-key path, the phases $\Delta_{K,K'}$ behave as high-entropy offsets relative to the authorized carrier. The phasor sum therefore concentrates as residual interference rather than as the authorized semantic overlap. This is the KPT collapse mechanism: the payload cipher may be changed or removed, but without the correct retrieval key the wave score does not preserve a useful semantic ranking signal.

The controlled ablation study in Section 5.2 identifies phase coherence as the isolation mechanism. Orthogonal projection alone gives no isolation. Phase rotation alone scales strongly with dimension. Disabling the decoy floor increases isolation but makes retrieval more brittle. The default therefore trades maximum isolation for robust retrieval.

3.3 Per-document scrambling

A shared hidden rotation is insufficient because orthogonal and monomial transforms preserve inner products when applied consistently across all documents. An attacker who sees all stored wave vectors can compute cosine similarity directly and recover the original pair geometry.

KPT therefore uses a row-specific scramble. For each document i, the system derives a permutation P_i and sign vector ξ_i, then stores

$$\widetilde{w}_i = \xi_i \odot P_i(w_i).$$

Since $P_i \neq P_j$ in general, cosine similarity between stored vectors no longer corresponds to plaintext similarity. The score path regenerates the scramble from the key and stable document routing data before computing the authorized overlap.

This stage is not for authorized recovery alone. Recovery would be possible with a shared keyed transform, but pair geometry at rest would still be visible by Proposition 4. The row-specific coordinate system is what breaks direct pairwise comparison between two stored rows before authorization.

In the measured implementation, wave-real pair correlation falls from 0.904 before scrambling to -0.004 after scrambling. All scrambled wave components have pair correlation below 0.01. Legacy and analysis paths that expose unscrambled routing components such as mode_weight and mode_energy retain weak residual signal, around pair correlation 0.07 and ARI 0.09. The v2 storage path therefore treats such routing values as derived scoring metadata rather than plaintext persisted secrets.

3.4 Post-hoc transformations

An attacker may try exponential stretching, PCA, neural inversion, or coordinate-sorted attacks on the stored representation. Per-document scrambling defeats pairwise coordinate operations because each row lives in a different coordinate system. Sign flips destroy sorted-value alignment for signed coordinates. Magnitude-multiset observables are a separate class: if an implementation preserves sorted absolute values, then this paper's inner-product and score-correlation claims do not cover that channel. The current handoff therefore treats magnitude leakage as a distinct implementation-scope risk, not as something resolved by the theorem chain below. For the reported chosen-plaintext reproduction, even with 500 chosen pairs, an MLP inversion achieves only cosine 0.005 on scrambled wave vectors.

Score compression is also part of the mechanism. Correct-key scores cluster in a narrow range around 0.557, while wrong-key scores are nearly flat around 0.012. Server-side monotone stretching would amplify information available to a score-oracle attacker. Client-side display normalization after authenticated retrieval is safe because it occurs after authorization.

3.5 Key derivation

The implementation uses HKDF with SHA-512 and separate domain parameters for projection, routing, phase shifts, and scrambling. This gives domain-separated key material and supports

arbitrary input key lengths. Effective security is bounded by the entropy of the supplied key and by the seed derivation, not by the continuous phase volume alone.

4 Mathematical Guarantees

4.1 Authorized unbinding

Proposition 1 (Exact authorized unbinding)**.** *Let $m = \xi \odot \alpha$ with $\xi_i \in \{\pm 1\}$ and $\alpha_i > 0$. Define*

$$\mathrm{Bind}_m(x)_i = x_i m_i, \qquad \mathrm{Unbind}_m(y)_i = y_i / m_i.$$

If $m_i \neq 0$ for every coordinate, then

$$\mathrm{Unbind}_m(\mathrm{Bind}_m(x)) = x.$$

Proof. Coordinatewise division cancels coordinatewise multiplication. □

Here m is the complete authorized coordinate mask: the sign vector ξ and the strictly positive envelope α are inverted together.

