DUETER: EFFICIENT DUAL-PATH ENCRYPTED SEARCH FOR RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Outsourcing retrieval-augmented generation (RAG) corpora and indexes to untrusted clouds creates a three-way conflict among retrieval fidelity, data confidentiality and scalability. Existing privacy-preserving RAG solutions either rely on single-path retrieval with inherently limited recall, or impose prohibitive cryptographic overhead. We propose DUETER, an efficient dual-path encrypted retrieval architecture that maps dense semantic similarity and compressed BM25 proxy onto a unified Euclidean ranking interface. Path-specific keyed indexes prune the search space to eliminate full-corpus scans, the cloud ranks encrypted candidates, and the client recomputes native scores before lightweight query-adaptive fusion. DUETER designs compartment conditional approximate distance-preserving encryption (CCADPE), which enables efficient ciphertext ranking without directly exposing the single global comparison frame of conventional DPE. Independent aliases further remove direct identifier equality across paths. Experiments on real-world datasets show that DUETER outperforms both single-path variants and improves recall over state-of-the-art (SOTA) baselines. At million-scale corpora, cloud-side query processing completes in roughly 1 second. These results establish a measured utility–efficiency–leakage architecture for scalable dual-path RAG. Our code is available on https://anonymous.4open.science/r/DuetER-0D0E.

1 INTRODUCTION

Despite prominent success in natural language processing, large language models (LLMs) suffer from outdated knowledge and inherent hallucinations (Ji et al. (2023); Achiam et al. (2023)). Retrieval-augmented generation (RAG) alleviates this limitation by grounding generation in evidence retrieved from an external corpus (Guu et al. (2020); Lewis et al. (2020); Izacard & Grave (2021); Borgeaud et al. (2022)), which has been widely used for search assistants, question-answering services, and enterprise knowledge systems.

Effective RAG systems typically combine two complementary retrieval paths. Semantic retrieval matches dense query and document embeddings, capturing paraphrases and conceptual similarity (Karpukhin et al. (2020)). Lexical retrieval, commonly based on BM25 (Robertson et al. (2009)), preserves exact evidence such as identifiers, names, and technical terms. Extensive studies have demonstrated that their combination yields superior retrieval effectiveness (Kuzi et al. (2020); Thakur et al. (2021)). However, practical RAG deployments commonly outsource document indexing and retrieval to untrusted clouds, raising severe privacy risks for private and sensitive data (Huang et al. (2023); Zeng et al. (2024)). In particular, healthcare RAG systems handling confidential clinical records may expose user privacy and sensitive database information through plaintext query and retrieval processes. This necessitates the migration of mature dual-path plaintext retrieval frameworks to secure encrypted scenarios for privacy-preserving RAG deployment.

Existing privacy-preserving RAG primarily optimize for privacy protection or retrieval efficiency. Differential privacy perturbs queries or embeddings (Dwork & Roth (2014); Grislain (2025); Yao & Li (2025)), but such perturbations are difficult to reconcile with exact lexical matching. RemoteRAG (Cheng et al. (2025)) and PRAG (Li et al. (2026)) use homomorphic encryption (HE) for private distance evaluation or ranking, whereas Pisces (Liang et al. (2026)) supports both semantic and lexical retrieval through cryptographic primitives, including oblivious pseudorandom functions (OPRF), oblivious key-value stores (OKVS), and secure multi-party computation (MPC) (Freedman

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al. (2005); Garimella et al. (2021); Bienstock et al. (2024)). These protections incur substantial computation or communication overhead. In contrast, ppRAG (Ye et al. (2025)) applies conditional approximate distance-preserving encryption (CADPE) (Fuchsbauer et al. (2022)) to dense embeddings, enabling near-plaintext cloud-side ranking, but supports only single-path semantic retrieval. Motivated by the above gaps, this work investigates a core problem: *Can encrypted dual-path retrieval preserve efficient cloud-side ranking while explicitly quantifying and constraining information leakage?*

