Pointing the Way, Hiding the Destination: Practical Private Dense Retrieval at Scale

Peichun Hua[1,2], Danyang Chen[1], Junan Zhang[1], Haifeng Sun[3], Jingyu Wang[3],
Diwen Xue[4], Mingyu Li[5], Yunming Xiao[1,2]

[1]*The Chinese University of Hong Kong, Shenzhen*
[2]*State Key Laboratory of Internet Architecture, Tsinghua University*
[3]*Beijing University of Posts and Telecommunications* [4]*The Chinese University of Hong Kong*
[5]*Institute of Software, Chinese Academy of Sciences*

Abstract

Hosted retrieval-augmented generation (RAG) and semantic search allow users to query valuable provider-held corpora, raising two competing demands: to hide each query and chosen result, yet reveal only the documents that the user is authorized to receive. Existing cryptographic approaches either make this costly by processing the entire corpus for every query, or sacrifice quality for efficiency by scanning a few clusters. We repurpose learned deep hashing as a private filter: a randomized binary code points the provider to a short candidate list, while encrypted reranking and oblivious key transfer protect the precise query and final selection. This shortlist short-circuits full-corpus cryptographic search without sacrificing retrieval quality: with 200–500 candidates, it closely matches full-corpus retrieval across five zero-shot corpora spanning 25K to 5.4M documents. On the full 2.68M-passage NQ corpus over a 10-Gbps link, our protocol only adds 0.73 seconds, or 10%, to a 128-token Qwen3-32B RAG pipeline. The released code satisfies directional metric differential privacy (DP) and substantially reduces embedding-inversion and property-inference leakage, demonstrating that a carefully learned shortlist can make private dense retrieval both accurate and practical.

1 Introduction

The proliferation of Retrieval-Augmented Generation (RAG) has transformed large language models (LLMs) from static artifacts into dynamic, knowledge-grounded reasoning engines [25, 27, 47, 67]. At the core of RAG lies dense retrieval: mapping documents and queries into a high-dimensional vector space and searching via nearest neighbors. As organizations deploy RAG over sensitive corpora such as legal archives, medical records, and proprietary knowledge bases, both the document collection and user queries become sensitive assets [14, 62, 74, 88]. Recent embedding inversion attacks [12, 35, 49, 61, 73, 90] have demonstrated that continuous embeddings are *not* opaque fingerprints, but rich, invertible semantic representations from which an adversary can reconstruct sensitive source text and infer private attributes.

Existing cryptographic approaches to private retrieval struggle to balance security with the scale and dimensionality of modern dense embeddings. Homomorphic encryption (HE) and multi-party computation (MPC) protocols [14, 48, 59] incur orders-of-magnitude of computational overhead, while ORAM-based methods [96] require multiple interactions and high user-side computation and storage [19]. Trusted execution environments (TEEs) avoid homomorphic scoring, while access-pattern [84] and microarchitectural leakage [6] remain.

This work. Our central design choice is to expose a coarse candidate pattern under directional metric differential privacy (mDP) [37] and reserve cryptographic computation for the resulting shortlist. An honest-but-curious *Corpus Owner* keeps its proprietary corpus with full-precision embeddings while additionally maintaining a lightweight binary index; an authorized *User* randomizes released hash codes that satisfy stringent mDP guarantees and encrypts the clean query representation. The resulting code makes the nearby query directions induce similar candidate-pattern distributions, making them hard to distinguish from one another and thereby protecting critical detailed privacy information (§3.3). The *Owner* performs a Hamming search on the corpus (N) and packed Brakerski–Fan–Vercauteren (BFV) homomorphic scoring over the resulting K candidates. The *User* decrypts the scores, selects locally, and uses active-secure k-out-of-K oblivious transfer (OT) to open at most k payloads without revealing its choices or accessing the other $K-k$ payloads.

Recovering retrieval quality under DP may require scoring $K = 500$–$3{,}000$ candidates, while a RAG request typically releases and bills only $k = 3$–10 documents. The gap between K and k makes OT central to this deployment model. Returning the shortlist would disclose hundreds of times more content; requesting the final documents directly would reveal the User's selection. Instead, our k-out-of-K transfer binds each round to the billable result count, while authenticated accounts and cumulative quotas govern repeated extraction. In our model, we assume that the *User* might want to actively

learn the payloads beyond its billed k per round and design our protocol to handle such threats (§6.4; Theorem 5).

Three empirical insights make our design highly practical. ❶ **Candidate-set redundancy persists across scales and domains.** A learned binary filter (Stage 1) needs only to preserve the documents that matter for the final ranking (Stage 2), while the ranking among the top documents is not critical. At $K = 500$, the two-stage pipeline (§5) retains 98.84%–100% of full-corpus NDCG@10 (ranking quality) across five zero-shot BEIR corpora [75] on both evaluated encoders. In particular, on Climate-FEVER with 5.4M documents, the shortlist excludes noisy high-scoring documents and raises NDCG@10 by 0.0012–0.0158. Even under DP protection, expanding the candidate pool up to 3,000 achieves a good balance between efficiency and search quality (§7.2.1). These corpora span scientific search, question answering, entity retrieval, and fact verification at scales from 25K to 5.4M documents (Figure 1), demonstrating strong generalizability of our approach.

❷ **Dense embeddings contain substantial precision redundancy.** We find that scalar int8 quantization preserves the pretrained ranking closely; our symmetric zero-point-free realization converts floating-point similarity into exact, low-depth integer dot products. Without this precision reduction, private real-valued scoring relies on approximate-number HE such as CKKS [15], higher-precision fixed-point BFV [40], or interactive fixed-point MPC [60]; int8 enables exact packed BFV scoring with a small plaintext modulus while preserving high ranking quality.

❸ **Model separation creates pipeline parallelism.** We adopt low-rank adaptation (LoRA) fine-tuning [34] of pretrained encoders to create the learned hash models efficiently. This ensures that the hash codes can select a high-recall subset without incurring the storage and memory burden of an entirely new model. Our protocol implementation carefully overlaps the pretrained scoring forward and BFV query encryption with Owner-side Hamming shortlist construction, then streams the candidate ciphertexts while encrypted scoring and key transfer advance. This successfully hides the latency and makes our protocol efficient under different network bandwidths.

We investigate randomized response, Gaussian, Rényi-DP-calibrated von Mises–Fisher (RDP-vMF), and exactly calibrated pure-vMF mechanisms that achieve mDP, extensively evaluating their privacy–utility trade-offs through retrieval quality, representation attacks, protocol latency, and end-to-end RAG. In summary, we make the following contributions:

1. **A leakage-aware retrieval protocol and security contract** that reveals a metric-DP candidate pattern, confines encrypted exact scoring to K candidates, protects the clean query against the Owner, and uses active-secure k-out-of-K OT to hide the final selection and bound payload recovery.
2. **A learned candidate-filter design and training recipe** that retrofits pretrained encoders via LoRA while preserving the original model as the Stage 2 scorer. It sustains near-lossless zero-shot retrieval across domains and corpus scales and improves over direct quantization, classical hashing, and supervised binary-retrieval baselines.
3. **A directional randomization mechanism and analysis** that gives the released shortlist code pure metric-DP protection, carries the guarantee through candidate generation by post-processing, and establishes its retrieval–privacy frontier against randomized response, Gaussian, and RDP-vMF mechanisms and representation attacks.
4. **A practical two-forward system realization** that combines int8 quantization, shallow packed BFV, key-only OT, and pipeline overlap in a networked prototype, reducing cryptographic work from N documents to a high-recall shortlist while retaining efficient million-scale retrieval and end-to-end RAG.

2 Background and Related Work

2.1 Dense Retrieval and Learned Hashing

RAG grounds language-model generation in retrieved evidence [25, 27, 67]. Its scalable first stage is usually a bi-encoder such as DPR, SBERT, E5, or BGE [11, 41, 69, 78]: queries and documents are encoded independently, then compared by inner product or cosine similarity [69]. Cross-encoders, instead, score query–document pairs jointly and are commonly reserved for reranking because their cost grows with the number of evaluated pairs [13, 16, 68].

Binary hashing compresses continuous representations into L-bit codes and replaces floating-point distance with XOR and popcount [56]. Classical methods include data-independent random-hyperplane, multi-probe, and cross-polytope LSH variants [2, 7, 57, 66], and data-dependent rotations such as

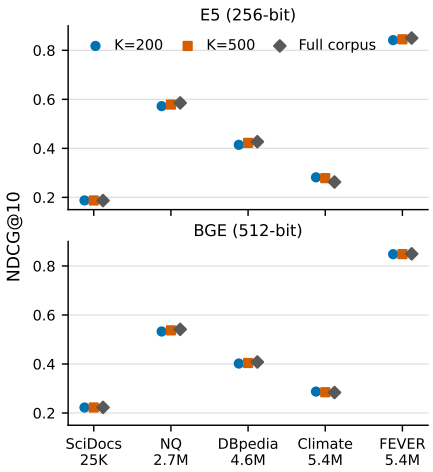

Figure 1: Cross-dataset quality of our pipeline. Each panel reports absolute Stage 2 NDCG@10 at $K \in \{200, 500\}$ alongside full-corpus pretrained retrieval. The five BEIR corpora span 25K–5.4M documents and four retrieval domains.

ITQ and IsoHash [28, 43]. Deep models jointly learn the representation and code: earlier work emphasized image retrieval with pairwise or center-based objectives [51, 53, 79, 89, 95], ranking-aware hashing directly optimized tie-aware AP and NDCG in Hamming space [30], and recent objectives couple discrimination with quantization or adapt transformer encoders to text [31, 33, 65]. Instead of preserving embedding geometry and search quality, we use a learned code only as a high-recall first-stage filter and retain the continuous representation for exact candidate reranking. This division gives the Owner a lightweight coarse index while reducing private similarity evaluation from all N documents to K candidates.

2.2 Representation Privacy

Dense embeddings preserve enough lexical and semantic information to support generation-based, search-based, and property-inference attacks [12, 35, 49, 61, 73, 90]. Our design randomizes the normalized pre-binarization representation under metric DP [37, 83] and then deterministically binarizes it; the released code inherits the same privacy bound by post-processing [23].

Dense retrievers normalize embeddings and rank them by cosine similarity, so their semantic neighborhoods lie naturally on the unit sphere and are measured by angle. We therefore instantiate metric DP over nearby directions of the learned coarse representation [8, 22, 23].

Definition 1 ((ε,δ)-Directional Privacy)**.** *Let d be a metric on the unit sphere and let $\rho > 0$. A randomized mechanism $\mathcal{M} : \mathbb{S}^{L-1} \to \mathcal{Y}$ satisfies (ε,δ)-directional privacy at radius ρ if, for every $u, u' \in \mathbb{S}^{L-1}$ with $d(u,u') \le \rho$ and every measurable $S \subseteq \mathcal{Y}$,*

$$\Pr[\mathcal{M}(u) \in S] \le e^{\varepsilon} \Pr[\mathcal{M}(u') \in S] + \delta. \tag{1}$$

This radius-bounded definition instantiates metric privacy [8, 83] on normalized directions. We use angular distance $d_\theta(u,u') = \arccos(u^\top u')$ as the primary metric and chord distance $d_c(u,u') = \|u-u'\|_2 = 2\sin(d_\theta(u,u')/2)$ when deriving the vMF density-ratio bound. The radius defines a neighborhood in representation space; an empirical query-pair distance study can further calibrate that neighborhood to paraphrase or intent-level relations.

Prior private-hashing methods randomize discrete codes with randomized response or randomize data-independent LSH functions [19, 26, 81]. Deep hashing instead exposes the continuous pre-sign representation as a natural randomization point, preserving coordinate margins and directional geometry. Gaussian perturbation operates on bounded $h \in [-1,1]^L$ under Euclidean adjacency [4], while von Mises–Fisher (vMF) perturbation operates on normalized $u = h/\|h\|_2$ under angular or chordal adjacency [82]; binarization then preserves their guarantees by post-processing [23].

Randomized response treats every bit alike and must compose privacy across the complete code because nearby continuous vectors can cross many low-margin sign boundaries. At $L = 256$ and $(\varepsilon = 16, \delta = 10^{-6})$, this calibration flips 46.1% of the bits and yields only 0.0306 mean NDCG@10 at $K = 3000$, versus 0.5360 for Gaussian and 0.5367 for RDP-vMF (Table 5).

2.3 Private Retrieval across Deployments

The ownership and trust boundary define the private-retrieval problem in different scenarios [85]. Table 1 organizes prior systems into four deployment models.

In *client-owned outsourcing*, classical searchable encryption [9, 20, 72] and vector systems [19, 54, 96] protect a corpus owner who queries an untrusted cloud. For example, MESS [19] randomizes LSH codes, maintains 64 HNSW shards with each item routed to 16, and reranks the aggregated candidates at the client; this yields 16× indexed-record replication. Retrieving the original documents from the cloud additionally requires private information retrieval (PIR) [17, 77] to hide the access pattern.

In the *public/shared-corpus* model, Tiptoe, Wally, PACMANN, and Speakeasy protect the query without a corpus-confidentiality goal [3, 32, 45, 94]. For example, Tiptoe [32] combines clustering with linearly homomorphic search across 45 servers to operate vector search on web-scale corpus, but its single-cluster pruning for efficiency significantly impacts the search quality compared to normal embedding search. PACMANN [94] improves this quality–latency tradeoff through graph search and client-preprocessing PIR, but with 100 million vectors, each client downloads 59.6 GB during setup, stores 2.9 GB of state, and exchanges another 399.4 MB to maintain that state after every query.

In *secret-shared outsourcing*, PRAG and P^{2}RAG protect distinct data owners and queriers by placing the database across MPC servers [59, 97]; PRAG assumes an honest majority, whereas P^{2}RAG uses two semi-honest non-colluding servers and avoids secure sorting through interactive bisection, but requires full-corpus work and trusted-dealer preprocessing, the cost of which is excluded from its benchmark.

Our target is the fourth model, a *provider-held proprietary corpus* serving an external querier [10, 14, 50, 52, 71]. Pisces [52] combines an oblivious SimHash filter with MPC scoring and PIR-to-share, and adds a cryptographic BM25 path; our learned filter releases a directionally private shortlist and concentrates cryptography on exact candidate scoring and selected payloads. PANTHER [50] co-designs PIR, secret sharing, garbled circuits, and homomorphic encryption for strong single-provider protection, but its cluster-wise PIR representation scales with both vector dimensionality and the number of probes. Its evaluation targets 96- and 128-dimensional vision embeddings, whereas modern RAG commonly uses substantially wider text embeddings and demands

enough probes to preserve near-lossless retrieval quality. In our evaluation (§7.3), PANTHER exceeds the memory limit of our server in a million-scale corpus. Within this model, our protocol keeps the corpus at a single provider for the owner, provides formal DP guarantees for the search pattern, cryptographically protects unselected corpus content from the user, and confines homomorphic scoring to the filtered candidates for efficiency and near-lossless quality.

3 Deployment and Threat Model

We target a proprietary corpus served directly by its Owner to an external authorized User. This section defines the two parties, states the information visible on each side, and introduces the attacks used to measure the released coarse code.

3.1 Parties and Trust Relations

Corpus Owner (Server). Holds the document collection, the binary hash index, the normalized embeddings used as plaintext HE operands, per-document content keys, and document payloads protected by authenticated encryption with associated data (AEAD). The Owner is *honest-but-curious*: it follows the prescribed computation and message schedule while attempting to infer query content, link queries, or profile Users from its protocol view. All server-side retrieval runs on Owner-controlled infrastructure.

User (Client). Holds a query, the agreed encoder and the hash model, the HE secret/public keys, and the DP parameters. The HE scoring guarantee applies to a conforming User that encrypts the prescribed bounded, canonically packed query; the active-secure OT guarantee additionally covers a malicious receiver attempting to recover more than k content keys.

3.2 Views, Leakage, and Assumptions

The protocol has four security goals. Metric DP protects the coarse query code released to the Owner, BFV semantic security protects the clean query from the Owner, ciphertext-simulatable BFV restricts a conforming User's scoring view to the prescribed K exact scores, and active-secure OT hides the User's selected positions while limiting payload-key recovery to k OT choices.

The Owner observes:

- Binary index $\{\mathbf{b}_d\}_{d\in[N]}$ and the resulting candidate set C_K;
- Metric-DP coarse query code $\tilde{\mathbf{b}}_q$;
- The HE ciphertext $\mathsf{Enc}(\bar{\mathbf{q}})$;
- The HE and sender-side OT transcripts;
- Authenticated identity, session and round identifiers, message lengths and timing, K, and the public result count k.

The Owner does *not* observe the User's plaintext query, the decrypted similarity scores, or which specific k indices the User selected via OT.

The User observes:

- The K scalar similarity scores;
- The K AEAD payload ciphertexts and the $k \times K$ masked content-key table for the candidate set;
- At most k content keys and the corresponding plaintext document payloads.

Under conforming execution, the application gives the User candidate-local scores and selected payloads while keeping the binary index, explicit corpus embeddings, Owner document identifiers, and unselected content keys within the Owner process. The compact randomized-evaluation path makes the evaluated ciphertext simulatable from the query ciphertext, the K scores, and public metadata, so its other slots and coefficients add no corpus-embedding information beyond this explicit score oracle (§6.3).

The security model assumes an authenticated confidential transport, under which a network observer learns message lengths and timing. Our measurement harness uses versioned framed TCP to expose and measure these metadata costs; a deployment places the same frames and the OT connection inside authenticated encrypted channels.