Proposition 2 (Exact recovery after row-specific scrambling)**.** *Let P be a permutation and P^{-1} its inverse. Then*

$$\mathrm{Unbind}_m\Big(P^{-1}\left(P(\mathrm{Bind}_m(x))\right)\Big) = x.$$

Proof. The inverse permutation restores coordinate order, and the previous proposition cancels the mask. □

These propositions establish that destructuring storage and preserving authorized search are compatible. The stored row can live in a keyed coordinate system, while the authorized query regenerates the inverse path.

4.2 Wrong-key residuals and decision bounds

Proposition 3 (Wrong-key residual decomposition)**.** *For masks m and n,*

$$\mathrm{Unbind}_m(\mathrm{Bind}_m(x) + \mathrm{Bind}_n(z)) = x + z \odot (n/m).$$

If $m = \xi \odot \alpha$ and $n = \xi' \odot \alpha'$, then

$$n/m = (\xi'/\xi) \odot (\alpha'/\alpha).$$

Proof. Coordinatewise division gives

$$\frac{x_i m_i + z_i n_i}{m_i} = x_i + z_i \frac{n_i}{m_i}.$$

Expanding sign and envelope factors gives the quotient expression. □

The wrong-key path is therefore structured interference. If the sign quotient ξ'/ξ behaves like independent signs and the envelope quotient α'/α is controlled, the residual term averages like noise in the score path; the observed wrong-key score offset is around 0.012 in the reported implementation. The companion wrong-key manuscript develops the probabilistic tail, false-accept, hard-decision, and prefix-process bounds for this residual family [8]. Once a clean lower bound and a wrong-key upper bound are certified, a threshold between them gives a decision-facing non-recovery certificate.

Operationally, this proposition says what remains after the correct inverse is applied to a mismatched component: the clean term survives, and the mismatched term is multiplied by a sign quotient and an envelope quotient. Later probability bounds must control that quotient family; the algebra alone is not a security proof.

4.3 Shared-transform failure

Proposition 4 (A shared monomial transform does not hide pair geometry)**.** *Let P be a permutation and $\xi \in \{\pm 1\}^d$. Define*

$$U(x) = \xi \odot P(x).$$

ξ is the same sign-mask object used above, but here it is shared globally rather than derived per row. Then for all $x, y \in \mathbb{R}^d$,

$$\langle U(x), U(y) \rangle = \langle x, y \rangle .$$

Proof. Permutations only reorder coordinates and shared sign flips cancel:

$$\sum_i (\xi_i x_{P(i)})(\xi_i y_{P(i)}) = \sum_i x_{P(i)} y_{P(i)} = \sum_i x_i y_i.$$

□

This proposition is the reason row-specific scrambling is required. A single global hidden monomial map would leave pair geometry visible at rest.

4.4 Phase decay and geometric depth

The sinc-squared phasor decay calculation applies to the coherence component. It is included because it gives the simplest quantitative picture of why phase mismatch suppresses coherent overlap: a small phase radius is close to the correct key, while a broad phase interval cancels coherent sums. In the current proof program, this is the coherence-layer calculation inside a broader wrong-key and margin theory.

Theorem 5 (Coherence phasor decay)**.** *For a uniform phase perturbation $\eta \sim \mathrm{Uniform}[-\delta, \delta]$, the expected phasor overlap is*

$$\mathbb{E}[e^{i\eta}] = \frac{\sin \delta}{\delta}.$$

When the coherence score is quadratic in phasor overlap, the coherence ratio is

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

Proof. The expectation is the Fourier transform of the uniform interval:

$$\mathbb{E}[e^{i\eta}] = \frac{1}{2\delta} \int_{-\delta}^{\delta} e^{it}\, dt = \frac{\sin \delta}{\delta}.$$

The implementation's coherence term is quadratic in the phasor overlap, giving the squared ratio. □

The theorem applies to the coherence component, whose score weight is 0.46 in the reported implementation. The full score also includes support and energy terms, which explain the observed wrong-key offset around 0.012. δ is a phase-radius parameter: small δ means the wrong path is close to the correct phase, and large δ means the wrong path ranges over a wider phase interval. The uniform interval is the clean reference model for this calculation, not a claim that every key-derived phase is literally sampled from this one-dimensional law.