We answer this question with DUETER, an efficient encrypted retrieval architecture for RAG that maps both paths to a common Euclidean ranking interface. For semantic retrieval, ℓ_2 normalization converts cosine similarity into Euclidean-distance ranking. For lexical retrieval, DUETER uses signed feature hashing (Weinberger et al. (2009)) and a maximum-inner-product-to-Euclidean (MIPS-to-L_2) reduction (Shrivastava & Li (2014)) to construct a compact BM25 proxy for cloud refinement, while retaining AEAD-protected term statistics for exact candidate BM25 recomputation. Each path uses our proposed CCADPE as an efficiency-oriented ranking primitive, allowing the cloud to rank encrypted candidates without HE or MPC, while partitioning the the released geometry into independently keyed compartment-local views. To avoid full-corpus scans, the semantic path adopts mean-residual spherical IVF, whereas the lexical path builds an HMAC-keyed inverted index with a cumulative-posting planner. After cloud-side ranking, the client locally decrypts and reranks the results, resolves randomized join capsules, and performs lightweight query-adaptive fusion. Independent keys and aliases separate the two server-visible retrieval views, while join logic and fusion parameters remain client-side. Our contributions are as follows:

- **Efficient encrypted dual-path retrieval with CCADPE.** We map both normalized semantic similarity and a compressed BM25 proxy to Euclidean distance, while preserving exact candidate BM25 scoring at the client. Our novel CCADPE enables efficient compartment-local cloud ranking, and supports independent representations, keys, indexes, and leakage profiles for two paths.
- **Path-specific candidate generation.** We develop mean-residual spherical IVF for anisotropic dense embeddings and an HMAC-keyed lexical index with cumulative-posting planning. We further characterize the approximation introduced by different stages.
- **Explicit leakage analysis.** We formalize the complete server-visible leakage and prove that the cryptographic wrapper adds no information beyond the declared view. Independent aliases remove direct identifier equality, while experiments quantify how compartmentalization degrades the evaluated geometry-based attacks.
- **Complete-corpus evaluation.** Experiments on real-world datasets show that DUETER retains over 98% of full-scan semantic nDCG@10 by scanning 13% of vector corpus. Its cumulative-posting planner improves relevance-candidate coverage by approximately 25%. Compared with SOTA baselines, DUETER achieves higher retrieval accuracy with millisecond-scale latency.

2 RELATED WORK

RAG with Hybrid Retrieval Practical RAG deployments rely on hybrid retrieval that combines complementary semantic dense and lexical sparse signals, where dense retrieval captures paraphrased meaning and lexical methods excel at exact-term, entity, and keyword matching (Karpukhin et al. (2020); Formal et al. (2021); Thakur et al. (2021)). Plaintext large-scale hybrid retrieval systems adopt coarse-to-fine indexing such as inverted files, LSH and product quantization for efficient candidate pruning (Datar et al. (2004); Jégou et al. (2011); Malkov & Yashunin (2020); Johnson et al. (2021)). While plaintext hybrid retrieval is well-studied, securely composing these two heterogeneous retrieval modalities under outsourced untrusted cloud settings remains an open challenge.

RAG with Privacy Protection Most privacy-preserving RAG leverages heavy cryptographic primitives including MPC, PSI, and HE (Cheng et al. (2025); Liang et al. (2026); Li et al. (2026)). DPE offers an alternative efficiency for cheap server-side ranking. ppRAG adapts this primitive for private dense RAG (Ye et al. (2025)), yet it only supports single-path semantic retrieval. Though Pisces (Liang et al. (2026)) realizes hybrid RAG, it does not investigate cross-path identifier-linkage leakage risks arising when merging results from heterogeneous retrieval paths.

3 PRELIMINARIES

3.1 COMPARTMENT CONDITIONAL APPROXIMATE DISTANCE-PRESERVING ENCRYPTION

CADPE enables efficient cloud-side coarse ranking over ciphertext vectors. However, it exposes one server-visible coordinate system and its induced metric relation. Therefore, we design compartment CADPE (CCADPE), which exposes a collection of independently keyed compartment-local coordinate views. The protocol directly evaluates approximate distances only between coordinates assigned to the same compartment and constructs no protocol-level cross-compartment distance interface. This removes an immediately shared global frame. Let $s_p > 0$ denote the CADPE scaling factor, $\beta_p > 0$ denote path-specific distance-tolerance parameter, $\varepsilon_{p,b}^{\mathrm{proj}}$ denote the within-compartment projection distortion, and $\delta_{p,b}^{\mathrm{cc}}$ denote its total failure probability. CCADPE provides the following two within-compartment properties.