Assumptions and scope. A trusted model-distribution step fixes the encoder, hash head, quantization parameters, and HE parameters shared by both parties. The HE scoring theorems apply to fresh symmetric BFV encryptions of bounded, canonically packed queries; the score quota governs cumulative exposure but does not establish consistency between the encrypted query and the coarse code. The guarantees assume uncompromised endpoints, authenticated identities, and authenticated confidential transport, with message lengths and timing treated as explicit leakage. They cover query privacy, exact and ciphertext-simulatable scoring for conforming inputs, and per-round payload access; Sybil resistance, availability, inference from the released exact scores, and malformed-ciphertext server privacy remain outside the model.

3.3 Empirical Privacy Attacks

We evaluate three attacks that recover text or sensitive attributes from an exposed representation. DP mechanism experiments target the released Stage 1 code, while the no-DP learned code and float embedding provide reference points.

Search-based inversion. ZSInvert [90] treats reconstruction as black-box optimization. Given a target representation r, an LLM proposes a beam of candidate texts, the target encoder maps each candidate to the same representation space, and similarity to r selects the next beam. The best candidate then seeds another refinement round. Float targets use cosine similarity, whereas binary targets use normalized hash similarity; the latter supplies only $L+1$ distinct Hamming-similarity values. The output is the highest-scoring reconstructed text.

Generation-based inversion. GEIA [49] learns an embedding-conditioned autoregressive decoder from auxiliary text–representation pairs. A learned projection maps the

Table 1: Private retrieval systems grouped by corpus ownership, querier role, and trusted infrastructure.

Scheme	Search	Infra.	Query privacy	Content privacy	Pattern privacy	Near-lossless	Fast online	Index / auxiliary state
Client-owned outsourcing—corpus owner is querier; cloud is adversary								
CGKO06 [20]	Keyword	1S	✓	–	✗	–	–	Inverted index
CLRZ18 [9], SOPK21 [72]	Keyword	1S	✓	–	DP	–	–	DP index
LZXL25 [54]	Graph	1S	△	–	✗	△	✓	Graph + 2 CT
Compass [96]	Graph	1S	✓	–	✓	✓	△	3.2–6.8× server
MESS [19]	Multi-graph	1S	DP	–	DP	△	△	16× HNSW index
Public/shared corpus—no corpus-confidentiality goal								
Tiptoe [32]	k-means	45S	✓	–	△	✗	✗	Cluster index
Wally [3]	k-means	Crowd	DP	–	DP	✗	✗	Cluster index
PACMANN [94]	Graph	1S+P	✓	–	△	△	✗	Graph + client hints
Secret-shared outsourcing—distinct owner and querier								
PRAG [97]	IVF	MPC	✓	✓	✓	△	✗	Database shares
P^{2}RAG [59]	Full scan	2NC	✓	△	✓	✓	✗	Two DB shares
Provider-held proprietary corpus—external querier								
Pisces [52]	SimHash + BM25	1S	✓	✓	△	△	✗	160N-CT + token OKVS
SANNS [10]	k-means	1S	✓	✓	✓	△	✗	DORAM
PANTHER [50]	k-means	1S	✓	✓	✓	△	✗	PIR + MPC
RemoteRAG [14]	ANN	1S	DP	△	DP	✓	△	ANN index
Ours	Hash scan	1S	DP	✓	DP	✓	✓	**32 B/doc filter**

Legend. ✓: cryptographic protection or full support; DP: formal differential privacy guarantee; △: empirical or partial protection/support; ✗: unsupported; –: inapplicable. Query privacy protects the query text and clean embedding from the search service. Content privacy limits the querier's plaintext payload recovery to its authorized results. Pattern privacy separately protects the service-visible retrieval trace: the search pattern reveals whether queries repeat, while the access pattern reveals which corpus items are touched or selected. Near-lossless denotes exact retrieval or at least 99% of the matched plaintext quality. Fast online denotes practical reported query-time latency at the evaluated scale; workloads and hardware differ. The final column reports each paper's native search structure beyond the embeddings. Our 32 B/doc figure is the 256-bit coarse hash index. 1S: one server; 2NC: two non-colluding servers; CT: ciphertext representation; OKVS: oblivious key–value store; P: per-client, database-dependent PIR preprocessing.

target representation into the decoder input space, and teacher-forced language-model training maximizes $\prod_t p_\phi(x_t \mid x_{<t}, r)$. At inference time, the trained decoder reconstructs each held-out target in one generation pass.

Property inference. The attacker obtains an auxiliary set of representation–attribute pairs [73], trains a classifier to predict the attribute from the exposed representation, and applies the selected classifier to held-out victim representations. We evaluate topic, sentiment, and authorship because they span coarse semantic content, affect, and fine-grained source identity. This attack can succeed without reconstructing the original text.

Section 7.4 reports attack outcomes, and the experimental setup specifies the models, datasets, splits, search budgets, and metrics. Appendix B.3 records the remaining optimization details. These experiments isolate representation leakage; §3.2 separately accounts for candidate identities and cross-round linkage in the Owner's protocol view.

4 Deep Hash Learning

The learned hash encoder turns a pretrained dense retriever into a high-recall candidate filter, while the original pretrained encoder remains the Stage 2 scorer. We focus here on the model architecture and the training signals that make this separation effective; Appendix B.1 gives the exact losses, discretization schedule, and optimization parameters.

4.1 Motivation and Architecture

A key challenge in deep hashing is preserving zero-shot candidate recall after adapting a pretrained encoder to a discrete space [31]. For text x, our hash model applies a linear head to a LoRA-adapted encoder [34] and emits

$$\mathbf{b}(x) = \operatorname{sign}(W \operatorname{pool}(E_{\text{LoRA}}(x))) \in \{-1, 1\}^L. \tag{2}$$

Here $W \in \mathbb{R}^{L \times d}$, where d is the encoder hidden dimension and L is the code length. We use mean pooling and $L = 256$ for E5-base-v2 [78], and [CLS]-token pooling and $L = 512$ for BGE-base-en-v1.5 [11]. The linear head and compact LoRA update specialize the pretrained representation for Hamming candidate recall without training another backbone from scratch.

Candidate generation and scoring use separate model states. The adapted encoder and hash head produce only the coarse code, while the unchanged pretrained encoder supplies the continuous query and document embeddings used by Stage 2. This separation lets Stage 1 reshape its geometry for Hamming search without shifting the final dense ranking. Online, both forwards reuse tokenization and input transfer, and the pretrained scoring forward overlaps Owner-side Hamming search as described in §5.2.

4.2 Encoder Tuning

We train the hash model on MS MARCO query–passage supervision using three functional groups:

$$\mathcal{L} = \mathcal{L}_{\text{retrieval}} + \mathcal{L}_{\text{transfer}} + \mathcal{L}_{\text{regularization}}. \tag{3}$$

Direct retrieval supervision brings relevant query–passage pairs together and pushes mined negatives away in the adapted continuous space; E5 additionally applies this supervision directly in Hamming space. Ranking transfer carries the adapted encoder's ordering into the deployed Hamming space. The remaining regularizers anchor the adapted representation to the frozen pretrained geometry and keep examples well spread before binarization. Appendix B.1 defines every component of Equation 3 and reports its model-specific weight.

4.3 Hard-Negative Training

A hard negative is a non-relevant passage that remains deceptively close to the query, making it more informative than a random passage for learning the candidate boundary. We mine these examples only from the MS MARCO training corpus: a broad lexical retrieval stage finds plausible candidates, a stronger reranker orders them, and training samples from the highest-ranked non-relevant passages. The reranker also suppresses likely unlabeled positives so that ambiguous passages do not become contradictory supervision. This source-only procedure teaches the hash model to preserve fine distinctions without adapting to any evaluation corpus.

E5 and BGE each train for 16 epochs. Training moves progressively from a smooth representation to the binary codes used at deployment; Appendix B.1 specifies this schedule together with the mining models, sample counts, learning rates, and remaining hyperparameters.

5 DP-Filtered Private Dense Retrieval

We now describe the complete two-party protocol. Throughout, K denotes the candidate budget (the number of documents that receive HE scoring) and k denotes the final result count returned to the User.

5.1 Offline Setup

Owner setup. The Owner runs two document encoders offline. The LoRA-adapted hash encoder and linear hash head produce $\mathbf{b}_i \in \{-1,1\}^L$ for candidate generation, while the unchanged pretrained encoder produces the normalized scoring embedding $\mathbf{z}_i \in \mathbb{R}^d$. A shared symmetric quantizer with zero point 0 and scale $a = \max_{i,j}|z_{i,j}|/127$ maps the pretrained embeddings to $\{-127,\dots,127\}^d$; the symmetric range excludes -128 and bounds every integer dot product by $127^2 d$. The Owner stores:

- A flat binary index $\{\mathbf{b}_i\}_{i=1}^N$ for Hamming-distance search;
- The quantized embeddings $\{\bar{\mathbf{z}}_i\}_{i=1}^N$ as Owner-local plaintext HE operands;
- A random 128-bit content key κ_i and an AEAD ciphertext P_i for each document. The plaintext contains its true-length field and is zero-padded to a 4096-byte boundary before one-shot ChaCha20–Poly1305 encryption under a 256-bit key derived from κ_i with HKDF–SHA-256, so ciphertext length reveals only the padded block count.

User and session setup. For encrypted candidate scoring, we instantiate single-instruction multiple-data (SIMD) batched BFV for the quantized integer dot products [1, 24], packing multiple score lanes into each ciphertext. The User obtains the LoRA-merged hash encoder, hash head, original pretrained encoder, tokenizer, and quantizer. It generates the BFV secret/public keys and required Galois keys, retains the secret key, and sends only public and evaluation material to the Owner. An authenticated session binds an identity, layout, K, k, and monotone round counter. The two parties establish the base-OT correlation state once per long-lived OT connection; each query advances the extension state to derive fresh rows rather than repeating base OT.

5.2 Online Protocol

Figure 2 summarizes the eight online stages described below; Appendix D.1 provides the complete message sequence.

For each query x, the following steps are executed:

1. **Hash forward and DP release.** The User tokenizes x once, transfers the retained token tensors to the GPU once, and first runs the LoRA-merged hash encoder. The hash head produces logits $\mathbf{z}_q \in \mathbb{R}^L$; this path does not compute or return an unused continuous scoring vector. Then, using the final training scale β, the User computes the bounded pre-binarization vector $\mathbf{h}_q^{\text{soft}} = \tanh(\beta \mathbf{z}_q)$, normalizes it to $\mathbf{u}_q = \mathbf{h}_q^{\text{soft}}/\|\mathbf{h}_q^{\text{soft}}\|_2$, samples $\mathbf{y}_q \sim \text{vMF}(\mathbf{u}_q, \kappa)$ with the pure calibration in Equation 7, and sends the round-bound coarse frame containing $\tilde{\mathbf{b}}_q = \text{sign}(\mathbf{y}_q)$.
2. **Hamming search and payload streaming.** The Owner validates the received frame $\tilde{\mathbf{b}}_q$ and starts Hamming search over $\{\mathbf{b}_i\}_{i=1}^N$. Once $C_K = (i_1,\dots,i_K)$ is fixed, it atomically reserves the query and per-document score exposure and immediately streams the round-bound AEAD ciphertexts $(P_{i_1},\dots,P_{i_K})$ in candidate order.
3. **Concurrent pretrained forward and BFV encryption.** While the Owner computes C_K, the User runs the unchanged pretrained encoder on the retained token tensors, producing $\mathbf{q} = \mathsf{Normalize}(E_{\text{pre}}(x))$. The User quantizes $\mathbf{q}$ with the shared scale, encrypts it as $\mathsf{Enc}(\bar{\mathbf{q}})$, and sends a separate scoring-query frame bound to the same session and round. Thus the serial prefix is the hash forward followed by $\max\{T_{\text{Hamming}}, T_{\text{pre}} + T_{\text{enc}}\}$, and payload transfer begins as soon as the Hamming branch produces C_K.
4. **Packed BFV scoring.** Once both C_K and $\mathsf{Enc}(\bar{\mathbf{q}})$ are ready, the Owner gathers the pretrained quantized candidate matrix $\bar{\mathbf{Z}}_K$ and computes $\mathsf{Enc}(\mathbf{s}) = \bar{\mathbf{Z}}_K \mathsf{Enc}(\bar{\mathbf{q}})$ using plaintext–ciphertext multiplication and rotation-based sum reduction while payload streaming continues. At polynomial degree 8192, the segmented *multi* layout places eight candidates in 1024-slot segments per result ciphertext at multiplicative depth 1, while the deployed *compact* layout collects up

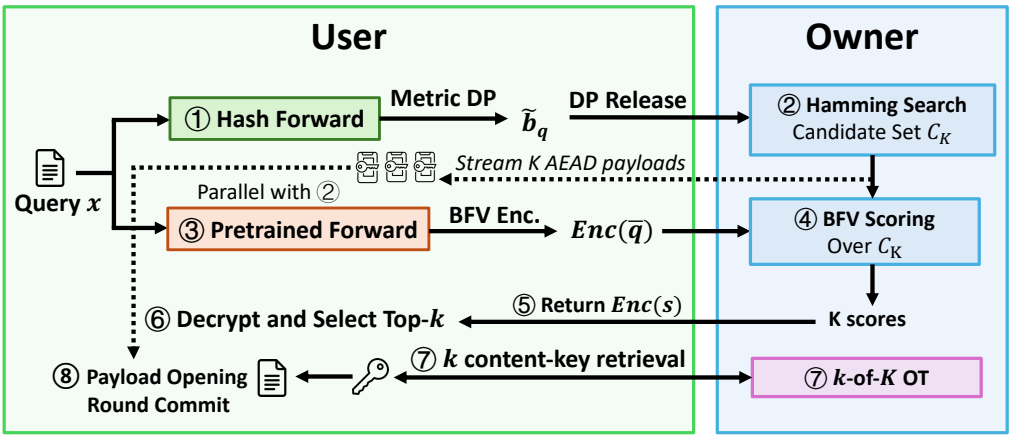

Figure 2: **Online retrieval pipeline.** Metric-DP Hamming shortlisting overlaps the pretrained scoring forward and query encryption. Once C_K is fixed, its AEAD ciphertexts stream during exact BFV scoring.

to 8192 scores in one ciphertext and applies PMultE/Rand randomized evaluation [36]. This path clears non-score slots and makes the evaluated ciphertext simulatable from the prescribed scores while preserving exact integer outputs (§6.3); it requires neither ciphertext–ciphertext multiplication nor bootstrapping.

5. **Encrypted-score return.** The Owner returns the encrypted scores in the same local order as C_K. Candidate position $j \in [K]$ is the only selection coordinate exposed to the User; Owner document identifiers are never transmitted.
6. **Decrypt and select top-k.** The User decrypts the exact integer score vector $\mathbf{s}$ and chooses local positions $c_1, \ldots, c_k \in [K]$.
7. **Active-secure k-out-of-K key retrieval.** The parties execute a batch of k active-secure 1-out-of-K Orrù–Orsini–Scholl (OOS) OT-extension transfers [64], one for each c_j. The receiver obtains one 128-bit OT key per row, while the sender derives K option keys per row and masks every candidate content key in a $k \times K$ table. The Owner sends this masked-key table after OT; row j and column c_j reveal $\kappa_{i_{c_j}}$ to the User, which opens the corresponding buffered payload. The OT structure limits a malicious receiver to at most one content key per row and hides each c_j from the Owner.
8. **Payload opening and round commit.** After authenticating and decrypting the selected AEAD payloads, the User sends a round-bound completion frame and the Owner commits the quota reservation. Any protocol or malicious-check failure poisons the affected daemon state and releases an uncommitted reservation.

5.3 Design Rationale

Metric DP. Hamming search requires the Owner to receive a coarse query code, which exposes the query's neighborhood and creates matching, property-inference, and linkability channels. Metric DP gives this released code a distance-calibrated indistinguishability guarantee while preserving efficient plaintext filtering.

Single-release utility recovery. MESS [19] recovers recall after discrete randomized response through 64 HNSW shards with separately trained IsoHash mappings, routing each item to 16 shards and searching every shard for candidates. This raises a relevant item's recovery probability to $1-(1-P_{\text{hit}})^{16}$; however, the hash codes appearing in 16 releases also compose privacy leakage across shards, leading to extremely large ε and almost null formal guarantee. Our learned filter, instead, releases one metric-DP code and recovers utility by enlarging K, leading to a good balance between privacy, efficiency, and search quality.

Table 2 gives an operational view of this coarse release. The closest DP-Hamming codes match isolated cues such as London, poppies, or tower without identifying the requested fact. The answer-bearing passage appears only at Hamming rank 1,439; encrypted clean-query scoring promotes it to rank 5, inside the User's hidden $k = 10$ selection. The large candidate set therefore preserves the answer while separating coarse code proximity from precise semantic relevance. Appendix B.2 presents three additional queries spanning factual counts, locations, and calendar rules.

Shallow HE. The scoring stage is a plaintext–ciphertext matrix–vector multiplication over pre-normalized vectors. Our exact-integer BFV path uses SIMD multiplication and rotation-based reduction at depth 1 in the multi layout or depth 2 in the communication-oriented compact layout. Local decryption handles top-k, eliminating encrypted comparison, ciphertext–ciphertext multiplication, and bootstrapping. The deployed compact path also clears non-output slots and randomizes evaluation according to ciphertext-simulatable BFV [36]: BFV hides the query from the Owner, while the randomized output reveals no Owner operand information to

Table 2: One NQ query (`test1054`) under the E5 pure-vMF operating point ($\varepsilon = 64$, $K = 3000$, $k = 10$, 256 bits). The DP-Hamming neighbors expose a broad mixture of query cues; Stage 2 recovers the answer-bearing passage.