The corresponding geometric key-depth calculation measures the continuous phase volume, not the entropy of the actual key.

$$B = D \log_2(\pi/\delta_c),$$

Configuration	D	Geometric depth
$d = 96, M = 8$	768	156 bits
$d = 384, M = 8$	3072	623 bits
$d = 768, M = 8$	6144	1245 bits

Table 1: Continuous phase-volume budget for three representative embedding widths with $M = 8$ modes. This is a geometry budget, not the entropy of the actual user key.

where $D = Md$ and δ_c is the critical phase radius defined by the chosen target coherence ratio. The table uses the same fixed target ratio for all rows; changing that target changes the bit figures. The following configurations are representative small, baseline, and large embedding widths for the $M = 8$ implementation:

This is not the same as key security. All phase parameters are derived from key material. The effective security is bounded by

$$\min(H(K), 256, B),$$

so a random 32-byte key is capped at 256 bits, while a 20-character alphanumeric password is roughly 119 bits. The 623-bit figure shows that the phase geometry is not the bottleneck for the $d = 384, M = 8$ configuration.

4.5 Dimension budget

The 623-bit entry above is not a dimension threshold. It is the phase-volume budget for the baseline $d = 384, M = 8$ configuration under the chosen coherence target. A deployment-facing dimension threshold is instead the smallest embedding width d that satisfies the selected coherence, decorrelation, and routing-stability budgets. Below such a bound, the same architecture cannot simultaneously claim leakage reduction, key-gated separation, and reliable routing under the same target budgets.

4.6 Ranking under monotone readout

Proposition 6 (Monotone readout invariance)**.** *Let $a_1, \ldots, a_M \in \mathbb{R}$ and let $\phi : \mathbb{R} \to \mathbb{R}$ be strictly increasing. Then ϕ preserves the finite ranking induced by $\{a_1, \ldots, a_M\}$, including argmax and top-k membership under fixed tie-breaking.*

Because ϕ is strictly increasing, it preserves every pairwise order comparison. This is why KPT scores should be treated as ordinal retrieval scores. Absolute score units may be compressed, but ranking and margin are the semantic objects used by retrieval.

5 Experiments

5.1 Setup

The benchmark suite uses AG News, 20 Newsgroups, and MS MARCO with sentence-transformer embeddings of dimension 384. Each 6k experiment uses 6,000 documents, 1,200 queries, and three random seeds. DCPE/SAP is evaluated at $\sigma \in \{0.05, 0.10, 0.20\}$. The 100,000-document scaling run uses AG News with the same embedding dimension. Unless otherwise stated, pair correlations are Pearson correlations over sampled document pairs from the same evaluation split.

Dataset	Type	Docs / Queries	Vectorizer
AG News	4-class classification	6,000 / 1,200	sentence-transformers
20 Newsgroups	20-class classification	6,000 / 1,200	sentence-transformers
MS MARCO	Passage retrieval	6,000 / 1,200	sentence-transformers

Table 2: Balanced 6k benchmark setup. Each dataset uses 6,000 indexed documents, 1,200 queries, and three seeds unless stated otherwise.

5.2 Ablation

The ablation disables individual components of the KPT transform and scoring pipeline across $d \in \{64, 128, 256, 384, 512, 768, 1024\}$. The main finding is that phase coherence is the isolation mechanism, while the decoy floor is the primary limiter. The table entries are isolation ratios between correct-key and wrong-key top-k score magnitudes; higher is more isolation, not necessarily better retrieval utility. The tested removals map to Section 3.1 as follows: top-k routing is the sparse routing step, base squaring is the base_overlap^2 score term, decoy floor is the auxiliary routing floor, all nonlinearities removes the nonlinear scoring stack, and pure phase keeps only the coherence component.

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

Table 3: Ablation of the implementation pipeline from Section 3.1. Disabling "top-k routing" removes the sparse routing step, while disabling "base squaring" removes the squared base-overlap term in the score. The 1024-D decoy-floor result shows that removing the floor can sharply increase isolation while making retrieval less robust.