Theorem 1 (Conditional comparison preservation) *Fix unit-norm vectors v_q^p, v_i^p, v_j^p in the same compartment b, and define $\Gamma_{p,b}^q = \frac{\beta_p}{2} + \varepsilon_{p,b}^{\mathrm{proj}} \left(2 + \frac{\beta_p}{2}\right)$. For a fixed, key-oblivious workload, except with probability $\delta_{p,b}^{\mathrm{cc}}$, if $\|v_q^p - v_i^p\|_2 < \|v_q^p - v_j^p\|_2 - 2\Gamma_{p,b}^q$, then $\|c_{q,b}^p - c_{i,b}^p\|_2 < \|c_{q,b}^p - c_{j,b}^p\|_2$.*

Theorem 2 (Stored-geometry distortion) *For two document pairs $(i,j),(k,\ell)$ in compartment b. For each $(u,v) \in \{(i,j),(k,\ell)\}$, define $d_{uv} = \|v_u^p - v_v^p\|_2$ and $d'_{uv,b} = \|c_{u,b}^p - c_{v,b}^p\|_2$, and let $\Gamma_{p,b}^{db} = \frac{3\beta_p}{4} + \varepsilon_{p,b}^{\mathrm{proj}} \left(2 + \frac{3\beta_p}{4}\right)$. Except with probability $\delta_{p,b}^{\mathrm{cc}}$, the following bound holds simultaneously for both document pairs $\left|\frac{d'_{uv,b}}{s_p} - d_{uv}\right| \le \Gamma_{p,b}^{db}$. Consequently, if $d_{ij} < d_{k\ell} - 2\Gamma_{p,b}^{db}$, then $d'_{ij,b} < d'_{k\ell,b}$.*

The details of CADPE and CCADPE can be found in Appendix A.1 and A.2, respectively. The proof of these properties is in Appendix B.1 and B.2.

3.2 DUAL-PATH REPRESENTATIONS

The dual-path representation unifies two heterogeneous retrieval signals—dense semantic similarity and sparse lexical BM25 proxy—into a single Euclidean distance ranking interface, enabling reuse of the CCADPE for cloud-side candidate refinement. For the semantic path, encoder outputs undergo L_2 normalization, exploiting the equivalence between cosine similarity and squared Euclidean distance for unit vectors to convert similarity-based ranking into distance ordering. For the lexical path, BM25 is formulated as a sparse inner product, compressed by signed feature hashing, and projected onto the unit sphere via a MIPS-to-L_2 transform. This unified abstraction preserves semantic cosine ordering and provides a compact approximation of native BM25 ordering, while enabling consistent CCADPE deployment across both paths. Full derivations are deferred to Appendix A.3.

4 SYSTEM AND THREAT MODEL

4.1 SYSTEM MODEL

The *owner* possesses the database and encrypts indexes. The authorized *client* holds retrieval keys and submits queries. The *cloud* stores encrypted documents and two path indexes, executes bucket lookup and ciphertext ranking, and returns candidates. The *cloud* is honest-but-curious. It faithfully executes the protocol but records all generated transcripts. The *owner* and the *client* are trusted.

4.2 SECURITY OBJECTIVES

We define two objectives centered on the retrieval phase of outsourced RAG

- **Security guarantees**. Confidentiality and authenticity for documents and exact returned features, pseudorandom path-local identifiers, and no explicit server-visible semantic-lexical join relation.
- **Efficiency**. Lightweight one-round server-side candidate pruning with client-local rank fusion, avoiding heavy cryptographic overhead from HE and MPC.