View	Rank / d_H	Text excerpt
Target	–	*Who made the poppies at Tower of London?*
DP-Ham.	1 / 65	*10 Downing Street:* "The terrace and garden were constructed in 1736 . . . "
DP-Ham.	2 / 69	*Anzac Day:* "Paper poppies are widely distributed . . . "
DP-Ham.	3 / 70	*Eiffel Tower:* "26 December 1888: Construction of the upper stage."
Stage 2	1439→5 / 88	*Blood Swept Lands and Seas of Red:* "The artist was Paul Cummins, with setting by stage designer Tom Piper."

a conforming secret-key User beyond the K prescribed scores.

Key-only OT. Returning all K documents would disclose the full candidate payload set, whereas transferring full documents inside OT would make the active-secure OT payload proportional to document size. The protocol therefore sends fixed-size 128-bit content keys through k 1-out-of-K choices and delivers AEAD ciphertexts on the ordinary channel. This composition keeps OT small, hides the selected positions from the Owner, and caps payload decryption at k choices; authenticated quotas separately govern the deliberately released score vector.

Owner-local execution. The Owner already holds the corpus and can retain the binary index on its own infrastructure. The index occupies $NL/8$ bytes (e.g., 283 MB for N=8.84M at L=256), allowing the Hamming scan and HE scoring to run without a query-processing intermediary.

6 Security Contract and Protocol Guarantees

This section establishes the protocol's four guarantees. **Views** (§6.1) defines the information released to each party. **Query privacy** (§6.2) proves metric privacy of the candidate pattern and computational privacy of the Owner's complete view. **Scoring privacy** (§6.3) proves exact BFV scores and ciphertext simulatability for a conforming User. **Payload access** (§6.4) formalizes the score oracle and the k-payload bound.

6.1 Explicit Per-Party Views

For one round, the Owner's explicit leakage is

$$\mathcal{L}_O = (\mathsf{id}, \mathsf{session}, \mathsf{round}, K, k, \mathsf{layout}, \mathsf{lengths}, \mathsf{timing}, \tilde{b}_q, C_K, \mathsf{accept/reject}). \tag{4}$$

Adjacent executions fix all fields in Equation 4 except the DP output $(\tilde{b}_q, C_K, \mathsf{accept/reject})$. Theorem 2 accounts computationally for the BFV and OT transcripts in the real view.

The application-level disclosure to a conforming User is

$$\mathcal{L}_U^{\text{target}} = (K, \mathbf{s}, \{|P_i| : i \in C_K\}, \{D_i : i \in S,\ |S| \le k\}, \mathsf{linkage}), \tag{5}$$

Here $\mathbf{s}$ is the candidate-local score vector, P_i is a padded payload ciphertext, and linkage records equality of recurring ciphertexts. Theorems 4 and 5 realize Equation 5 for HE scoring and payload recovery, respectively.

6.2 Directional Candidate Privacy and Query Privacy

For query x, let $h(x) = \tanh(\beta z(x)) \in [-1,1]^L$ and $u(x) = h(x)/\|h(x)\|_2 \in \mathbb{S}^{L-1}$. Given u, the mechanism samples $Y \sim \mathrm{vMF}(u, \kappa)$, whose density is $p_u(y) = C_L(\kappa)\exp(\kappa u^\top y)$, and releases $M_\kappa(u) = \mathrm{sign}(Y)$.

Theorem 1 (Pure directional metric privacy)**.** *For every $\kappa \ge 0$, every $u, u' \in \mathbb{S}^{L-1}$, and every output event S,*

$$\Pr[M_\kappa(u) \in S] \le \exp\big(\kappa d_c(u,u')\big)\Pr[M_\kappa(u') \in S]. \tag{6}$$

Consequently, for angular radius $\rho \in (0,\pi]$, choosing

$$\kappa = \frac{\varepsilon}{2\sin(\rho/2)} \tag{7}$$

gives $(\varepsilon, 0)$-directional privacy within that radius.

Proof. The vMF normalizer is independent of its mean direction, so for every y, $\log(p_u(y)/p_{u'}(y)) = \kappa(u-u')^\top y \le \kappa\|u-u'\|_2$. Integration gives the bound for Y, and sign binarization preserves it by post-processing. Finally, $d_c(u,u') = 2\sin(d_\theta(u,u')/2) \le 2\sin(\rho/2)$ inside the angular radius. □

Equation 6 is the global metric-DP guarantee under chordal distance [8]; Equation 7 is its angular-radius corollary.

Theorem 2 (Computational directional privacy of the Owner view)**.** *Suppose a conforming User generates the coarse and encrypted queries, BFV is indistinguishable under chosen-plaintext attack (IND-CPA), OT is receiver-private against its sender, and adjacent executions have identical auxiliary fields in Equation 4. For every probabilistic polynomial-time (PPT) distinguisher A and $\eta = \kappa d_c(u,u')$ such that e^η is polynomially bounded in the security parameter,*

$$\Pr[A(\mathsf{View}_O(x)) = 1] \le e^\eta \Pr[A(\mathsf{View}_O(x')) = 1] + \mathrm{negl}(\lambda). \tag{8}$$

Proof sketch. Receiver privacy simulates the choice-dependent OT messages, and BFV IND-CPA replaces the clean-query ciphertext by an encryption of zero. The residual view is $(M_\kappa(u), C_K, \mathsf{accept/reject})$, where the last two components are post-processing. Theorem 1 supplies the factor e^η; polynomially bounded e^η absorbs the hybrid losses into $\mathrm{negl}(\lambda)$. □

For T adaptive releases, sequential composition replaces η in Equation 8 by $\eta_T = \kappa \sum_{t=1}^{T} d_c(u_t, u'_t)$, conditioned on the fixed auxiliary leakage.

6.3 Exact Scoring and HE Privacy Scope

Let the canonically packed query $\bar{\mathbf{q}}$ and every candidate $\bar{\mathbf{z}}_i$ lie in $[-B,B]^d$.

Lemma 3 (Exact integer scoring)**.** *If BFV decryption succeeds for the configured circuit and*

$$t > 2dB^2, \tag{9}$$

every decoded anchor equals $s_i = \sum_{j=1}^{d} \bar{z}_{ij}\bar{q}_j \in \mathbb{Z}$.

Proof. Since $|s_i| \le dB^2 < t/2$, centered reduction modulo t is injective on every possible score. Appendix A.3 proves that both layouts place this residue at each decoded anchor. □

The deployed $B = 127$, $d = 768$, and $t > 24{,}774{,}144$ satisfy Equation 9; BFV noise correctness remains an independent decryption condition.

BFV IND-CPA hides $\bar{\mathbf{q}}$ from the Owner. To hide the Owner's plaintext operands from the decrypting User, the compact path instantiates ciphertext-simulatable BFV [36]. Write $R_t = \mathbb{Z}_t[X]/(X^n+1)$, identify ring elements with coefficient vectors, and let $D_{\Lambda,\sigma}$ denote the discrete Gaussian on lattice coset Λ. For $\mu \in R_t$, it samples $\widehat{\mu} \leftarrow D_{\mu+t\mathbb{Z}^n,\sigma}$ and evaluates

$$\begin{aligned} \mathsf{PMultE}(ct,\mu) &= ct \cdot \widehat{\mu} + (e,0), \\ \mathsf{Rand}(pk) &= e_2 pk + (e_0,e_1), \end{aligned} \tag{10}$$

where $e \leftarrow \lfloor D_{\mathbb{R}^n,\tau} \rceil$, $e_1, e_2 \leftarrow D_{\mathbb{Z}^n,\sigma_r}$, and $e_0 \leftarrow \lfloor D_{\mathbb{R}^n,\tau_r} \rceil$. One $\mathsf{Rand}(pk)$ precedes the public linear rotation–mask–addition subcircuit; the compact mask leaves $\mathbf{s}$ in its canonical slots and zero elsewhere.

Theorem 4 (HE server-input privacy for conforming Users)**.** *Let ct_q be a fresh symmetric BFV encryption of a bounded, canonically packed query, and let the randomized-evaluation parameters satisfy the correctness and smoothing conditions of ciphertext-simulatable BFV [36]. Under the corresponding ring-learning-with-errors (RLWE) assumption, the compact scoring view of a conforming secret-key User is computationally simulatable as*

$$\mathsf{View}_U^{\mathrm{HE}} \approx_c \mathsf{Sim}(pk, sk, ct_q, \mathbf{s}, \mathsf{metadata}), \tag{11}$$

where $\mathbf{s}$ contains the prescribed K exact scores. Consequently, the evaluated ciphertext coefficients and all non-score slots reveal no information about the candidate embeddings beyond $\mathbf{s}$ and public metadata.

Proof sketch. PMultE error simulatability and Rand masking replace each group ciphertext by one generated from ct_q and its group scores [36]. Applying the public linear subcircuit preserves indistinguishability, and a hybrid over groups yields Equation 11 because the final plaintext is $\mathsf{Encode}(\mathbf{s}, 0, \ldots, 0)$. □

Appendix D.2 specifies the samplers and BFV parameters. The theorem assumes a fresh symmetric, canonical query ciphertext; malformed ciphertexts require a well-formedness proof. The scores $\mathbf{s}$ remain explicit leakage.

6.4 Exact-Score Exposure and Payload Access

The score oracle and quota ledger satisfy

$$\begin{aligned} \log_2 |\operatorname{supp}(\mathbf{s})| &\le K \log_2(2dB^2+1), \\ r_i \le R &\Longrightarrow \operatorname{rank}(Q_i) \le \min\{R, d\}, \end{aligned} \tag{12}$$

where $\mathbf{s}_i = Q_i \bar{\mathbf{z}}_i$ contains the r_i scores released for document i. The first bound grows linearly with K; the second counts linear observations but does not bound inference from representation priors. Authentication is required because Sybil identities reset R.

Theorem 5 (Per-round payload-key bound)**.** *Assume the OOS extension realizes active-secure 1-out-of-K OT for each of k rows [64], the OT-key mask is pseudorandom, content keys are independent, and the payload cipher is authenticated encryption. Except with negligible probability, a malicious receiver completing one accepted round recovers at most k distinct candidate content keys and therefore at most k distinct candidate payloads, independently of K.*

Proof sketch. Active receiver security reveals at most one option key per row. Pseudorandom masking hides every unchosen content key, and authenticated-encryption confidentiality hides its payload; summing over k rows proves the bound. □

Across T accepted rounds, the payload bound composes to Tk; corpus enumeration remains possible if the accepted selections eventually cover it. Scores, padded lengths, stable-ciphertext linkage, and cross-round inference remain explicit leakage. Thus increasing K enlarges Equation 12 but not the per-round payload cap.

7 Evaluation

Our evaluation asks four questions: whether the learned filter outperforms classical alternatives while preserving retrieval across models, domains, and corpus scales; whether the resulting evidence supports end-to-end RAG quality; how shortlist size and two-forward overlap determine protocol cost; and how privacy calibration and candidate budget jointly determine retrieval quality and representation exposure.

Table 3: Two-forward retrieval on five BEIR corpora spanning 25K–5.4M documents. Stage 1 uses the learned hash model; Stage 2 reranks its candidates with the original pretrained encoder. Pretr. Fl. is matched full-corpus retrieval, and each $\Delta@K$ is Stage 2@K minus Pretr. Fl.

			Stage 1: Bin. Recall@K		**Stage 2: NDCG@10**				
Model	**Dataset**	**Docs**	**K=200**	**K=500**	**S2@200**	**S2@500**	**Pretr. Fl.**	$\Delta@200$	$\Delta@500$
E5-base-v2	SciDocs	25K	.4331	.5469	.1875	.1874	.1870	+.0005	+.0004
(256-bit)	NQ	2.7M	.8884	.9271	.5723	.5787	.5854	−.0132	−.0068
	DBpedia-Entity	4.6M	.4549	.5474	.4144	.4224	.4271	−.0127	−.0047
	Climate-FEVER	5.4M	.5300	.6154	.2818	.2785	.2627	+.0192	+.0158
	FEVER	5.4M	.9368	.9484	.8417	.8451	.8501	−.0084	−.0050
BGE-base	SciDocs	25K	.5211	.6467	.2224	.2225	.2228	−.0004	−.0003
(512-bit)	NQ	2.7M	.8923	.9349	.5326	.5372	.5414	−.0088	−.0042
	DBpedia-Entity	4.6M	.4764	.5679	.4018	.4041	.4081	−.0063	−.0040
	Climate-FEVER	5.4M	.5935	.6774	.2874	.2848	.2836	+.0038	+.0012
	FEVER	5.4M	.9429	.9517	.8480	.8483	.8495	−.0015	−.0012

7.1 Experimental Setup

Encoder, Datasets, and Metrics. We train E5-base-v2 [78] and BGE-base-en-v1.5 [11] hash models on MS MARCO passage ranking [63] and evaluate zero-shot transfer on five BEIR corpora [75]: SciDocs [18], Natural Questions (NQ) [44], DBpedia-Entity [29], Climate-FEVER [21], and FEVER [76], following the standard BEIR zero-shot protocol. The corpora span from 25,657 to 5.4M documents. E5-base-v2 uses 256-bit codes, and BGE-base-en-v1.5 uses 512-bit codes. Stage 1 reports Recall@K over the relevance judgments for the learned Hamming filter; Stage 2 reranks exactly those candidates with the unchanged pretrained encoder and reports NDCG@10. Full-corpus retrieval with the same pretrained encoder is presented as a reference. Appendix B.1 gives the detailed training recipe and hyperparameters.

Representation Exposure. Search-based inversion evaluates the embedding-guided search stages of ZSInvert [90] on 100 randomly sampled MS MARCO documents. Llama-3.1-8B-Instruct [55] generates a width-50 beam for six search rounds, scored by float cosine or normalized hash similarity; we report mean verifier cosine and attack success at cosine 0.8. Generation-based inversion trains a GEIA [49] DialoGPT-medium [93] decoder on PersonaChat [91] for 10 epochs and reports token F1 and verifier cosine on held-out passages. Following the property-inference threat model of Song and Raghunathan [73], we evaluate AG News topic [92], IMDB sentiment [58], and 50-way 20 Newsgroups authorship [46] labels. Five stratified 60/20/20 splits separate attacker training, model selection, and victim testing; validation macro F1 selects the best model among logistic regression, a two-layer MLP, and LightGBM [42], and we report victim-test macro F1 averaged across the five splits. Appendix B.3 gives the complete attack flow, decoding limits, optimization hyperparameters, and preprocessing.

Following metric- and directional-DP evaluation conventions [8, 82], we report each operating point by its protected space, radius, and (ε, δ) parameters. Definition 1 defines the generic radius ρ; here, we denote ρ_h to be the Euclidean radius on the bounded pre-sign vector, and ρ_θ to be the angular radius on its normalized direction. The RDP-vMF comparison maps $\rho_h = 2$ to $\rho_\theta = 2\arcsin(1/\sqrt{L})$. The retrieval sweeps and property-inference table use $\varepsilon \in \{8, 16, 32, 64\}$ at $\rho_h = 2$; the inversion sweeps additionally include tighter budgets and the $\rho_h = 6.32$ setting. Following standard approximate-DP calibration [23], Gaussian and RDP-vMF set $\delta = 10^{-6}$, below the inverse of every evaluated query-set size, while pure-vMF provides $\delta = 0$. The randomized attack plots use the same calibrated mechanisms as the retrieval comparison, and Theorem 1 gives the angular calibration for the protocol's pure-vMF release.

7.2 Retrieval Quality

7.2.1 Main Retrieval Results

A concise shortlist preserves quality and may prune distractors. Table 3 evaluates the deployed two-forward path at $K \in \{200, 500\}$. Recall@K measures how much judged-relevant material survives the learned filter; Stage 2 NDCG@10 measures the ranking obtained when the original pretrained model scores only those candidates. The full-corpus column uses the same scorer, and $\Delta@200$ and $\Delta@500$ isolate the candidate filtering capability of the hash model.

At $K = 500$, both encoders retain 98.84–100.21% of full-corpus NDCG@10 on SciDocs, NQ, DBpedia-Entity, and FEVER. On Climate-FEVER, the shortlist even improves NDCG@10 by 0.0158 for E5 and 0.0012 for BGE. We hypothesize that the learned filter can act as a coarse semantic denoiser that removes spurious high-scoring distractors that Stage 2 would otherwise place ahead of relevant ones.

We observe only a slight increase in NDCG@10 (0.0064

Table 4: E5 candidate-filter baselines averaged over the five corpora in Table 3. Every method uses exact Hamming Stage 1 and the same pretrained E5 Stage 2 scorer. Direct sign uses 768 embedding coordinates; remaining methods use 256 bits.

	Recall@K		NDCG@10	
Method	$K = 200$	$K = 500$	$K = 200$	$K = 500$
Direct sign (768b)	.5046	.5669	.4183	.4338
Random-hyperplane LSH [7]	.3068	.3743	.2893	.3283
Super-Bit LSH [39]	.3257	.3916	.3008	.3366
PCA-sign	.5802	.6398	.4469	.4534
ITQ [28]	.6080	.6789	.4494	.4562
IsoHash [43]	.6098	.6786	.4520	.4573
BPR† [86]	.6125	.6813	.4473	.4540
Learned filter (ours)	**.6486**	**.7170**	**.4595**	**.4624**

†For a fair comparison, we manually reimplement BPR using the same E5 representation, 256-bit budget, MS MARCO training data, symmetric Hamming candidate search, and evaluation pipeline as our method.

and 0.0080, respectively) as K increases from 200 to 500, while several easier pairs are already saturated at $K = 200$. We therefore report both budgets here and benchmark latency through $K = 3000$, leaving larger candidate pools available for the differential-privacy operating points evaluated next.