5.3 KPT versus DCPE/SAP

DCPE/SAP denotes the distance-comparison-preserving baseline from [2]; σ is its tested noise parameter.

	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 $\sigma = 0.05$	0.393	0.763	0.340	0.911	0.379	0.714
DCPE $\sigma = 0.10$	0.076	0.397	0.095	0.702	0.063	0.345
DCPE $\sigma = 0.20$	0.006	0.135	0.013	0.326	0.006	0.111

Table 4: Retrieval utility versus pair-geometry leakage. Recall is Recall@10; pair_corr is the Pearson correlation between plaintext pair similarities and the observable pair similarities for the tested representation.

KPT achieves Recall@10 at least 0.906 with pair correlation at most 0.029. DCPE/SAP does not reach Recall@10 above 0.4 while keeping pair correlation below 0.7 at the tested noise

levels. DCPE/SAP has a continuous noise parameter, so untested intermediate points may exist, but the mechanism is designed to preserve distances and therefore has an inherent leakage floor.

5.4 Key sensitivity

Twelve key variants at Levenshtein distances 0 through 27 were tested. Distance 0 gives full retrieval. Distance at least 1 gives complete collapse, with isolation ratios between 48x and 228x. Attack success is not correlated with edit distance; the transition is qualitative rather than gradual.

In a random-key validation with 10,000 keys, the wrong-key score has $\mu = 0.0065$, $\sigma = 0.0044$, and maximum 0.0196, compared with a correct-key score 0.5577. The gap is about 124σ.

5.5 Multi-user demonstration

In a 500-document AG News demonstration, Alice's key retrieves the business document relevant to "Stock market crash" with score 0.558. Bob's wrong key returns an unrelated world-news document with score 0.012. Public-only access returns a random document with score 0.060. The same stored corpus is therefore searched in different semantic spaces depending on the key.

5.6 Scaling to 100,000 documents

Scale	KPT R@10	KPT corr	enc/doc	DCPE R@10	DCPE 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

Table 5: AG News scaling run. The KPT pair-correlation column reports the measured observable pair correlation under the same sequential sampling protocol; enc/doc is CPU encoding time per document.

KPT recall remains stable and slightly improves at scale. Under controlled sequential sampling, pair correlation remains around 0.001 to 0.002, consistent with the larger three-seed balanced-evaluation bound 0.029.

5.7 Residual routing observables

The score path needs stable candidate structure after scrambling. Routing features such as mode weights and mode energies describe that structure; they are metadata, not payload text and not standalone security primitives. A PCA pass on routing features can describe linear dimension, but it does not by itself imply low leakage, mutual-information control, or pair-correlation control. For that reason the PCA result is not used as a security claim in this handoff.

The residual risk is instead reported directly where it matters. Legacy or analysis paths that expose mode weight and mode energy retain weak pair and cluster signal, around pair correlation 0.07 and ARI 0.09. The v2 storage path avoids treating those values as persisted plaintext side information, but the broader routing-observable family remains an implementation-scope leakage surface to reduce or formally bound.

5.8 Vectorizer independence

KPT was tested across three embedding models without retraining or parameter adjustment.

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

Table 6: Vectorizer robustness check. The same KPT parameters were used without retraining or model-specific adjustment.

The embedding model dominates wall-clock time. The reported transform-only encoding overhead is about 0.05 ms per document in the local CPU benchmark; end-to-end latency also depends on embedding inference, filtering, batching, and hardware.

5.9 Quantization

Production vector databases often quantize embeddings. We quantize wave vectors while keeping mode_weight at float32, because it is used by the routing and score path. This experiment is included because score compression makes fine ordering sensitive to low precision: the mean score can stay stable while Recall@10 collapses.

Precision	Recall@10	Score mean	Storage/doc
float64	0.980	0.5566	54.1 KB
float32	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

Table 7: Quantization sensitivity for the wave representation. Mean score alone is not sufficient; retrieval ranking is the relevant utility object. Storage/doc reports the tested wave-representation footprint, not a fully optimized record format.

int16 fixed-point preserves retrieval at half the float32 storage. At 12 bits and below, and for float16 in this implementation, ranking degrades because correct-key scores are compressed within a narrow margin. The mean score remains stable, but fine-grained ordering is lost.