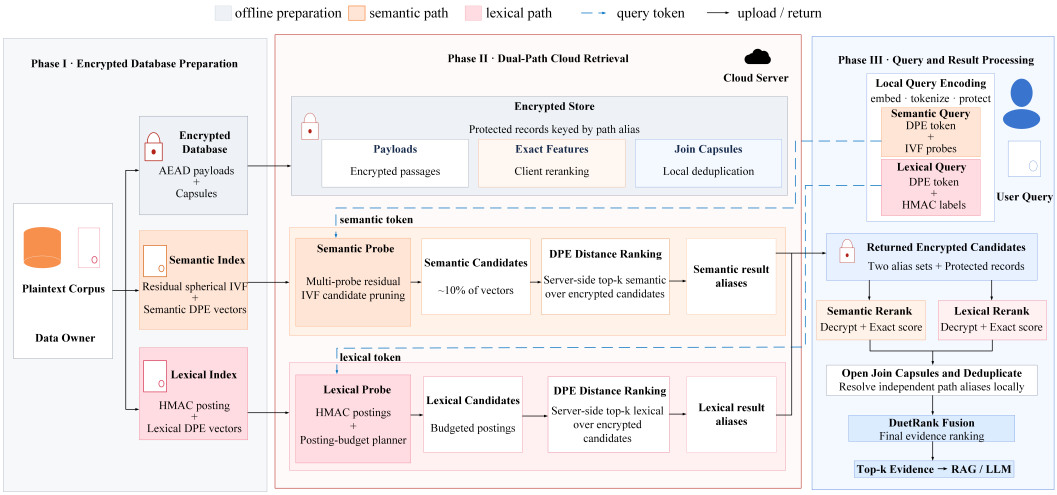

Figure 1: Framework of DUETER.

4.3 LEAKAGE PROFILE

We decompose leakage into CCADPE-inherent and protocol-induced components. For path $p \in \{s, l\}$, let $\mathcal{B}_p$ denote its compartment set and $\mathcal{H}_{p,b}$ the local handle space of compartment $b \in \mathcal{B}_p$. After T adaptive queries, the concrete coordinate view is

$$\mathsf{G}_p^{\mathsf{cmp}}(\mathcal{D}, \mathbf{q}) = \Big(\big\{ \big(\tau_{p,b}, \mathcal{H}_{p,b}, \{(h, c^p_{h,b}, \nu^p_{h,b}) : h \in \mathcal{H}_{p,b}\}\big) : b \in \mathcal{B}_p \big\}, \\ \big\{ \big(\mathsf{QCmp}_{p,t}, \{c^p_{q_t,b} : b \in \mathsf{QCmp}_{p,t}\}\big) : 1 \le t \le T \big\} \Big),$$

where b is an opaque but equality-visible compartment label, $c^p_{h,b} \in \mathbb{R}^{r_c}$ is the uploaded local coordinate of handle h, and $\nu^p_{h,b} \in \{0,1\}^\lambda$ denotes any public per-copy DPE randomizer or nonce, with $\nu^p_{h,b} = \perp$ when no separate value is emitted. The value $c^p_{q_t,b} \in \mathbb{R}^{r_c}$ is the local coordinate of query q_t, and $\mathsf{QCmp}_{p,t}$ contains the compartments for which query coordinates are transmitted. The view $\mathsf{G}_p^{\mathsf{cmp}}$ reveals complete numerical coordinates, within-compartment distances, and local neighborhood orders. For the surrounding protocol, setup reveals

$$\mathcal{L}_p^{\mathsf{idx,cmp}} = \big(N, d_p, r_c, |\mathcal{B}_p|, \mathsf{pp}_p, \mathsf{Len}_p, \mathsf{CmpInc}_p, \mathsf{PostInc}_p, \mathsf{PostVal}_p, \mathsf{BktSize}_p, \mathsf{Upd}_p\big).$$

pp_p denotes public parameters including index dimensions, compartment counts, probe counts, local return depths, quantizer width, and padding bounds. CmpInc_p records opaque compartment-handle incidence, while $\mathsf{PostInc}_p$ records keyed posting-handle incidence. $\mathsf{PostVal}_l$ contains quantized lexical posting weights, whereas $\mathsf{PostVal}_s = \perp$. $\mathsf{BktSize}_p$ exposes compartment and posting sizes, and Upd_p records the visible update pattern. For adaptive query t, operational leakage is

$$\mathcal{L}_{p,t}^{\mathsf{op,cmp}} = \big(\mathsf{SearchEq}_{p,t}, \mathsf{QCmp}_{p,t}, \mathsf{Acc}_{p,t}, \mathsf{Cand}_{p,t}, \mathsf{DistTr}_{p,t}, \mathsf{Rsp}_{p,t}, \mathsf{Time}_{p,t}, \mathsf{Fetch}_{p,t}\big).$$