The learned filter outperforms classical and supervised hashing baselines. Table 4 organizes candidate filters by the information used to construct their codes. Direct sign is a parameter-free, one-bit quantization of each pretrained coordinate. Data-independent LSH comprises random-hyperplane LSH [7] and Super-Bit LSH [39], which orthogonalizes the random projections. Unsupervised data-dependent hashing includes PCA-sign and the learned rotations of ITQ [28] and IsoHash [43]; we fit each transformation on MS MARCO passage embeddings and transfer it across corpora. BPR [86] represents a recent supervised learning-to-hash method through a pairwise ranking objective, while our filter jointly adapts the encoder and hash head with retrieval and ranking supervision.

Table 4 isolates candidate-filter quality by holding exact Hamming search and the pretrained Stage 2 scorer fixed. Our filter improves mean Recall@500 by 0.0357 and mean Stage 2 NDCG@10 by 0.0084 over BPR; it also improves mean Stage 2 NDCG@10 by 0.0051 over the strongest unsupervised baseline. The full-precision scorer can repair ordering only among documents retained by Stage 1, making candidate recall the more direct measure of hash quality.

7.2.2 Retrieval under Differential Privacy

Following the comparison methodology of Biswas et al. [5], we evaluate randomized response, analytic Gaussian, and RDP-vMF at common $(\varepsilon, \delta = 10^{-6})$ targets and the protocol's formal pure-vMF mechanism under pure metric DP. Randomized response composes privacy across the full binary code; RDP-vMF calibrates its Rényi-divergence curve at the angular boundary defined above; Gaussian calibrates its analytic profile to Euclidean radius $\rho_h = 2$ on the bounded pre-sign representation. Pure-vMF uses the exact directional calibration in Theorem 1.

Continuous randomization preserves retrieval quality. Table 5 reports randomized response at $K = 3000$ and selects the continuous mechanisms' smallest evaluated K that reaches approximately 99% retention at $\varepsilon = 16$ or 64, while retaining the $K = 3000$ endpoints at tighter budgets. RDP-vMF reaches 99.2% at $K = 1000$, Gaussian reaches 99.2% at $K = 2000$, and formal pure-vMF reaches 99.4% at $K = 2000$. With 128-token Qwen3-32B generation [87], their absolute protection costs are 0.37, 0.73, and 0.73 seconds (**9.97%** of the plaintext pipeline); the $K = 3000$ endpoint adds at most 1.10 seconds. The per-dataset K curves and complete E5 and BGE sweeps appear in Appendix C.

7.2.3 End-to-End RAG Quality

Protected retrieval preserves end-to-end RAG quality. We evaluate 500 Natural Questions queries over the full 2.68M-passage BEIR corpus [75]. The E5 scorer supplies the top five passages to Qwen3-32B, which returns a short answer under greedy decoding; exact match and token F1 use the NQ-Open answer aliases. Table 6 shows that every two-forward operating point remains within 0.2 EM and 0.20 F1 of full-corpus float retrieval. Pure-vMF at $(\varepsilon = 64, K = 3000)$ reaches 50.0 EM and 62.89 F1, compared with 50.0 and 63.09 for the float reference. Appendix C.1 evaluates client-side cross-encoder reranking of the authorized payloads.

7.3 Efficiency Evaluation

We evaluate four sources of systems cost: the two-forward pipeline, candidate-set scaling, complete RAG latency, and the online latency relative to prior private-retrieval systems. Appendices D.2 and D.3 give the implementation and measurement details.

Pipeline overlap absorbs almost all two-forward overhead. We measure the complete online path on SciDocs ($N = 25{,}657$, $d = 768$, $k = 10$, $K = 500$) over a 10-Gbps link, spanning shared tokenization, dual BF16 encoder forward passes, randomized compact BFV scoring, active-secure OT, and final payload decryption. The two encoders reuse the same token tensors, and the first forward computes only the hash logits. Model-stage and cryptographic results are means over 50 queries. Figure 3 expands Steps 1–4 of the online protocol (§5.2). After Step 1 sends the coarse frame, Owner-side Hamming search (Step 2) and the User's pretrained forward plus BFV encryption (Step 3) run concurrently. For E5, Step 2 finishes at 10.89 ms and starts payload streaming before Step 3 finishes at 13.26 ms, so Step 4 waits for Step 3. For BGE, Step 3 finishes at 13.34 ms before Step 2 fixes C_K at 14.23 ms, so Step 4 instead waits for Step 2. The complete paths take 198.91 and 199.88 ms, respectively.

Table 5: Representative two-forward DP operating points. NDCG@10 averages E5 and BGE on SciDocs, NQ, and FEVER. RAG latency uses Qwen3-32B with 128 output tokens on NQ; Δ is the absolute protection cost over the plaintext two-forward pipeline.

Release	Guarantee	(ε, K)	NDCG@10	Ret.	RAG s	Δs
Full-corpus float	Reference	(∞, N)	.5394	100.0%	7.302	—
Plaintext filter	None	$(\infty, 500)$	.5363	99.4%	7.305	0
Randomized response	(ε, δ)-DP	$(16, 3000)$	.0306	5.7%	8.403	+1.098
Gaussian	(ε, δ)-DP	$(8, 3000)$	.5084	94.3%	8.403	+1.098
Gaussian	(ε, δ)-DP	$(16, 2000)$	.5350	99.2%	8.033	+0.728
RDP-vMF	(ε, δ)-DP	$(8, 3000)$	.5187	96.2%	8.403	+1.098
RDP-vMF	(ε, δ)-DP	$(16, 1000)$	.5350	99.2%	7.676	+0.371
Pure-vMF	Pure metric DP	$(32, 3000)$	.5229	96.9%	8.403	+1.098
Pure-vMF	Pure metric DP	$(64, 2000)$	.5359	99.4%	8.033	+0.728

Table 6: End-to-end RAG quality on 500 NQ queries. Hit is answer-alias coverage in 5 passages supplied to Qwen3-32B.

Retrieval	(ε, K)	Hit	EM	F1
Full-corpus float	(∞, N)	88.4	50.0	63.09
No-DP hash	$(\infty, 500)$	87.8	50.2	63.12
Gaussian	$(16, 3000)$	88.2	50.2	63.06
RDP-vMF	$(16, 3000)$	87.8	50.2	63.00
Pure-vMF	$(64, 3000)$	87.8	50.0	62.89

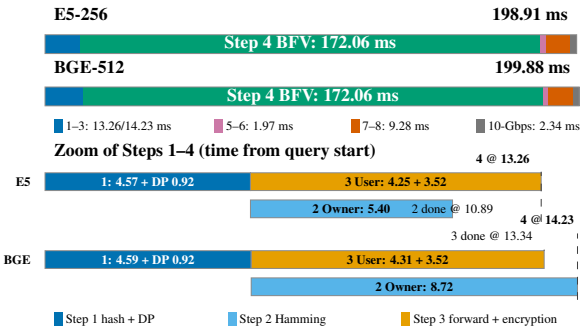

Figure 3: Protected critical-path latency at $K = 500$. After Step 1, Owner-side Step 2 and User-side Step 3 overlap, hiding 5.40 of 7.77 ms for E5 and all 7.83 ms for BGE before Step 4. Times are milliseconds; §5.2 defines the steps.

Candidate budget directly controls compute and communication. We sweep K from 200 to 3000 on SciDocs, measuring the complete randomized-evaluation protocol and projecting its exact serialized traffic at 100 Mbps, 1 Gbps, and 10 Gbps. As Table 7 shows, compute grows from 104.1 ms at $K = 200$ to 1096.2 ms at $K = 3000$, while traffic grows from 1.64 to 13.64 MB. At $K = 500$, total latency is 198.9 ms at 10 Gbps and 430.9 ms at 100 Mbps; at $K = 3000$, these values rise to 1.107 and 2.187 seconds. Compact BFV scoring returns one score ciphertext throughout this range, so the remaining growth comes from candidate scoring and the payload and masked-key traffic, which scale with K. The smallest candidate budget that meets the retrieval-quality target therefore gives the best operating point.

Table 7: Protected retrieval before generation versus candidate budget on SciDocs. Compute includes query DP and the complete randomized-evaluation protocol except transfer; bandwidth columns add the exact serialized traffic.

			Protected latency (ms)		
K	Compute	Traffic (MB)	100 Mbps	1 Gbps	10 Gbps
200	104.1	1.64	235.6	117.3	105.4
500	196.6	2.93	430.9	220.0	198.9
1000	376.1	5.07	781.7	416.6	380.1
2000	729.8	9.35	1478.2	804.7	737.3
3000	1096.2	13.64	2187.3	1205.3	1107.1

Protection adds little latency to a full RAG pipeline. We place the full 2.68M-passage E5 index on one A100 and Qwen3-32B on a second A100, which generates 128 tokens in 7.296 seconds. The plaintext filter and full-corpus float baselines take 7.305 and 7.302 seconds end to end. Figure 4 removes this shared generation time and reports the incremental protection cost: $K = 500$ adds 0.190 seconds (2.6%), the representative $K = 1000$ and $K = 2000$ operating points add 0.371 (5.1%) and 0.728 seconds (10.0%), and $K = 3000$ adds 1.098 seconds (15.0%). The compact view makes the candidate-budget scaling visible without redrawing the same generation bar for every operating point; Appendix D.3 reports the measurement scope and output-length sensitivity.

At matched quality, our protocol is fastest at every evaluated corpus scale. Figure 5 compares online latency against the recent P^{2}RAG [59], RemoteRAG [14], and PANTHER [50] at 100 Mbps using the same 768-dimensional E5-base-v2 embeddings, top-10 output, and hardware setup. Our protocol uses pure-vMF at $\varepsilon = 64$ and the smallest dataset-specific K retaining at least 99% of float NDCG@10: 292 for SciDocs, 104 for Touché, and 1956 for NQ-1M. It takes 0.298, 0.160, and 1.456 seconds on the three corpora, respectively; Touché is faster than SciDocs because it has a smaller

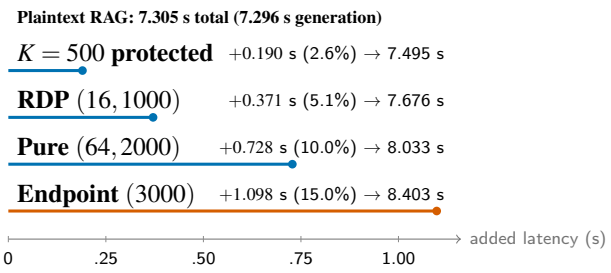

Figure 4: Incremental latency over plaintext 128-token RAG. Qwen3-32B generation takes 7.296 seconds on one A100 (7.305 seconds total); labels report added latency, percentage overhead, and protected total.

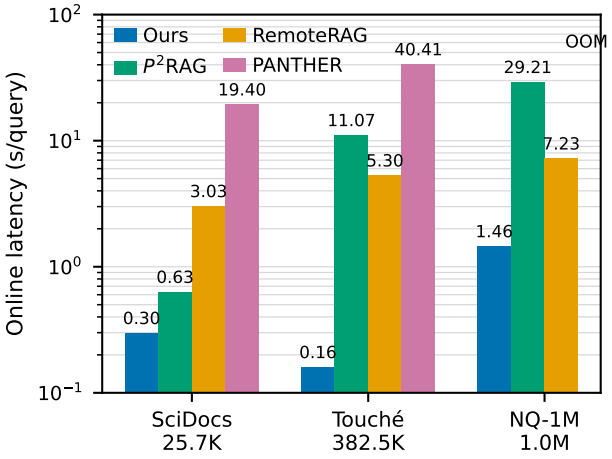

Figure 5: Matched online protocol latency on three corpus scales using E5-base-v2, top-10 output, and a 100-Mbps link. Our protocol uses pure-vMF at $\varepsilon = 64$ and the smallest dataset-specific K that retains at least 99% of float NDCG@10. PANTHER exhausts 256 GB of memory in NQ-1M.

candidate set. These latencies are 2.1×, 69.2×, and 20.1× faster than P^{2}RAG and 10.2×, 33.1×, and 5.0× faster than RemoteRAG. PANTHER takes 19.40 and 40.41 seconds on SciDocs and Touché and exhausts 256 GB of memory during its NQ-1M PIR answer. The P^{2}RAG implementation uses 192 hardware threads and excludes trusted-dealer preprocessing, whereas our BFV scorer uses only 16 threads. Therefore, these choices favor its reported online latency. Appendix D.4 provides the matched benchmark contract, baseline implementations, and corresponding 1-Gbps results.

7.4 Representation Exposure

Binarization and DP provide defense in depth against inversion. Table 8 separates the two layers: the learned hash creates a discrete bottleneck that reduces leakage even without DP, and metric-DP randomization adds a formally calibrated second layer. Table 9 illustrates the aggregate trend on a common target. The float reconstruction recovers the named organization, its nonprofit status, and its activity; the learned hash turns the foundation-like acronym into a banking organization; and the randomized code produces an unrelated geographic passage. Appendix B.4 reports the complete decoded strings for this target and three additional cases.

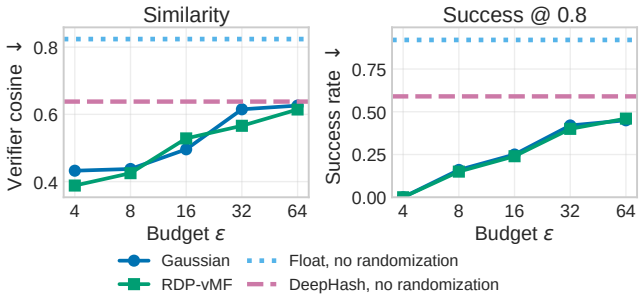

Figure 6: Search-based embedding inversion under Gaussian and RDP-vMF randomization at $\rho_h = 2$. The left panel reports mean verifier cosine similarity and the right panel reports attack success at a threshold of 0.8. Horizontal lines show the float and no-DP learned-code references.

Table 8: Embedding-inversion measurements for E5-base-v2; lower is better. Hash and pure-vMF releases use 256-bit codes, and pure-vMF provides metric DP. Search success uses a verifier-cosine threshold of 0.8.

Metric	Float	No-DP hash	Pure-vMF ($\varepsilon = 64$)
Search cosine	.824	.638	.423
Search success	.920	.590	.325
Gen. token F1	.545	.368	.205
Gen. cosine	.784	.584	.328

Table 9: One search-based inversion case shared across the evaluated releases. Excerpts are shortened; Appendix B.4 gives the complete outputs and three additional cases. Verifier cosine measures semantic agreement with the target.

Release	Target or reconstruction excerpt	Cosine
Target	"Welcome to the U.S. High School Bowling Foundation ... promotes the growth of high school bowling."	–
Float	"US Bowling Foundation. The nonprofit organization ... 501(C) ..."	.903
Learned hash	"US Bank is affiliated by National Bank Union ..."	.464
Gaussian, $\varepsilon = 8$	"The country Australia encompasses expansive territories ..."	−.060

Tighter randomization suppresses embedding inversion. Figure 6 shows that both mechanisms suppress successful

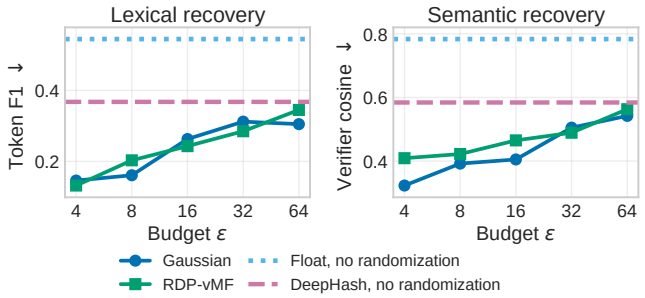

Figure 7: Generation-based embedding inversion under Gaussian and RDP-vMF randomization.

search-based reconstruction at tighter privacy budgets, degrading smoothly when ε relaxes. Generation-based inversion exhibits the same privacy-budget response (Figure 7): randomized releases reveal less lexical and semantic information than the learned code, with stronger suppression at smaller ε.

Table 10: E5 property-inference macro F1 for the validation-selected best attacker; lower is better. Entries average five stratified splits.

Release	ε	Topic	Sentiment	Authorship	Mean
Gaussian	8	.6001	.4763	.0862	.3875
Gaussian	16	.6303	.5402	.1002	.4235
Gaussian	32	.7024	.5849	.1303	.4725
Gaussian	64	.8221	.6575	.1875	.5557
RDP-vMF	8	.5980	.5021	.0727	.3909
RDP-vMF	16	.6251	.5528	.0924	.4234
RDP-vMF	32	.6713	.6057	.1235	.4668
RDP-vMF	64	.8325	.6515	.1550	.5463
Learned hash	∞	.8774	.7383	.2240	.6133
Float reference	∞	.8872	.7986	.3876	.6911

Randomization further weakens property inference. Property inference confirms the same layered effect (Table 10): learned hashing removes attribute signal relative to the float representation, and randomization further weakens topic, sentiment, and authorship inference as ε tightens. Gaussian and RDP-vMF yield similar attack leakage and utility trade-offs. We leave mechanisms that provide better utility–privacy trade-offs under the same ε guarantee to future work.

8 Concluding Remarks

We presented a practical two-party private dense-retrieval design for provider-held corpora serving external users. Concentrating private computation on a high-recall shortlist preserves retrieval quality and practical latency at million-document scale, while layered protections limit query and selection leakage and enforce the per-query payload allowance. Evaluations against embedding-inversion and property-inference attacks show that the released representation reduces reconstruction fidelity and attribute leakage. The resulting deployment reconciles interests that usually conflict: legitimate users receive privacy-preserving, precise semantic search, while corpus owners retain control over valuable content and align disclosure with the service's billing model.