5.10 Key-gated clustering

This benchmark tests one downstream task that depends on pairwise similarity: KMeans on AG News score-matrix rows for 1,000 documents, using class labels as the external clustering reference. It gives:

Space	ARI	NMI
Plaintext	1.000	1.000
Authorized correct-key scores	1.000	1.000
Wrong-key scores	0.040	0.125
Routing side information	0.001	0.021
Scrambled wave at rest	-0.000	0.017

Table 8: Downstream clustering on pairwise score-derived features. Authorized scores preserve the task structure; wrong-key, routing-only, and scrambled at-rest views do not recover useful clusters in this benchmark.

The kNN neighbor overlap is 89.7% with the correct key and 0.9% with a wrong key, close to the random baseline at $k = 10$.

6 Adversarial Analysis

Two independent red-team analyses, one cryptographic and one ML-oriented, were translated into empirical reproductions on 2,000 documents. The reproductions use the same threat split as Section 2.2: the attacker sees stored artifacts or wrong-key scores, but not the retrieval key or plaintext embeddings. We report correlation, cosine, ARI, and kNN overlap against random or plaintext baselines; "no recovery" means the reproduced attack did not recover useful semantic geometry under the stated sample and query budgets.

6.1 Known-plaintext attack

The attack trains a neural model from plaintext/encrypted pairs. The reproduced MLP reaches cosine 0.228 at $N = 800$ and similarity correlation 0.048. This did not recover useful semantic geometry in the tested regime.

6.2 Score-oracle attack

The attack attempts to reconstruct a Gram matrix from cross-key score profiles: score vectors produced by replaying queries through wrong-key readouts. With 200 queries over 2,000 documents, Gram-matrix correlation is 0.024, and kNN overlap is 0.006 versus a random baseline 0.005. Wrong-key score variance is around 10^{-12}, leaving little signal to aggregate in the tested regime. Although absolute variance is small, rankings can still expose stable document offsets because top-k disclosure depends on relative order rather than calibrated score magnitude.

We also separate wrong-key semantic failure from wrong-key distribution leakage. Wrong-key scores are not assumed to induce a uniformly random ranking over documents: a readout can contain a document-dependent offset, so some documents act as low-score hubs under many wrong keys or unrelated queries. The corresponding hubness diagnostic counts top-k frequencies

$$h_i = \Pr_{q \sim \mathcal{Q},\, K' \sim \mathcal{W}}[i \in \mathrm{Top}_k(S_{K'}(q, \cdot))]$$

over wrong-key readouts. The relevant summary statistics are entropy $H(h)$, top-k frequency skew, correlation with plaintext degree, correlation with labels or clusters, and kNN overlap induced by the hubness ranking. These values are oracle leakage diagnostics, not evidence of semantic recovery unless they correlate with plaintext semantic neighborhoods.

Hubness diagnostic	Local smoke value
Normalized entropy $H(h)/\log n$	0.997
Top-k frequency skew	1.36
Correlation with plaintext degree	-0.212
Label/cluster NMI from hubness bins	0.213
Plaintext-degree top-k overlap	0.000

Table 9: Wrong-key hubness smoke diagnostic on the local conformance corpus ($n = 24$, 8 queries, 100 wrong keys, $k = 3$). The table measures the distribution-leakage surface introduced above; it is not used as the large-scale semantic-recovery claim, which is reported by the 2,000-document score-oracle numbers.

6.3 Statistical leakage

The attack argues that a small pair correlation becomes significant at scale. A controlled sweep $N = 50$ to 2,000 on routing-derived side information shows pair correlation oscillating around 0.00 ± 0.01. Routing-only homophily matches random chance. The caveat is that legacy exposed mode weight and mode energy retain weak residual signal, documented above.

6.4 Chosen plaintext

Three chosen-plaintext strategies were tested with 500 chosen pairs: least-squares reconstruction gives cosine 0.001 on wave-real; cluster injection gives ARI -0.002; and MLP inversion gives cosine 0.005. The per-document scramble makes the mapping document-dependent and prevents generalization.