SearchEq tracks equality among repeated keyed labels, QCmp records transmitted compartment labels, Acc denotes the accessed compartments and postings, Cand are the per-compartment top-t_p outputs, DistTr records the within-compartment distances and orders used for refinement, Rsp describes padded response shapes, Time records query ordering and cross-path synchronization, and Fetch records the encrypted evidence locations ultimately requested after client-side fusion. The distance trace is retained explicitly for transparency, despite being derivable from $\mathsf{G}_p^{\mathsf{cmp}}$. We define the overall concrete leakage adopted in our simulation theorem as

$$\mathcal{L}^{\mathsf{conc}} = \big(\mathsf{G}_s^{\mathsf{cmp}}, \mathsf{G}_l^{\mathsf{cmp}}, \{\mathcal{L}_p^{\mathsf{idx,cmp}}\}_{p\in\{s,l\}}, \{\mathcal{L}_{p,t}^{\mathsf{op,cmp}}\}_{p\in\{s,l\},\, t\le T}\big).$$

Definition 1 (Cross-path leakage) *Let* $\mathsf{View}_s^{\mathsf{cmp}}$ *and* $\mathsf{View}_l^{\mathsf{cmp}}$ *denote the two complete Compartment-DPE path transcripts. Cross-path leakage is information efficiently derivable from* $(\mathsf{View}_s^{\mathsf{cmp}}, \mathsf{View}_l^{\mathsf{cmp}})$ *that is not efficiently derivable from either view alone under its declared leakage function.*

5 PROPOSED METHOD

5.1 ENCRYPTED INDEX CONSTRUCTION

Mean-Squared Residual Spherical IVF Index for Semantic Path. Granite embeddings exhibit an anisotropic distribution: a dominant shared global direction degrades standard LSH-based inverted-file lookup and can lead to nearly full-scan search overhead. To mitigate this issue, we center residuals on the corpus mean $\mu = \frac{1}{N}\sum_i v_i^s$ and unit-normalize as $z_i = \frac{v_i^s - \mu}{\|v_i^s - \mu\|_2}$. It then learns K unit centroids $\{m_1, \ldots, m_K\}$ by spherical k-means over residual vectors z_i and assigns each document to its r centroids with largest inner products $m_j^\top z_i$. Each bucket j is stored under an opaque label $\tau_j^s = \mathsf{HMAC}_{K_s^{\text{bucket}}}(s\|\texttt{ivf}\|j)$. At query time, the client retains μ and the precomputed centroids, residualizes the query embedding, and submits labels corresponding to its n_{probe} nearest centroids. For both paths, compartment records are constructed according to Algorithm 4. For document i stored in compartment b of path $p \in \{s, l\}$, the resulting record is $\mathsf{Rec}_{i,b}^p = \left(a_{i,b}^p, c_{i,b}^p, V_{i,b}^p, J_{i,b}^p\right)$, where $a_{i,b}^p$ is a compartment-local alias, $c_{i,b}^p$ is the cloud-ranking, $V_{i,b}^p$ encrypts the path-native reranking payload, and $J_{i,b}^p$ is the rerandomized join capsule. The semantic payload contains the normalized embedding, and the lexical payload contains keyed token IDs with multiplicities and document length. Each semantic cell stores a separate record for every assigned document copy. The cloud returns a fixed t_s entries per probed cell without comparing coordinates across cells; the client deduplicates their union and reranks it using the original normalized embeddings. We call this construction *keyed residual spherical IVF*.