References

[1] Martin Albrecht, Melissa Chase, Hao Chen, Jintai Ding, Shafi Goldwasser, Sergey Gorbunov, Shai Halevi, Jeffrey Hoffstein, Kim Laine, Kristin Lauter, Satya Lokam, Daniele Micciancio, Dustin Moody, Travis Morrison, Amit Sahai, and Vinod Vaikuntanathan. Homomorphic encryption security standard. Technical report, HomomorphicEncryption.org, Toronto, Canada, November 2018.

[2] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya P. Razenshteyn, and Ludwig Schmidt. Practical and optimal LSH for angular distance. In *Advances in Neural Information Processing Systems 28*, pages 1225–1233, 2015.

[3] Hilal Asi, Fabian Boemer, Nicholas Genise, Muhammad Haris Mughees, Tabitha Ogilvie, Rehan Rishi, Guy N. Rothblum, Kunal Talwar, Karl Tarbe, Ruiyu Zhu, and Marco Zuliani. Scalable private search with wally. *CoRR*, abs/2406.06761, 2024.

[4] Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 394–403. PMLR, 2018.

[5] Sayan Biswas, Mark Dras, Pedro Faustini, Natasha Fernandes, Annabelle McIver, Catuscia Palamidessi, and Parastoo Sadeghi. Comparing privacy notions for protection against reconstruction attacks in machine learning. *ArXiv preprint*, abs/2502.04045, 2025.

[6] Jo Van Bulck, Marina Minkin, Ofir Weisse, Daniel Genkin, Baris Kasikci, Frank Piessens, Mark Silberstein, Thomas F. Wenisch, Yuval Yarom, and Raoul Strackx. Foreshadow: Extracting the keys to the intel SGX kingdom with transient out-of-order execution. In *27th USENIX Security Symposium*, pages 991–1008. USENIX Association, 2018.

[7] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 380–388. ACM, 2002.

[8] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *international symposium on privacy enhancing technologies symposium*, pages 82–102. Springer, 2013.

[9] Guoxing Chen, Ten-Hwang Lai, Michael K. Reiter, and Yinqian Zhang. Differentially private access patterns for searchable symmetric encryption. In *IEEE Conference on Computer Communications, INFOCOM 2018*, pages 810–818. IEEE, 2018.

[10] Hao Chen, Ilaria Chillotti, Yihe Dong, Oxana Poburinnaya, Ilya Razenshteyn, and M. Sadegh Riazi. SANNS: Scaling up secure approximate *k*-nearest neighbors search. In *29th USENIX Security Symposium*, pages 2111–2128. USENIX Association, 2020.

[11] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[12] Yiyi Chen, Heather C. Lent, and Johannes Bjerva. Text embedding inversion security for multilingual language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7808–7827. Association for Computational Linguistics, 2024.

[13] Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Ying Shan, Xiaojuan Qi, and Weiming Hu. How to make cross encoder a good teacher for efficient image-text retrieval? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26984–26993. IEEE, 2024.

[14] Yihang Cheng, Lan Zhang, Junyang Wang, Mu Yuan, and Yunhao Yao. Remoterag: A privacy-preserving LLM cloud RAG service. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 3820–3837. Association for Computational Linguistics, 2025.

[15] Jung Hee Cheon, Andrey Kim, Miran Kim, and Yong Soo Song. Homomorphic encryption for arithmetic of approximate numbers. In Tsuyoshi Takagi and Thomas Peyrin, editors, *Advances in Cryptology—ASIACRYPT 2017*, volume 10624 of *Lecture Notes in Computer Science*, pages 409–437. Springer, 2017.

[16] Justin Chiu and Keiji Shinzato. Cross-encoder data annotation for bi-encoder based product matching. In Yunyao Li and Angeliki Lazaridou, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*, pages 161–168. Association for Computational Linguistics, 2022.

[17] Benny Chor, Eyal Kushilevitz, Oded Goldreich, and Madhu Sudan. Private information retrieval. *Journal of the ACM (JACM)*, 45(6):965–981, 1998.

[18] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: document-level representation learning using citation-informed transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics, 2020.

[19] Haoyu Cui, Zengpeng Li, Tien Tuan Anh Dinh, and Mei Wang. MESS: Fast and private semantic search on multi-graph HNSW. *CoRR*, abs/2607.28999, 2026.

[20] Reza Curtmola, Juan A. Garay, Seny Kamara, and Rafail Ostrovsky. Searchable symmetric encryption: Improved definitions and efficient constructions. In *Proceedings of the 13th ACM Conference on Computer and Communications Security*, pages 79–88. ACM, 2006.

[21] Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614, 2020.

[22] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[23] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[24] Junfeng Fan and Frederik Vercauteren. Somewhat practical fully homomorphic encryption. *IACR Cryptology ePrint Archive*, 2012:144, 2012.

[25] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM, 2024.

[26] Natasha Fernandes, Yusuke Kawamoto, and Takao Murakami. Locality sensitive hashing with extended differential privacy. In Elisa Bertino, Haya Schulmann, and Michael Waidner, editors, *Computer Security - ESORICS 2021 - 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4-8, 2021, Proceedings, Part II*, Lecture Notes in Computer Science, pages 563–583. Springer, 2021.

[27] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2023.

[28] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 817–824. IEEE, 2011.

[29] Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. DBpedia-Entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268. ACM, 2017.

[30] Kun He, Fatih Çakir, Sarah Adel Bargal, and Stan Sclaroff. Hashing as tie-aware learning to rank. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4023–4032. IEEE Computer Society, 2018.

[31] Liyang He, Zhenya Huang, Cheng Yang, Rui Li, Zheng Zhang, Kai Zhang, Zhi Li, Qi Liu, and Enhong Chen. A survey on deep text hashing: Efficient semantic text retrieval with binary representation. *ArXiv preprint*, abs/2510.27232, 2025.

[32] Alexandra Henzinger, Emma Dauterman, Henry Corrigan-Gibbs, and Nickolai Zeldovich. Private web search with tiptoe. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023*, pages 396–416. ACM, 2023.

[33] Jiun Tian Hoe, Kam Woh Ng, Tianyu Zhang, Chee Seng Chan, Yi-Zhe Song, and Tao Xiang. One loss for all: Deep hashing with a single cosine similarity based learning objective. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24286–24298, 2021.

[34] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[35] Yu-Hsiang Huang, Yuche Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. Transferable embedding inversion attack: Uncovering privacy risks in text embeddings without model queries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4193–4205, 2024.

[36] Intak Hwang, Seonhong Min, and Yongsoo Song. Ciphertext-simulatable HE from BFV with randomized evaluation. Cryptology ePrint Archive, Paper 2025/203, 2025.

[37] Jacob Imola, Amrita Roy Chowdhury, and Kamalika Chaudhuri. Metric differential privacy at the user-level via the earth-mover's distance. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 348–362, 2024.

[38] Yuval Ishai, Joe Kilian, Kobbi Nissim, and Erez Petrank. Extending oblivious transfers efficiently. In *Advances in Cryptology—CRYPTO 2003*, volume 2729 of *Lecture Notes in Computer Science*, pages 145–161. Springer, 2003.

[39] Jianqiu Ji, Jianmin Li, Shuicheng Yan, Bo Zhang, and Qi Tian. Super-bit locality-sensitive hashing. In *Advances in Neural Information Processing Systems 25*, pages 108–116, 2012.

[40] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha P. Chandrakasan. GAZELLE: A low latency framework for secure neural network inference. In William Enck and Adrienne Porter Felt, editors, *27th USENIX Security Symposium, USENIX Security 2018*, pages 1651–1669. USENIX Association, 2018.

[41] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online, 2020. Association for Computational Linguistics.

[42] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pages 3146–3154, 2017.

[43] Weihao Kong and Wu-Jun Li. Isotropic hashing. In *Advances in Neural Information Processing Systems 25*, 2012.

[44] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[45] Vihan Lakshman, Xiaochen Zhu, Alexandra Henzinger, Henry Corrigan-Gibbs, and Emma Dauterman. Speakeasy: Billion-scale two-server private semantic search. In *2nd Workshop on Vector Databases, VecDB@VLDB 2026*, 2026.

[46] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339. Morgan Kaufmann, 1995.

[47] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[48] Dong Li, Qingguo Lü, Xiaofeng Liao, Tao Xiang, Jiahui Wu, and Junqing Le. Avpmir: Adaptive verifiable privacy-preserving medical image retrieval. *IEEE Transactions on Dependable and Secure Computing*, 21(5):4637–4651, 2024.

[49] Haoran Li, Mingshi Xu, and Yangqiu Song. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040, Toronto, Canada, 2023. Association for Computational Linguistics.

[50] Jingyu Li, Zhicong Huang, Min Zhang, Cheng Hong, Jian Liu, Tao Wei, and Wenguang Chen. PANTHER: Private approximate nearest neighbor search in the single server setting. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 365–379. ACM, 2025.

[51] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. In Subbarao Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 1711–1717. IJCAI/AAAI Press, 2016.

[52] Xiaojian Liang, Lushan Song, Shishuai Du, Weicheng Zhu, Tan Li Hui Faith, Jun Jie Sim, Haibing Jin, Zhenghao Wu, Yingting Liu, Xin Zhang, Jiang-Ming Yang, and Pu Duan. Pisces: Cryptography-based private retrieval-augmented generation with dual-path retrieval. In *International Conference on Learning Representations*, 2026.

[53] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. *Int. J. Comput. Vis.*, 127(9):1217–1234, 2019.

[54] Yingfan Liu, Yandi Zhang, Jiadong Xie, Hui Li, Jeffrey Xu Yu, and Jiangtao Cui. Privacy-preserving approximate nearest neighbor search on high-dimensional data. In *41st IEEE International Conference on Data Engineering, ICDE 2025*, pages 3017–3029. IEEE, 2025.

[55] Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

[56] Xiao Luo, Haixin Wang, Daqing Wu, Chong Chen, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. A survey on deep hashing methods. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–50, 2023.

[57] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe LSH: Efficient indexing for high-dimensional similarity search. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 950–961. ACM, 2007.

[58] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*,

pages 142–150. Association for Computational Linguistics, 2011.

[59] Yulong Ming, Mingyue Wang, Jijia Yang, Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. P^2RAG: Efficient privacy-preserving RAG service supporting arbitrary top-k retrieval. *CoRR*, abs/2603.14778, 2026.

[60] Payman Mohassel and Yupeng Zhang. SecureML: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy*, pages 19–38. IEEE Computer Society, 2017.

[61] John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. Text embeddings reveal (almost) as much as text. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore, 2023. Association for Computational Linguistics.

[62] Nguyen Linh Bao Nguyen, Wanlun Ma, Viet Vo, Alsharif Abuadbba, Minghong Fang, Jun Zhang, and Yang Xiang. Five queries are enough: Query-efficient and surrogate-free membership inference attacks on RAG via entailment. In *35th USENIX Security Symposium (USENIX Security 26)*. USENIX Association, 2026.

[63] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches at NIPS 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[64] Michele Orrù, Emmanuela Orsini, and Peter Scholl. Actively secure 1-out-of-n OT extension with application to private set intersection. Cryptology ePrint Archive, Paper 2016/933, 2016.

[65] Zijing Ou, Qinliang Su, Jianxing Yu, Ruihui Zhao, Yefeng Zheng, and Bang Liu. Refining BERT embeddings for document hashing via mutual information maximization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Findings of ACL, pages 2360–2369. Association for Computational Linguistics, 2021.

[66] Ninh Pham and Tao Liu. Falconn++: A locality-sensitive filtering approach for approximate nearest neighbor search. In *Advances in Neural Information Processing Systems 35*, 2022.

[67] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

[68] Muhammad Arslan Rauf, Mian Muhammad Yasir Khalil, Weidong Wang, Qingxian Wang, Muhammad Ahmad Nawaz Ul Ghani, and Junaid Hassan. BCE4ZSR: bi-encoder empowered by teacher cross-encoder for zero-shot cold-start news recommendation. *Inf. Process. Manag.*, 61(2):103686, 2024.

[69] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[70] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, NIST Special Publication, pages 109–126. National Institute of Standards and Technology (NIST), 1994.

[71] Sacha Servan-Schreiber, Simon Langowski, and Srinivas Devadas. Private approximate nearest neighbor search with sublinear communication. In *43rd IEEE Symposium on Security and Privacy, SP 2022*, pages 911–929. IEEE, 2022.

[72] Zhiwei Shang, Simon Oya, Andreas Peter, and Florian Kerschbaum. Obfuscated access and search patterns in searchable encryption. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021*. The Internet Society, 2021.

[73] Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna, editors, *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 377–390. ACM, 2020.

[74] Tingting Tang, James Flemings, Yongqin Wang, and Murali Annavaram. Differentially private retrieval-augmented generation. *CoRR*, abs/2602.14374, 2026.

[75] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Joaquin Vanschoren and

Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[76] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819. Association for Computational Linguistics, 2018.

[77] Sennur Ulukus, Salman Avestimehr, Michael Gastpar, Syed A Jafar, Ravi Tandon, and Chao Tian. Private retrieval, computing, and learning: Recent progress and future challenges. *IEEE Journal on Selected Areas in Communications*, 40(3):729–748, 2022.

[78] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022.

[79] Liangdao Wang, Yan Pan, Cong Liu, Hanjiang Lai, Jian Yin, and Ye Liu. Deep hashing with minimal-distance-separated hash centers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23455–23464. IEEE, 2023.

[80] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[81] Yimu Wang, Shiyin Lu, and Lijun Zhang. Searching privately by imperceptible lying: A novel private hashing method with differential privacy. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2700–2709, 2020.

[82] Benjamin Weggenmann and Florian Kerschbaum. Differential privacy for directional data. In Yongdae Kim, Jong Kim, Giovanni Vigna, and Elaine Shi, editors, *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 1205–1222. ACM, 2021.

[83] Xinpeng Xie, Chenyang Yu, Yan Huang, Yang Cao, and Chenxi Qiu. A decade of metric differential privacy: Advancements and applications. *ArXiv preprint*, abs/2502.08970, 2025.

[84] Yuanzhong Xu, Weidong Cui, and Marcus Peinado. Controlled-channel attacks: Deterministic side channels for untrusted operating systems. In *2015 IEEE Symposium on Security and Privacy*, pages 640–656. IEEE Computer Society, 2015.

[85] Timofey Yaluhin. SoK: Confidential transformer inference and retrieval-augmented generation. Cryptology ePrint Archive, Paper 2026/1544, 2026.

[86] Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. Efficient passage retrieval with hashing for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 979–986. Association for Computational Linguistics, 2021.

[87] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

[88] Mengyu Yao, Ziqi Zhang, Ning Luo, Shaofei Li, Yifeng Cai, Xiangqun Chen, Yao Guo, and Ding Li. Connect the dots: Knowledge graph-guided crawler attack on retrieval-augmented generation systems. In *35th USENIX Security Symposium (USENIX Security 26)*. USENIX Association, 2026.

[89] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis E. H. Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3080–3089. IEEE, 2020.

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

[91] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics, 2018.

[92] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, pages 649–657, 2015.

[93] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics, 2020.

[94] Mingxun Zhou, Elaine Shi, and Giulia Fanti. PACMANN: Efficient private approximate nearest neighbor search. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025.

[95] Han Zhu, Mingsheng Long, Jianmin Wang, and Yue Cao. Deep hashing network for efficient similarity retrieval. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2415–2421. AAAI Press, 2016.

[96] Jinhao Zhu, Liana Patel, Matei Zaharia, and Raluca Ada Popa. Compass: Encrypted semantic search with high accuracy. In Lidong Zhou and Yuanyuan Zhou, editors, *19th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2025, Boston, MA, USA, July 7-9, 2025*, pages 915–938. USENIX Association, 2025.

[97] Guy Zyskind, Tobin South, and Alex Pentland. Don't forget private retrieval: Distributed private similarity search for large language models. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 7–19, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Ethical Considerations

This work proposes a defense solution in a privacy-sensitive two-party scenario, which aims to *reduce* privacy exposure in dense retrieval against attacks documented in the literature [12, 35, 49, 61, 73, 90]. In particular, *no* new attack algorithms beyond the reproduction and adaptation of existing work are presented.

Dual-use of attack reproduction. We re-implement published embedding-inversion attacks [49,73,90] to measure the exposure of the representations without protection and confirm the concerning leakage among them. The experiments reproduce existing attack capabilities on public benchmarks and pretrained encoders; they use *no* proprietary corpus or production system.

Use of human-generated data. All datasets are publicly available research benchmarks released for academic use [46, 58, 75, 91, 92]. The study collects *no* new data and involves *no* interaction with human subjects.

Security boundary. The formal query guarantee covers an honest-but-curious Owner observing the quantized randomized code, candidate set, and protocol metadata. For a conforming secret-key User, compact ciphertext-simulatable BFV restricts the scoring ciphertext to the explicit K-score oracle, while active-secure OT limits a malicious receiver's per-query payload-key recovery. We note that the paper source code only presents a research artifact that meets the declared security guarantees and it might miss additional consideration that must be in place for real-world deployment. For example, commercial deployment shall additionally bind identities, protect endpoint keys, authenticate and encrypt both transport channels, and address side channels beyond the measured message lengths and timing.

Responsible disclosure. The evaluation reproduces public attacks and addresses the general bi-encoder retrieval pipeline rather than a product-specific vulnerability, so coordinated vendor disclosure is not applicable.