6.5 Collusion

Two users with different keys combine protected views of the same documents. The strategies below concatenate side channels, concatenate wave vectors, use coordinatewise wave products, or apply canonical-correlation alignment (CCA).

Strategy	Correlation
Concatenate routing side information	-0.002
Concatenate wave vectors	0.001
Wave product	0.000
CCA alignment	-0.001

Without per-document scrambling, concatenating wave vectors yielded correlation 0.944, a critical vulnerability. Row-specific scrambling reduces this to noise because each key produces a different per-document coordinate system.

7 Limitations

These limitations are scoped to the storage model and experiments above. They are not exceptions to the headline claim; they are the boundary between the retrieval-semantics protection studied here and the surrounding deployment system.

No full reduction. The paper provides analytical guarantees, formalized proof components, and reproduced attacks, but not a full reduction from end-to-end KPT security to a standard hard problem.

Score compression. Correct-key scores cluster at approximately 0.557 ± 0.001. Retrieval remains good, but absolute score differences are small. KPT scores should be treated as ordinal rankings rather than calibrated cardinal similarities.

Cross-key score bias. Wrong-key scores are not assumed to induce a uniformly random ranking over documents. In the current implementation, wrong-key readouts can contain a document-dependent offset: some documents act as low-score hubs and appear near the top under many wrong keys or unrelated queries. This effect does not by itself imply semantic recovery; in the reproduced attacks of Section 6, wrong-key hubness did not correlate enough with plaintext semantic neighborhoods to recover useful rankings or clusters. However, it is a separate oracle observable. An adaptive adversary could aggregate many wrong-key or unrelated-query readouts to estimate hubness, routing artifacts, norm/energy artifacts, or tenant/corpus

structure. We therefore treat cross-key hubness as an operational leakage surface rather than as part of the protected semantic readout. Deployments should rate-limit wrong-key score access, avoid exposing raw wrong-key scores, apply tenant filters before KPT scoring, and monitor repeated failed readout attempts.

Mode routing leakage. mode_weight and mode_energy are not scrambled in the same way as wave channels in legacy or analysis paths. They show weak residual cluster signal because similar documents route through similar modes; see Section 5.7.

Magnitude observables. The theorem chain and reproduced attacks focus on score geometry, pair correlation, and signed coordinate representations. They do not bound every observable of the stored row. In particular, any implementation that preserves the sorted multiset of coordinate magnitudes exposes a distinct attack surface that must be closed by the implementation or stated outside the claim. The v2 path adds a secret preconditioning/scrambling layer for this channel, but the paper does not yet provide a full theorem for all magnitude observables.

Private side artifacts. KPT protects the retrieval representation only under the stated storage model. If a companion private codebook, hidden segment file, seed file, or key-bearing artifact stores plaintext, then that artifact must be encrypted and managed as secret material. The KPT transform does not protect data that is separately serialized in plaintext.

Precision sensitivity. KPT requires at least 16-bit precision for wave vectors in the tested implementation. mode_weight remains float32 in the reported quantization run because it participates in routing and score stabilization; the stored-size comparison therefore applies to wave-vector precision, not to a fully quantized record format.

Operational assumptions. KPT does not protect traffic patterns, metadata, unrestricted score-oracle access, or compromised keys. Rate limits and ordinary access controls remain deployment requirements.

8 Discussion

KPT is access control over retrieval semantics. The same stored object supports authorized search under the correct key and denies useful semantic search under the wrong key. No secondary plaintext index is required, and the score operator works directly on the stored keyed representation.

The comparison with AES is therefore not "KPT versus encryption." AES is one standard choice for protecting payload bytes; another authenticated cipher could serve the same role. KPT protects retrieval semantics. With payload encryption alone, semantic search requires decrypt-all scanning or trusted execution. With KPT alone, payload bytes are not protected. A practical deployment composes KPT with a separate payload-protection layer.