A Proof Details and Auxiliary Calibration

A.1 Analytic Gaussian Baseline

The additive Gaussian baseline acts on the bounded, unnormalized hash-head vector $h \in [-1,1]^L$ rather than on the direction protected by Theorem 1. For $G_\sigma(h) = h + \mathcal{N}(0, \sigma^2 I)$, two inputs at Euclidean distance $D > 0$ have the exact privacy profile [4]

$$\delta_G(\varepsilon, D, \sigma) = \Phi\left(\frac{D}{2\sigma} - \frac{\varepsilon\sigma}{D}\right) - e^{\varepsilon}\Phi\left(-\frac{D}{2\sigma} - \frac{\varepsilon\sigma}{D}\right), \quad (13)$$

with $\delta_G = 0$ at $D = 0$. The profile increases with D, so calibration at Euclidean radius ρ_h solves $\delta_G(\varepsilon, \rho_h, \sigma) \le \delta_0$ at the boundary.

Table 11 gives the numerically solved scales used by our implementation for $\rho_h = 2$ and $\delta_0 = 10^{-6}$. Each regression test substitutes the result into Equation 13 and checks the target profile, and the retrieval and attack sweeps use these calibrated scales.

Gaussian and vMF results protect different input spaces: G_σ uses Euclidean adjacency on h, whereas M_κ uses angular or chord adjacency on $u = h/\|h\|_2$. Their numerical privacy budgets must therefore be reported with the protected space and radius, not as a mechanism ranking under an unspecified common adjacency.

Table 11: Analytic Gaussian scales for bounded pre-sign representations at Euclidean radius $\rho_h = 2$ and $\delta_0 = 10^{-6}$.

ε	8	16	32	64
σ	1.3059	0.7372	0.4324	0.2638

A.2 Approximate-vMF Calibration

For mean directions separated by angle θ, rotate coordinates so that $u = e_1$ and $u' = \cos\theta\, e_1 + \sin\theta\, e_2$. Under $Y \sim \mathrm{vMF}(u, \kappa)$, the privacy-loss random variable is

$$L(Y) = \kappa\big((1-\cos\theta)Y_1 - \sin\theta Y_2\big). \tag{14}$$

The corresponding pre-binarization hockey-stick divergence is

$$\begin{aligned}\delta_{\mathrm{vMF}}(\varepsilon, \theta, \kappa) = &\Pr_u\big[L(Y) > \varepsilon\big] \\ &- e^{\varepsilon} \Pr_u\big[L(Y) < -\varepsilon\big].\end{aligned} \tag{15}$$

A radius-ρ approximate guarantee requires the supremum of Equation 15 over every $\theta \in [0, \rho]$. Our quadrature evaluates the boundary, while certified approximate calibration additionally requires establishing the maximizing angle. The protocol uses the exact pure calibration of Theorem 1; the RDP-vMF sweep serves as the empirical mechanism comparison.

A.3 Packed-Layout Invariants

At degree 8192, BFV batching provides two rows of 4096 slots. Each row is divided into four 1024-slot segments, so a score group contains eight candidates. Query encryption repeats the d-coordinate vector in every segment and fills the remaining slots with zero; Owner encoding places one candidate in each corresponding segment.

After componentwise plaintext–ciphertext multiplication, rotations by $1, 2, \ldots, 512$ and additions place each segment's dot product at its anchor. The multi layout returns that group ciphertext directly, yielding $\lceil K/8 \rceil$ ciphertexts; it specifies correctness only at the anchors and does not zero the remaining slots. The compact layout multiplies by a plaintext mask that retains only the eight anchors, rotates group g to residue g mod 1024, and adds 1024 groups per output ciphertext. Different groups then occupy different residues at every segment anchor, yielding $\lceil K/8192 \rceil$ ciphertexts. The final partial group and partial compact output contain only zero-padded dummy candidates.

These layout invariants establish where each modular score is decoded; Equation 9 establishes that its centered residue is the intended signed integer. Circuit correctness additionally assumes sufficient BFV noise budget, which the implementation checks empirically through successful decryption and oracle equality over both concrete layouts. Circuit privacy and malicious-input validity are separate properties governed by the scope in §6.3.

A.4 Cryptographic Hybrid Details

For Theorem 2, begin with the real Owner view for a conforming User. Receiver privacy of active OOS OT replaces the receiver-choice-dependent messages with a simulated transcript for the same sender inputs. BFV IND-CPA replaces the valid encryption of the clean int8 query by an encryption of zero; applying the public scoring circuit and serializing its output cannot increase the Owner's distinguishing advantage. The remaining query-dependent plaintext is $M_\kappa(u)$ and its Hamming-search and quota post-processing. Directional privacy gives Equation 8, while the auxiliary fields in Equation 4 are held fixed. Theorem 2 therefore governs the Owner view; §6.3 separately specifies the secret-key User view.

For Theorem 5, active receiver security restricts each row to one OT option key. In the random-oracle hybrid, the hash of every unchosen option key is independent of the receiver's view, making its masked content key uniform. Replacing the unselected derived payload keys by random keys then reduces disclosure of any remaining plaintext to the confidentiality of ChaCha20–Poly1305; ciphertext modification is rejected by its authentication check. Across k rows, the receiver obtains at most k content keys, with duplicate selections yielding fewer distinct payloads.

B Experimental Details

B.1 Experimental Setup Details

We provide the training and evaluation details needed to reproduce the two-forward retrieval results in Table 3.

Evaluation Corpora. The five BEIR corpora originate from SciDocs [18], Natural Questions [44], DBpedia-Entity [29], FEVER [76], and Climate-FEVER [21]. Climate-FEVER adapts FEVER's claim–evidence methodology from artificially constructed general-domain claims to real-world climate claims and includes disputed evidence. Their dataset corpus are largely overlapping, but they have distinct query distributions.

Metrics. NDCG@k is the standard normalized discounted cumulative gain at rank k, which rewards relevant documents appearing higher in the ranked list and normalizes against the ideal ordering. Recall@k is the fraction of gold-relevant documents appearing in the top-k results. In particular, the former is sensitive to ranking among the top-k, while the latter is only sensitive to whether the relevant documents appear, regardless of their specific ranking.

Models and Hard Negatives. E5-base-v2 uses mean pooling and a 256-bit hash head, whereas BGE-base-en-v1.5 uses its CLS representation and a 512-bit head. We retrieve 512 BM25 [70] candidates per MS MARCO training query, rerank them with the "ms-marco-MiniLM-L6-v2"[1] cross-encoder [80], retain the top 32, form a pool from the five

[1] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

highest-ranked non-relevant passages, and sample three negatives per query. A mined negative is removed from the binary ranking loss when its cross-encoder score is within 0.5 of the labeled positive. We also tried stronger cross-encoders, larger mining pools, or more hard-negatives per query during our experiments, but none of them gives noticeably better results, while more hard-negatives even harm the evaluated quality.

Training Objective. A batch contains queries q_i, labeled passages p_i, and mined-negative sets $\mathcal{N}_i$ for $i \in \{1,\dots,B\}$. Let $\mathbf{a}_x$ and $\mathbf{a}_x^0$ denote the normalized pooled representations of text x from the adapted and frozen encoders, respectively. The hash head produces logits $\mathbf{z}_x = W\,\mathrm{pool}(E_{\mathrm{LoRA}}(x))$ and training code $\mathbf{h}_x = \tanh(\beta \mathbf{z}_x)$; deployment uses $\mathbf{b}_x = \mathrm{sign}(\mathbf{z}_x)$, where $\mathrm{sign}(t) = +1$ for $t \geq 0$ and -1 otherwise. We write

$$\begin{aligned} s^a(q,d) &= \mathbf{a}_q^\top \mathbf{a}_d, \\ s^0(q,d) &= (\mathbf{a}_q^0)^\top \mathbf{a}_d^0, \\ s^h(q,d) &= \mathbf{h}_q^\top \mathbf{h}_d / L. \end{aligned} \tag{16}$$

and let $c(q,d)$ be the cross-encoder relevance logit. For a score function s, candidate set $\mathcal{D}$, and temperature τ, define

$$P_{i,\tau}^{s,\mathcal{D}}(d) = \frac{\exp(s(q_i,d)/\tau)}{\sum_{d' \in \mathcal{D}} \exp(s(q_i,d')/\tau)}. \tag{17}$$

Let $\mathcal{D}_B = \{p_j\}_{j=1}^B \cup \bigcup_j \mathcal{N}_j$ be all passages in the batch. The implemented groups in Equation 3 expand as

$$\begin{aligned} \mathcal{L}_{\mathrm{retrieval}} &= \mathcal{L}_{\mathrm{InfoNCE}} + \lambda_{\mathrm{bin}} \mathcal{L}_{\mathrm{bin}}, \\ \mathcal{L}_{\mathrm{transfer}} &= \lambda_{\mathrm{rankKD}} \mathcal{L}_{\mathrm{rankKD}}, \\ \mathcal{L}_{\mathrm{regularization}} &= \lambda_{\mathrm{floatKD}} \mathcal{L}_{\mathrm{floatKD}} + \lambda_{\mathrm{GOR}} \mathcal{L}_{\mathrm{GOR}}. \end{aligned} \tag{18}$$

The continuous retrieval term is

$$\mathcal{L}_{\mathrm{InfoNCE}} = -\frac{1}{B} \sum_{i=1}^{B} \log P_{i,\tau}^{s^a,\mathcal{D}_B}(p_i). \tag{19}$$

For false-negative margin m, we retain $\mathcal{M}_i = \{n \in \mathcal{N}_i : c(q_i,n) < c(q_i,p_i) - m\}$ and define $I = \{i : \mathcal{M}_i \neq \varnothing\}$. The ranking term applied to the deployed binary codes is

$$\begin{aligned} \ell_i &= \tau_b \log \sum_{n \in \mathcal{M}_i} e^{(s^h(q_i,n) - s^h(q_i,p_i))/\tau_b}, \\ \mathcal{L}_{\mathrm{bin}} &= \frac{1}{|I|} \sum_{i \in I} \mathrm{softplus}(\ell_i/\tau_b). \end{aligned} \tag{20}$$

Here $\mathrm{softplus}(u) = \log(1+e^u)$. E5 uses this direct Hamming-space term; BGE obtains stronger candidate recall from RankKD alone and sets $\lambda_{\mathrm{bin}} = 0$.

RankKD (ranking knowledge distillation) transfers the adapted continuous ranking over all in-batch passages:

$$\mathcal{L}_{\mathrm{rankKD}} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{KL}\left(P_{i,\tau}^{s^a,\mathcal{D}_B} \,\middle\|\, P_{i,\tau}^{s^h,\mathcal{D}_B} \right). \tag{21}$$

RankKD transfers batch-wide ordering without using labels from an evaluation corpus; the cross-encoder contributes difficult local examples through hard-negative mining.

The geometry terms stabilize the representation that supplies these rankings. For the batch collections $\mathcal{X}_Q = \{q_i\}_i$, $\mathcal{X}_P = \{p_i\}_i$, and $\mathcal{X}_N = \biguplus_i \mathcal{N}_i$, FloatKD (float-embedding distillation) aligns the adapted and frozen encoders:

$$\mathcal{L}_{\mathrm{floatKD}} = \frac{1}{3} \sum_{G \in \{Q,P,N\}} \frac{1}{|\mathcal{X}_G|} \sum_{x \in \mathcal{X}_G} \left[1 - \cos\left(\mathbf{a}_x, \mathbf{a}_x^0\right) \right]. \tag{22}$$

For any batch collection $\mathcal{X}$, GOR (global orthogonal regularization) uses the spread-out penalty

$$R_{\mathrm{GOR}}(\mathcal{X}) = \frac{1}{|\mathcal{X}|(|\mathcal{X}|-1)} \sum_{\substack{x,y \in \mathcal{X} \\ x \neq y}} \cos^2(\mathbf{h}_x, \mathbf{h}_y), \tag{23}$$

We use $\mathcal{L}_{\mathrm{GOR}} = [R_{\mathrm{GOR}}(\mathcal{X}_Q) + R_{\mathrm{GOR}}(\mathcal{X}_P) + R_{\mathrm{GOR}}(\mathcal{X}_N)]/3$. FloatKD retains the pretrained geometry during LoRA adaptation, whereas GOR discourages different batch examples from collapsing to similar directions before binarization.

Discrete Optimization and Hyperparameters. The reported models use the differentiable code $\mathbf{h} = \tanh(\beta \mathbf{z})$ throughout training; β rises linearly from 1 to 2.5 during the first quarter and then remains fixed, and deployment applies $\mathrm{sign}(\mathbf{z})$. We apply LoRA to every encoder linear layer with rank 16, $\alpha = 16$, and dropout 0.05. Each epoch contains 300 steps at batch size 128, and both models train for 16 epochs. The encoder learning rate is 2×10^{-6} for E5 and 5×10^{-6} for BGE, while the hash-head rate is 2×10^{-4} for both. InfoNCE has unit weight; $\lambda_{\mathrm{bin}} = 0.8$ for E5 and 0 for BGE. We set $\lambda_{\mathrm{rankKD}} = \lambda_{\mathrm{floatKD}} = \lambda_{\mathrm{GOR}} = 1$, with $\tau = 0.05$, $\tau_b = 0.1$, and $m = 0.5$. Evaluation uses an exponential moving average with decay 0.999.

Two-Forward Evaluation. Stage 1 runs the trained hash encoder to retrieve exact Hamming top-K candidates. Stage 2 independently runs the original pretrained encoder and reranks only those candidates by float similarity; it never uses the hash model's continuous representation. Both forwards share tokenization and host-to-device transfer. E5 uses the standard "query:" and "passage:" prefixes, BGE uses its standard query instruction and no document prefix, and maximum query and document lengths are 48 and 512 tokens. We evaluate in BF16 with the LoRA weights merged into the Stage 1 encoder.

DP Retrieval Sweeps. We evaluate E5 and BGE on SciDocs, NQ, and FEVER at $K \in \{200, 500, 1000, 2000, 3000\}$. Gaussian, RDP-vMF, and pure-vMF use $\varepsilon \in \{8, 16, 32, 64\}$; Gaussian and RDP-vMF set $\delta = 10^{-6}$, while pure-vMF uses the exact calibration in Theorem 1. Each randomized query code is searched by exact Hamming distance, and the unchanged pretrained encoder reranks the resulting candidates.

Hash Baselines. Table 4 uses the same cached pretrained E5 query and corpus embeddings and the same exact Hamming and Stage 2 routines as Table 3. Direct sign thresholds

all 768 embedding coordinates at zero. Random-hyperplane LSH [7] uses a fixed 256-column Gaussian projection; Super-Bit LSH [39] orthogonalizes all 256 columns as one maximum-depth Super-Bit. PCA-sign, ITQ [28], and IsoHash [43] are fitted on normalized embeddings of the first 20,000 MS MARCO corpus passages, with 50 ITQ and 100 IsoHash rotation updates, then applied without target-corpus fitting. Every Stage 2 score uses the unchanged full-precision pretrained E5 embedding, so the reported differences come only from candidate membership in Stage 1.

B.2 Qualitative Candidate-Set Cases

Table 12 extends Table 2 with three NQ queries under the same E5 pure-vMF operating point. The nearest Hamming items preserve broad cues—constitutional amendments, elevation and battlefields, or quarters of a year—but do not supply the requested count, location, or calendar rule. The answer-bearing passages occur much deeper in the coarse ranking and move into the User's top ten only after clean-query Stage 2 scoring.

Together with Table 2, these cases show the intended resolution split: coarse-code proximity can reveal a subject or isolated lexical cues, while exact answer selection depends on the clean representation evaluated inside the protected scoring stage.

B.3 Attack Implementation Details

We specify the attacker's observations, optimization procedure, and output for each evaluated attack.

Search-Based Inversion. We evaluate the embedding-guided adversarial-decoding and iterative-refinement stages of ZSInvert [90]. Given a released target representation, the attacker initializes Llama-3.1-8B-Instruct [55] with the prompt "Write a factual passage." At each decoding step, the LLM proposes its ten most likely next tokens for every active partial passage; the target encoder maps each expansion into the released representation space, similarity to the target ranks the expansions, and the best 50 remain in the beam. The highest-scoring completed passage becomes the seed for the next round, whose prompt asks for a factual passage similar to that seed. We run six search rounds with at most 80 token steps per round. Float targets use cosine similarity, while hash targets use normalized hash dot product, an affine transformation of Hamming distance. A separate MiniLM verifier [80] measures semantic similarity between the final reconstruction and the private target text; we report the mean over 100 sampled targets.

Generation-Based Inversion. Following GEIA [49], the attacker collects auxiliary PersonaChat [91] text–representation pairs and trains an embedding-conditioned DialoGPT-medium [93] decoder. A learned linear layer maps each released representation to the decoder hidden dimension and prepends it as a pseudo-token before the ground-truth token embeddings; teacher-forced cross-entropy then trains both the projection and causal LM to predict the original passage. We train a separate decoder for each released representation for 10 epochs at a learning rate of 1×10^{-5} and a batch size of 16. At test time, the held-out representation alone conditions width-5 beam decoding for at most 50 new tokens, producing one reconstruction per target; Table 8 reports the test-split mean.

Property Inference. Following Song and Raghunathan [73], the attacker receives an auxiliary collection of Stage 1 releases with known attributes, fits supervised probes on those representation–attribute pairs, and predicts the attribute of independently held-out victim releases. We test four-way topic classification on AG News [92], binary sentiment on IMDB [58], and 50-way closed-set authorship attribution on 20 Newsgroups [46]. For authorship, the label is the email address in the `From:` header among the 50 most prolific authors; we remove the complete header block before encoding the message body. Every sample is encoded as the deployed Stage 1 query release using the current E5 hash checkpoint, the "query:" prefix, a 48-token limit, and the mechanism under evaluation. Each of five seeds creates stratified 60/20/20 attacker-training, validation, and victim-test partitions. Logistic regression, a two-hidden-layer MLP, and LightGBM [42] are fitted on the attacker-training split; validation macro F1 selects the architecture, and Table 10 reports its victim-test macro F1 averaged across seeds.

B.4 Qualitative Inversion Cases

Table 13 reports the complete decoded strings for the case in Table 9 and three additional targets. Across the four cases, the float release supports reconstruction of specific entities, facts, and relations. The learned hash usually retains a broad topic or lexical association while dropping identifying details. Gaussian randomization removes even that stable association, producing outputs unrelated to the target.

Cases A and B show how a low-bit release can preserve form-level cues such as an organizational acronym while losing the entity itself. Case C retains the broad calorie-deficit relation but changes the quantities and time scale, while Case D retains chemistry vocabulary but drops the defining electron-transfer and redox relation. In every case, the Gaussian reconstruction switches to a different subject, matching the near-zero verifier cosine.

C Additional Retrieval Results

Figure 8 shows how the candidate budget recovers ranking quality at the representative $\varepsilon = 32$ operating point. Gaussian approaches the no-DP curve by $K = 1000$ on all three datasets, while pure-vMF continues to benefit from larger candidate pools on NQ and FEVER. The horizontal float references

Table 12: Three additional NQ candidate-set cases under E5 pure-vMF ($\varepsilon = 64$, $K = 3000$, $k = 10$, 256 bits). Stage 2 entries show coarse Hamming rank → clean-float rank.

Case	View	Rank / d_H	Target or passage excerpt
A	Target	–	*How many amendments to the Constitution have there been?*
	DP-Ham.	1 / 60	*First Amendment to the United States Constitution:* "The civil rights of none shall be abridged . . . "
	DP-Ham.	2 / 60	*Limited government:* "The Ninth and Tenth Amendments . . . "
	DP-Ham.	3 / 60	*Article Three of the United States Constitution:* "Hamilton continues . . . "
	Stage 2	1742→1 / 82	*List of amendments to the United States Constitution:* "Thirty-three amendments . . . have been proposed . . . Twenty-seven . . . are part of the Constitution."
B	Target	–	*Where is the world's highest battlefield located?*
	DP-Ham.	1 / 68	*Operation Market Garden:* "The country was wooded and rather marshy . . . two important hills . . . represented some of the highest ground in the Netherlands."
	DP-Ham.	2 / 70	*List of elevation extremes by country:* "The Dead Sea is the lowest point on Earth."
	DP-Ham.	3 / 70	*Geography of China:* "Tallest mountain peaks."
	Stage 2	2317→2 / 87	*Siachen Glacier:* "The glacier's region is the highest battleground on Earth . . . Pakistan and India . . . maintain a permanent military presence."
C	Target	–	*Explain what happens to the extra quarter of a day each calendar year.*
	DP-Ham.	1 / 67	*Calendar year:* "The calendar year can be divided into four quarters . . . "
	DP-Ham.	2 / 69	*Fiscal year:* "The Financial year is split into the following four quarters."
	DP-Ham.	3 / 71	*Accounting period:* "The end of the fiscal year would move one day earlier . . . "
	Stage 2	555→8 / 88	*Leap year:* "Adding one extra day in the calendar every four years compensates for . . . almost 6 hours."

expose both the remaining candidate loss and the point at which increasing K saturates.

Tables 15 and 16 report every mechanism, privacy budget, dataset, and candidate budget used in the retrieval evaluation. Both encoders exhibit the same operating pattern: approximate mechanisms saturate at smaller K, while formal pure-vMF requires a larger pool at tighter budgets and converges toward the float reference as ε increases.

C.1 Cross-Encoder Compatibility

The pretrained E5 scorer produces the top ten Natural Questions passages, matching the protocol's $k = 10$ payload allowance. The User opens those passages through OT, locally reranks them with the MS MARCO MiniLM-L-6-v2 cross-encoder [80], and supplies the top five to Qwen3-32B. The cross-encoder consumes the clean query and authorized plaintext payloads entirely on the User side.

Table 14 shows that client-side reranking raises answer-bearing context coverage by 0.6–1.4 points and improves EM by at least 3.0 points. Every protected variant remains within 0.4 EM and 0.35 F1 of the cross-encoder float reference, so the learned filter composes with a standard retrieve–rerank–generate stack while retaining the protocol's payload-access bound.

D Protocol Implementation and End-to-End Latency

This appendix presents the complete message sequence (§D.1), protocol implementation (§D.2), latency methodology (§D.3), and matched comparison with RemoteRAG [14], PANTHER [50], and P^{2}RAG [59] (§D.4) [14, 50, 59].

D.1 Complete Message Sequence

Figure 9 expands the compact flow in Figure 2 into the complete authenticated round, including session setup, round binding, quota reservation, OT extension, payload delivery, and commit.

D.2 Protocol Implementation

BFV scoring. Following the Homomorphic Encryption Standard's 128-bit classical-security parameters [1], the Microsoft SEAL implementation uses polynomial-modulus degree 8192, a 25-bit batching plaintext modulus, and coefficient-modulus chains totaling 109 bits for the multi layout and 180 bits for the compact layout, within the recommended 218-bit ceiling. The exact signed-int8 contract clamps both operands to $[-127, 127]$, giving $|\langle \tilde{\mathbf{q}}, \tilde{\mathbf{z}} \rangle| \le 127^2 d$ and exact agreement with an int32 oracle for $d = 768$. The multi layout uses coefficient-modulus bits $(40, 40, 29)$ and 1024-slot segments, placing eight candidates in each output ciphertext with ten rotate-and-add steps. The compact layout uses $(50, 40, 40, 50)$, masks each segment anchor, and rotates anchors into dense residues, reducing score traffic at the cost of a second multiplicative level. Layout-specific Galois keys contain only the required rotation steps. The protected compact path follows Hwang et al.'s PMultE and Rand construction [36]: it samples each plaintext lift from the appropriate modulo-t discrete-Gaussian coset, adds the prescribed rounded continuous-Gaussian multiplication error, and adds one public-key randomizer before the public rotation, mask, and addition subcircuit. The discrete sampler uses

Table 13: Complete search-based inversion outputs for four MS MARCO targets. Case A is the example summarized in Table 9; Cases B–D are additional examples. Cosine is measured by the independent verifier.

Case	Release	Complete target or reconstruction	Cosine
A	Target	Welcome to the U.S. High School Bowling Foundation. The U.S. High School Bowling Foundation (USHSBF) is an IRS 501 (c)(3) non-profit organization that actively promotes the growth of high school bowling throughout the country.	–
	Float	US Bowling Foundation. The nonprofit organization. USABF or HS, is an (non profit.)501, 501(C). US bowling (foundation) or the bowling (foundation) (U.s)	.903
	Learned hash	US Bank is affiliated by National Bank Union(UFBG,NFUB or more informlly refererd UBFGU) through UBFBU-UnitedStatebannkinngunitiationthough there name oficialally stands out too - unitesstatebunkfinfinitation. However alot offpeople recognizeit with thershorthandletternamr ofunsttobbankor	.464
	Gaussian, $\varepsilon = 8$	The country Australia encompasses expansive territories spanning over seven major regions with diverse geology, resulting form varied processes across historical timescale having several distinct natural environments within different territoories - Northern Territories comprising land that shares the Timoresse Sea and extends from southern coast to Indonesian borders as well northern territory known as Western Artesia, along the Great Victoria river system in addition the southeastern coastline forms extensive shoreline along	−.060
B	Target	What does FGA stand for? What does FGA mean? This page is about the various possible meanings of the acronym, abbreviation, shorthand or slang term: FGA.	–
	Float	FAA acronym definition abbreviation. What meaning? short name F.A. Association meaning? What stands: stands for.	.697
	Learned hash	The Aviation Alphabet Standards for pilots. A commonly accepted and well-diffused standardization was put forward with these phrases - Uniform - Mike Victor Lindberqg (L), Si-gmah Si-bay, Fo-sure-Fo-sho Fos- ter (R).	.204
	Gaussian, $\varepsilon = 8$	An Intergregaler Frost (also sometimes spelled Yuki-On) event periodically transpires across Arctic archipelagos located within northern Norway towards summations occurring after equinocial solar events.	.031
C	Target	One pound of fat is 3,500 calories. If you simply eat 500 calories less per day, then in seven days that adds up to a 3,500 calorie deficit and you'd have lost one pound of fat. Fitness is science, not magic. (7 days x 500 calorie deficit = 3,500)	–
	Float	seven days calories deficit per pound fat approximately =500 calorie workouts three * day eating one-pound less * is 500 calorie one day * equals fat one day loss * (500* (1 pound * (7 days * (	.879
	Learned hash	The ideal calorie deficit for fat and muscular pounds weightloss is approximately three pounds or so within eight weeks at rate equivalent of nearly one-further half pound in weight each seven-days.	.732
	Gaussian, $\varepsilon = 8$	Antarctica Ross Land Sea ice field exhibits groundbreaking discoveries showcasing glaucial layers with crucial data suggesting expansive glacier formations reaching up to fifty seven-and twenty-two hundred-thoussd-year intervals dating around eleven and ninety eight miiiillion-earth years into the distant prehistirical.	−.004
D	Target	Alkali metals react with nonmetals to form ionic compounds. In these types of reactions, the alkali metal gives up its outermost electron to a nonmetal that is greedy for electrons. Reactions like this that involve an element exchanging an electron with another is called an oxidation-reduction or redox reaction.	–
	Float	Reactive alkalide metals reaction nonmetal oxido make other. An Ionic forms bonds elements ox and compounds oxidic and metal nonreducible react with. Alternatively, reactive acid nonmetals make other compounds.	.824
	Learned hash	Throughout chemistry ions form and transform, resulting in diverse mechanisms of conductivity modification as atoms exhibit changes of state along the way.	.413
	Gaussian, $\varepsilon = 8$	The selenicerids (specific member from order of braniophorous gastreoids including species classified under Thylakocysticida) exhibit unique characteristics across various subtaxons found residing throughout multiple ecosystems such as cold sea floors adjacent volcanic regions and along with their habitat ranging near geothermic hot vent zones, abysses of ocean and river-mouth habitats located primarily close by island nations and	.052

Table 14: End-to-end RAG with client-side cross-encoder reranking on 500 NQ queries. The opened top ten are reranked to the five passages supplied to Qwen.

Retrieval	(ε, K)	Hit	EM	F1
Full-corpus float	(∞, N)	89.2	53.2	65.36
No-DP hash	$(\infty, 500)$	88.6	53.6	65.68
Gaussian	$(16, 3000)$	89.0	53.6	65.71
RDP-vMF	$(16, 3000)$	89.2	53.2	65.35
Pure-vMF	$(64, 3000)$	89.0	53.2	65.34

a high-precision 192-bit cumulative distribution table (CDT) truncated at eight times its Gaussian parameter; tables are constructed once when the Owner daemon starts and reused across rounds. The User creates a fresh symmetric query ciphertext, while the Owner holds only public and rotation keys. Compact masking makes every non-score plaintext slot zero, and randomized evaluation makes the full ciphertext view simulatable as stated in Theorem 4. The multi and fresh-zero paths remain comparison variants.

Long-lived HE roles. A User daemon retains the BFV context, public key, and secret key and handles symmetric query encryption and parallel score decryption; an Owner daemon retains the public context, public key, Galois keys, and randomized-evaluation samplers and handles candidate scoring. Both daemons load their key material once, bind each command to an exact round identifier, and terminate on malformed input or computation failure. The default harness assigns 16 OpenMP threads to Owner scoring and 8 to User decryption, with one SEAL evaluator, decryptor, or batch encoder per worker where required.

Active-secure key transfer and payload protection. The OT backend uses libOTe's OOS active-secure 1-out-of-N extension [64] with a 40-bit statistical check and a 16-bit option index, supporting project candidate budgets up to $K = 60000$. Following the IKNP OT-extension organization [38], the parties establish a small public-key base-OT correlation once per long-lived connection and derive each query's k fresh extension rows with symmetric-key work. Each query advances the

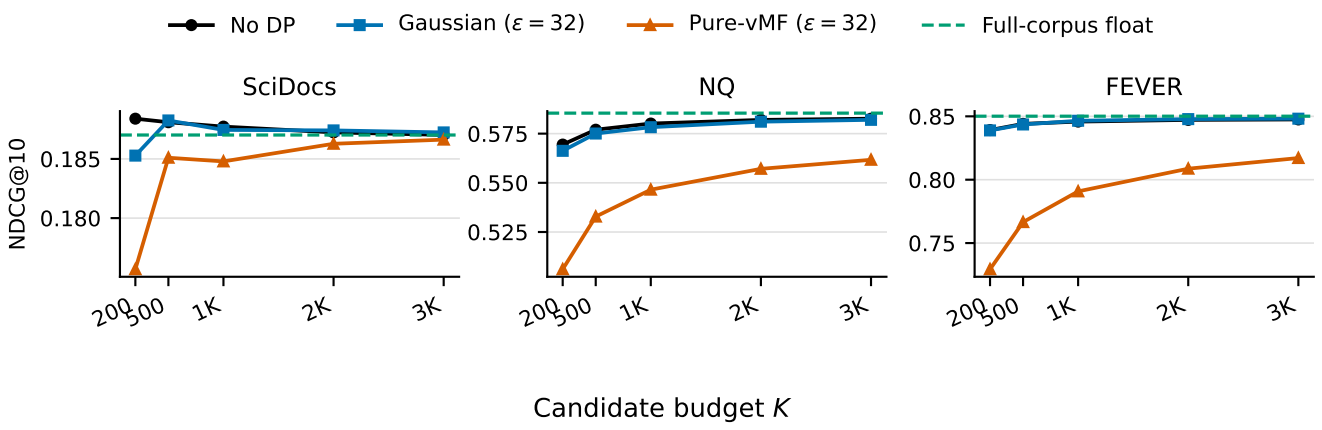

Figure 8: E5 two-forward retrieval versus candidate budget. Gaussian and formal pure-vMF use $\varepsilon = 32$; the full-corpus float reference is independent of K.

internal PRG state and performs a nonzero-challenge malicious check. The C++ sender derives and masks the full $k \times K$ option table before serialization, so raw option keys remain inside the Owner process. Each table entry XOR-masks a 128-bit content key with the first 128 bits of SHA-256 applied to a domain-separated OT key. Document plaintexts contain an encrypted true-length field, are padded in 4096-byte units, and are protected by one-shot ChaCha20–Poly1305 under an HKDF–SHA-256-derived key.

Network path and state. The network design exchanges versioned frames carrying a session identifier, message type, round identifier, HE layout, K, and a length-capped payload. Session establishment binds the layout and K. A COARSE frame starts Hamming search while the User computes the pretrained representation; after C_K is fixed, the Owner sends PAYLOADS on the ordinary channel while the User sends the separate SCORING-QUERY frame and the Owner computes SCORES. The parties then execute OT, the Owner sends OT-MASKED, and the User opens the selected buffered payloads before DONE. TCP delayed-ACK batching is disabled for these latency-sensitive control frames. Header fields and the 64-MiB payload cap are validated before payload allocation or blocking reads, and either-side failure closes the session and poisons reusable cryptographic state. With one 4096-byte payload block per candidate, the payload frame supports $K \leq 16256$; this bound exceeds the candidate budgets evaluated in the paper.

D.3 Latency Measurement Methodology

We use two complementary measurement experiments. The corpus-level comparison runs all four protocols against the same E5-base-v2 query and document embeddings on SciDocs, Webis-Touché, and NQ-1M. A separate two-process harness exercises our complete Owner/User message path after query-embedding generation and checks every recovered payload byte-for-byte against the retrieval oracle. Following private-retrieval benchmarking practice, we separate reusable key/index preprocessing from steady-state online latency while including every per-query encryption, HE evaluation, OT extension, payload transfer, and serialized byte [50, 59].

Latency and traffic accounting. Let T_{post} contain encrypted-score return, local decryption and selection, OT, and masked-key delivery. The pipelined critical path is $T_{\text{hash}} + T_{\text{DP}} + \max\{T_{\text{Hamming}} + T_{\text{payload}}, \max(T_{\text{Hamming}}, T_{\text{pre}} + T_{\text{enc}}) + T_{\text{score}} + T_{\text{post}}\} + T_{\text{open}}$, where transfer terms are charged at the evaluated bandwidth. Payload bytes therefore overlap scoring rather than appearing as a serial suffix. Figure 3 reports this path, while Figure 4 adds Qwen3-32B generation. Query-side Gaussian, RDP-vMF, and pure-vMF release leave the subsequent message flow unchanged; the DP rows use the largest measured release time, 0.92 ms for vMF.

Measurement scope. Long-lived User and Owner BFV daemons keep secret and public evaluation material in their respective roles, while the active-secure libOTe daemons reuse base-OT state across rounds. The protocol-stage experiment uses $N = 25{,}657$ synthetic rows and checks the complete HE-score, OT-key, and AEAD-payload chain against an int32 and byte-for-byte oracle. The RAG experiment measures 20 full-corpus float queries, 30 exact Hamming queries over all 2.68M NQ codes, and five generation prompts with 559–1178 input tokens; Qwen3-32B occupies 61.5 GiB. Network latency is computed from exact serialized byte counts; Figure 3 uses 10 Gbps and Table 7 reports the 100-Mbps, 1-Gbps, and 10-Gbps projections.

Hardware. Measurements were run on an AMD EPYC 9654 server with NVIDIA A100 GPUs under Linux 6.14. Encoder timing uses the default BF16, merged LoRA weights, 10 warm-up queries, and 50 measured queries. The BFV Owner uses 16 OpenMP threads, the User decryptor uses 8, and each cryptographic cell discards one full warm-up round before 50 measured rounds; the $K = 500$ breakdown uses 50 measured rounds.

Table 17: Qwen3-32B generation length and the $K = 500$ protection overhead. Times are seconds per query.