The security-utility tradeoff is controlled partly by the decoy floor. The default supports robust retrieval with isolation ratio around 47x. Removing the floor can increase isolation dramatically, up to 15,905x at $d = 1024$, but makes retrieval more brittle. This is analogous to a noise parameter, but on a different axis: KPT trades coherence reserve against isolation.

The same mechanism should apply beyond top-k search to operations that depend on pairwise similarity. This paper tests clustering as one such downstream task; recommendation, deduplication, anomaly detection, topic tracking, and RAG context selection remain deployment hypotheses. In a multi-tenant system, different users can search the same physical index with different keys and observe different semantic spaces.

Property	AES-GCM	KPT	DCPE/SAP	Compass
Search without decryption	No	Yes	Yes	Yes
Recall@10 at 100k	N/A	0.993	0.352	~ 1.0
pair_corr	0	≤ 0.029	≥ 0.714	0
Formal proof	Yes	search-core stack	Yes (RoR)	Yes
Overhead	decrypt-all	0.05 ms/doc	< 1 ms/doc	10–100x
Vectorizer agnostic	Yes	Yes	Yes	No

Table 10: High-level comparison. KPT values are measurements from this paper; DCPE/SAP and Compass entries mix cited design properties with the tested baseline where available. Compass security and overhead are not the same observable as KPT's at-rest pair-correlation metric.

Brute-force KPT scoring is linear in the number of candidate documents. In multi-tenant deployments, a user or tenant filter can restrict scoring to the authorized subset. Larger deployments should use a two-stage design: coarse authorized filtering followed by KPT scoring on the candidate set. End-to-end latency is hardware-, embedding-, batching-, and index-dependent, so the paper reports transform overhead rather than a deployment latency claim.

9 Conclusion

We present KPT as a key-bound retrieval layer for protected vector data. The correct key restores the geometry needed for semantic ranking; wrong keys produce structured residual interference instead of a usable semantic readout; and row-specific scrambling prevents stored wave channels from preserving obvious pair geometry at rest.

The empirical results match the proposed mechanism. Across three datasets, KPT achieves Recall@10 at least 0.906 with at-rest observable pair correlation at most 0.029. At 100,000 documents, Recall@10 reaches 0.993 with pair correlation 0.002. Per-document scrambling reduces at-rest wave leakage from 0.904 to -0.004, and five reproduced attacks did not recover useful semantic geometry in the tested regimes.

KPT should be deployed as part of a layered system: an authenticated cipher protects payload content, KPT protects retrieval semantics, and operational controls protect keys, metadata, and query access. In that role, it gives a lightweight, mathematically grounded way to make stored embeddings searchable with the right key and semantically unusable without it.

References

[1] Cao et al. *OSNIP: Obfuscated semantic null space injection*, 2026.

[2] Georg Fuchsbauer, Riddhi Ghosal, Adam O'Neill, and Krzysztof Pietrzak. *Approximate distance-comparison-preserving symmetric encryption*. Security and Cryptography for Networks, 2022.

[3] Jiaqi He et al. *STEER: Transform before you query*, 2025.

[4] Yifan Liu et al. *Privacy-preserving approximate nearest neighbor search*, 2025.

[5] John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. *Text Embeddings Reveal (Almost) As Much As Text*. EMNLP 2023; arXiv:2310.06816. https://arxiv.org/abs/2310.06816

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

[7] Nino Wagensonner. *KPT-HGA v1: Permutation-Scrambled Scaled Rademacher Unbinding*. Companion manuscript, 2026.

[8] Nino Wagensonner. *Foundations of Wrong-Key Laws for Keyed Search: Kernel Geometry, Difference Profiles, and Adaptive Security*. Companion manuscript, 2026.

[9] Jinhao Zhu, Liana Patel, Matei Zaharia, and Raluca Ada Popa. *Compass: Encrypted Semantic Search with High Accuracy*. IACR ePrint 2024/1255, 2024. https://eprint.iacr.org/2024/1255

[10] Wei Zou, Isack Lee, Yangsibo Huang, Peter Hartigan, and Yunfang Chen. *PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models*. USENIX Security Symposium, 2025. https://www.usenix.org/conference/usenixsecurity25/presentation/zou