Tokens	Generation	Plaintext	Protected	Overhead
1	.298	.307	.497	61.9%
32	2.024	2.033	2.223	9.3%
128	7.296	7.305	7.495	2.6%
256	14.438	14.447	14.637	1.3%

D.4 Matched Protocol Comparison

Common benchmark contract. All corpora use the same normalized 768-dimensional E5-base-v2 [78] vectors and top-10 output: 1,000 SciDocs queries over 25,657 documents, 49 Webis-Touché queries over 382,545 documents, and all 3,452 NQ queries over NQ-1M, which retains every judged-relevant NQ document and fills the remaining positions from the corpus's original order. Query encoding is excluded because it is identical across methods; online cryptographic setup, candidate search, secure scoring, selection, and serialized traffic are included. Long-lived keys and sessions exclude one-time setup. Figure 5 demonstrates the exact traffic at 100 Mbps, while Table 18 uses 1 Gbps; the P^{2}RAG projection additionally charges 0.1 ms RTT for each declared online round because it requires two non-colluding servers.

Table 18: Matched online latency at 1 Gbps in seconds per query. Ours uses the smallest K retaining at least 99% of float NDCG@10: 292, 104, and 1956, respectively. P^{2}RAG includes both client–server and inter-server transfer.

Method	SciDocs	Touché	NQ-1M
Ours	.152	.071	.796
P^{2}RAG	.155	2.254	5.012
RemoteRAG	3.001	5.256	7.167
PANTHER	8.126	20.389	OOM

Our protocol. We use pure-vMF at $\varepsilon = 64$, randomized-evaluation compact BFV scoring, and active-secure 10-out-of-K key transfer. After unit-resolution refinement at each 99% boundary, the smallest evaluated budgets are $K = 292$ on SciDocs (0.18533 versus 0.18702), $K = 104$ on Touché (0.24753 versus 0.24954), and $K = 1956$ on NQ-1M (0.62768 versus 0.63385). Each value uses the measured 32-thread full-index Hamming scan and the matched candidate-bound cryptographic path.

P^{2}RAG [59]. We use the authors' official implementation[2], compile the authors' retrieval kernel and expose only N, d, and the bisection depth as runtime parameters. We set $k' = 16$ so its revealed set contains the requested top-10, measure five online compute runs, and combine their median with the exact user–server and inter-server byte formulas from the paper's Table 2. Its kernel uses all 192 hardware threads, compared with 16 threads for our BFV scorer, and the trusted-dealer preprocessing remains offline as specified by P^{2}RAG; these two factors favor its online result.

RemoteRAG [14]. We reproduce the paper's spherical-cap shortlist geometry and Paillier encrypted cosine with a 1024-bit modulus, `gmpy2` modular exponentiation, and process-parallel query encryption and inner products. At $\varepsilon = 15360$ and $k = 10$, the mean shortlist sizes are 627.3 on SciDocs, 1,581.4 on Touché, and 2,178 on the five measured NQ-1M queries; the first two corpora retain 100% of the float top-10 in the shortlist. We report the median of three long-lived-key queries on SciDocs and Touché and five on NQ-1M.

PANTHER [50]. We reuse the authors' artifact[3]. We compile the authors' random-client/random-server secure path, and quantize the common E5 vectors with a corpus-calibrated 9-bit affine map. SciDocs uses 8,074 bounded clusters, a 1,561-item stash, a maximum cluster size of 10, and 1,280 probes. Touché uses four bounded-cluster levels with 28,198, 10,280, 3,205, and 1,016 clusters, a 14,432-item stash, a maximum cluster size of 20, and 512/256/128/64 probes. These settings obtain 99.22% and 99.18% float top-10 agreement, respectively. NQ-1M uses 49,720 and 63,928 bounded clusters, a 37,125-item stash, a maximum cluster size of 20, and 792/396 probes, attaining 99.0% float top-10 agreement over 100 queries. The 768-dimensional, 20-point cluster layout expands its 113,648 PIR records to 15,480 32-bit elements each; the implementation maps 1,188 probes to 1,782 cuckoo bins and generates their SEAL replies in parallel while retaining the encoded database. This answer-stage memory peak exhausts the 256-GB host after database construction and the distance, argmin, and garbled-circuit stages complete, so we report OOM rather than an extrapolated latency. The completed SciDocs and Touché values are medians of three secure runs with measured traffic.

[2] https://github.com/myl7/p2rag
[3] https://github.com/AntCPLab/OpenPanther

Table 15: Complete E5 two-forward NDCG@10 under query-side DP. Randomized response, Gaussian, and RDP-vMF use $\delta = 10^{-6}$; pure-vMF gives the formal pure metric-DP guarantee.

Dataset	Mechanism	ε	$K = 200$	$K = 500$	$K = 1000$	$K = 2000$	$K = 3000$	Float
SciDocs	No DP	∞	.1884	.1881	.1877	.1872	.1870	.1870
	Randomized response	8	.0102	.0203	.0308	.0474	.0596	.1870
	Randomized response	16	.0111	.0235	.0380	.0536	.0700	.1870
	Randomized response	32	.0250	.0458	.0632	.0870	.1057	.1870
	Randomized response	64	.0566	.0766	.0992	.1262	.1407	.1870
	Gaussian	8	.1558	.1685	.1784	.1829	.1844	.1870
	Gaussian	16	.1857	.1868	.1873	.1880	.1876	.1870
	Gaussian	32	.1853	.1883	.1875	.1874	.1872	.1870
	Gaussian	64	.1885	.1871	.1875	.1872	.1873	.1870
	RDP-vMF	8	.1630	.1776	.1820	.1834	.1854	.1870
	RDP-vMF	16	.1833	.1850	.1872	.1869	.1866	.1870
	RDP-vMF	32	.1860	.1871	.1877	.1874	.1872	.1870
	RDP-vMF	64	.1874	.1877	.1880	.1874	.1870	.1870
	Pure-vMF	8	.0486	.0750	.0944	.1199	.1363	.1870
	Pure-vMF	16	.1210	.1447	.1599	.1732	.1804	.1870
	Pure-vMF	32	.1757	.1851	.1848	.1863	.1866	.1870
	Pure-vMF	64	.1846	.1854	.1857	.1869	.1868	.1870
NQ	No DP	∞	.5694	.5770	.5801	.5820	.5826	.5854
	Randomized response	8	.0003	.0005	.0008	.0015	.0043	.5854
	Randomized response	16	.0018	.0025	.0041	.0069	.0096	.5854
	Randomized response	32	.0044	.0094	.0155	.0264	.0326	.5854
	Randomized response	64	.0379	.0619	.0873	.1170	.1401	.5854
	Gaussian	8	.4440	.4845	.5072	.5271	.5389	.5854
	Gaussian	16	.5519	.5657	.5708	.5761	.5778	.5854
	Gaussian	32	.5663	.5750	.5783	.5811	.5820	.5854
	Gaussian	64	.5689	.5781	.5800	.5821	.5825	.5854
	RDP-vMF	8	.4645	.5015	.5257	.5434	.5522	.5854
	RDP-vMF	16	.5589	.5695	.5752	.5765	.5778	.5854
	RDP-vMF	32	.5681	.5761	.5796	.5822	.5827	.5854
	RDP-vMF	64	.5701	.5764	.5800	.5821	.5832	.5854
	Pure-vMF	8	.0254	.0414	.0592	.0852	.1043	.5854
	Pure-vMF	16	.2353	.2971	.3443	.3985	.4219	.5854
	Pure-vMF	32	.5061	.5328	.5465	.5571	.5617	.5854
	Pure-vMF	64	.5596	.5685	.5732	.5781	.5801	.5854
FEVER	No DP	∞	.8392	.8441	.8458	.8472	.8475	.8501
	Randomized response	8	.0007	.0008	.0013	.0022	.0024	.8501
	Randomized response	16	.0000	.0005	.0011	.0024	.0039	.8501
	Randomized response	32	.0039	.0060	.0103	.0162	.0226	.8501
	Randomized response	64	.0343	.0542	.0762	.1037	.1264	.8501
	Gaussian	8	.6029	.6675	.7083	.7410	.7561	.8501
	Gaussian	16	.8161	.8255	.8319	.8380	.8412	.8501
	Gaussian	32	.8389	.8436	.8465	.8478	.8481	.8501
	Gaussian	64	.8396	.8435	.8470	.8480	.8482	.8501
	RDP-vMF	8	.6484	.7037	.7406	.7690	.7867	.8501
	RDP-vMF	16	.8246	.8332	.8390	.8428	.8452	.8501
	RDP-vMF	32	.8385	.8427	.8460	.8485	.8488	.8501
	RDP-vMF	64	.8396	.8439	.8464	.8476	.8483	.8501
	Pure-vMF	8	.0187	.0331	.0503	.0713	.0890	.8501
	Pure-vMF	16	.2753	.3489	.4098	.4722	.5097	.8501
	Pure-vMF	32	.7293	.7665	.7908	.8087	.8171	.8501
	Pure-vMF	64	.8260	.8350	.8400	.8434	.8447	.8501

Table 16: Complete BGE two-forward NDCG@10 under query-side DP. Randomized response, Gaussian, and RDP-vMF use $\delta = 10^{-6}$; Pure-vMF gives the formal pure metric-DP guarantee.

Dataset	Mechanism	ε	$K = 200$	$K = 500$	$K = 1000$	$K = 2000$	$K = 3000$	Float
SciDocs	No DP	∞	.2233	.2225	.2227	.2229	.2228	.2228
	Randomized response	8	.0098	.0222	.0344	.0552	.0752	.2228
	Randomized response	16	.0169	.0335	.0486	.0672	.0861	.2228
	Randomized response	32	.0266	.0473	.0671	.0953	.1111	.2228
	Randomized response	64	.0635	.0984	.1295	.1534	.1731	.2228
	Gaussian	8	.2073	.2180	.2202	.2228	.2230	.2228
	Gaussian	16	.2221	.2231	.2229	.2230	.2229	.2228
	Gaussian	32	.2227	.2226	.2223	.2228	.2227	.2228
	Gaussian	64	.2227	.2224	.2230	.2229	.2228	.2228
	RDP-vMF	8	.2174	.2220	.2218	.2230	.2233	.2228
	RDP-vMF	16	.2222	.2227	.2232	.2231	.2229	.2228
	RDP-vMF	32	.2222	.2224	.2230	.2229	.2228	.2228
	RDP-vMF	64	.2228	.2227	.2230	.2227	.2228	.2228
	Pure-vMF	8	.0580	.0822	.1094	.1434	.1651	.2228
	Pure-vMF	16	.1530	.1784	.1947	.2097	.2169	.2228
	Pure-vMF	32	.2124	.2199	.2199	.2218	.2221	.2228
	Pure-vMF	64	.2218	.2228	.2233	.2232	.2229	.2228
NQ	No DP	∞	.5360	.5391	.5392	.5401	.5402	.5414
	Randomized response	8	.0000	.0003	.0010	.0035	.0039	.5414
	Randomized response	16	.0005	.0021	.0032	.0073	.0086	.5414
	Randomized response	32	.0038	.0071	.0118	.0230	.0304	.5414
	Randomized response	64	.0258	.0417	.0574	.0789	.0932	.5414
	Gaussian	8	.4765	.4988	.5107	.5217	.5247	.5414
	Gaussian	16	.5253	.5322	.5364	.5376	.5385	.5414
	Gaussian	32	.5341	.5369	.5379	.5396	.5400	.5414
	Gaussian	64	.5348	.5375	.5387	.5392	.5396	.5414
	RDP-vMF	8	.4911	.5093	.5174	.5251	.5278	.5414
	RDP-vMF	16	.5305	.5366	.5385	.5393	.5390	.5414
	RDP-vMF	32	.5342	.5380	.5387	.5398	.5400	.5414
	RDP-vMF	64	.5338	.5371	.5387	.5395	.5401	.5414
	Pure-vMF	8	.0238	.0345	.0503	.0690	.0837	.5414
	Pure-vMF	16	.2158	.2679	.3127	.3496	.3758	.5414
	Pure-vMF	32	.4746	.4969	.5088	.5206	.5252	.5414
	Pure-vMF	64	.5231	.5317	.5343	.5372	.5389	.5414
FEVER	No DP	∞	.8448	.8470	.8480	.8483	.8488	.8495
	Randomized response	8	.0001	.0006	.0011	.0017	.0021	.8495
	Randomized response	16	.0003	.0010	.0017	.0042	.0055	.8495
	Randomized response	32	.0025	.0055	.0088	.0160	.0198	.8495
	Randomized response	64	.0258	.0430	.0619	.0888	.1035	.8495
	Gaussian	8	.7609	.7884	.8046	.8188	.8236	.8495
	Gaussian	16	.8398	.8442	.8458	.8472	.8479	.8495
	Gaussian	32	.8450	.8463	.8476	.8483	.8488	.8495
	Gaussian	64	.8463	.8474	.8478	.8482	.8487	.8495
	RDP-vMF	8	.7942	.8129	.8237	.8336	.8370	.8495
	RDP-vMF	16	.8422	.8451	.8469	.8483	.8486	.8495
	RDP-vMF	32	.8453	.8473	.8479	.8483	.8485	.8495
	RDP-vMF	64	.8459	.8477	.8479	.8482	.8485	.8495
	Pure-vMF	8	.0184	.0307	.0437	.0647	.0806	.8495
	Pure-vMF	16	.2781	.3540	.4133	.4766	.5186	.8495
	Pure-vMF	32	.7579	.7875	.8027	.8177	.8248	.8495
	Pure-vMF	64	.8359	.8433	.8450	.8466	.8471	.8495

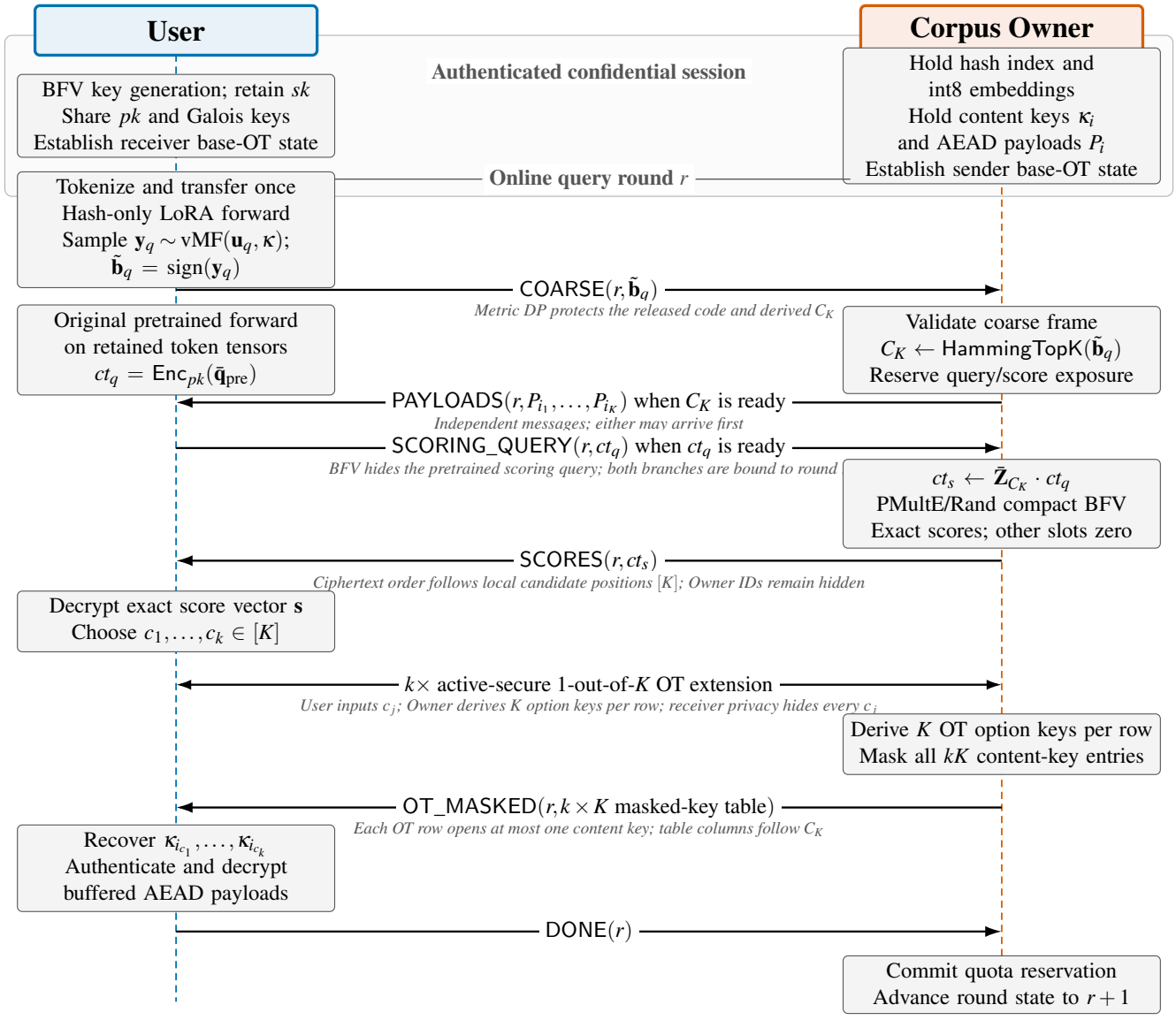

Figure 9: Two-party message sequence for one authenticated query round. The coarse frame starts Owner-side Hamming search while the User runs the pretrained forward and BFV encryption. Payload streaming begins when C_K is ready, the scoring query is sent when ct_q is ready, and either message may arrive first; active-secure OT later hides the selected positions and releases only their content keys.