Collision-free Keyed BM25 Tail Index for Lexical Path. Generic Euclidean LSH performs poorly after the lexical MIPS-to-L_2 transformation, since the residual coordinate contributes no relevance evidence. Let the transformed document and query be $P(\hat{x}_i) = [u_i; r_i]$ and $Q(\hat{y}_q) = [\bar{y}_q; 0]$, respectively, where $u_i = \hat{x}_i/M$, $\bar{y}_q = \hat{y}_q/\|\hat{y}_q\|_2$, and $r_i = \sqrt{1 - \|u_i\|_2^2}$. We construct a collision-free keyed tail index over original sparse BM25 coordinates, while retaining compressed transformed vectors for encrypted L_2-based candidate refinement. For each term t, the owner derives the opaque posting label $\tau_t^l = \mathsf{HMAC}_{K_l^{\text{bucket}}}(l\|t)$. Document i is assigned to $b_l(i) = \mathsf{HMAC}_{K_l^{\text{part}}}(i) \bmod K_l^{\text{cmp}}$, and its posting for term t stores $\left(a_{i,b_l(i)}^l, b_l(i), \mathsf{Quant}(x_{i,t})\right)$. Here, $\mathsf{Quant}(x_{i,t})$ is the b_q-bit uniformly quantized BM25 coordinate. A corpus-wide bound U determines its step size Δ_q, while its approximation error is analyzed in Appendix B.3. In contrast, $V_{i,b_l(i)}^l$ contains an AEAD-protected lexical payload W_i and is available only to the client for final BM25 reranking. Fixed-count term probing is unstable under highly variable posting sizes. For visible posting size f_t, corpus size N, and global posting budget B_l, terms are sorted by priority $\pi(t) = \frac{\log(1+N/f_t)}{f_t^{0.5}}$, and greedily admitted until $\sum_{t\in S_q} f_t \le B_l$. For the posting union, the cloud forms the quantized proxy $\widetilde{S}_l(i) = \sum_{t\in S_q} \mathsf{Dequant}(\mathsf{Quant}(x_{i,t}))$ and retains its top C_l documents. The survivors are grouped by $b_l(i)$ and ranked against the corresponding compartment-local query coordinate. The cloud emits a fixed local top-t_l per active partition, after which the client unions the results, decrypts $V_{i,b_l(i)}^l$, and computes exact BM25 within the returned set.

Path-local Aliases and Join Capsules. For each path and compartment, the owner derives $a_{i,b}^p = \mathsf{PRP}_{K_{p,b}^{alias}}(i\|b)$. Semantic replicas in different IVF cells therefore receive unrelated aliases, while a lexical record receives an alias local to its keyed partition. Path-native reranking payloads and join capsules are independently rerandomized for every stored copy with fresh nonces. It further produces the join capsules $J_{i,b}^p = \mathsf{AEAD}_{K_J}(i\|\mathsf{loc}_i\|\mathsf{version}_i; \mathsf{nonce}_{i,b}^p)$. Consequently, identical documents yield distinct aliases and disjoint join-capsule ciphertexts across paths and compartments. Only the client can decrypt join capsules and perform cross-path result deduplication. Document contents and metadata are protected via AEAD with HKDF-derived per-document keys $K_i = \mathsf{HKDF}(K_{doc}, i\|\mathsf{version}_i)$, producing ciphertexts $X_i = \mathsf{AEAD}_{K_i}(d_i; \mathsf{metadata}_i)$.

5.2 QUERY PREPARATION

The client builds the semantic vector $v_q^s = \mathsf{Normalize}(E(q))$ and lexical proxy $v_q^l = Q(Sy_q)$, using index-consistent feature mappings. It residualizes v_q^s to select the nearest semantic IVF cells and

Algorithm 1: Encrypted dual-path index construction

Input: Document collection $\mathcal{D}$, semantic encoder E, lexical statistics
Output: Cloud indexes $\mathcal{I}_s, \mathcal{I}_l$ and encrypted contents

foreach $d_i \in \mathcal{D}$ **do**
- Compute the normalized semantic vector v_i^s, BM25 vector x_i, and exact lexical payload W_i;
- Transform x_i into the lexical Euclidean representation v_i^l;
- Residualize v_i^s, assign nearest semantic cells, and add one compartment record to each assigned cell;
- Assign v_i^l to its keyed lexical partition and insert its compartment record with $V_i^l = AEAD(W_i)$;
- For every $t \in \mathrm{supp}(x_i)$, insert $(a_{i,b_l}^l, b_l, \mathrm{Quant}(x_{i,t}))$ into posting list τ_t^l;
- Encrypt d_i and its metadata into the document store;

generates the corresponding HMAC labels, while query terms are mapped to keyed lexical posting labels. For each compartment b of path p, it produces $c_{q,b}^p = \mathsf{CmpQEnc}_{p,b}(v_q^p)$. For the lexical path, it contains a local coordinate for each partition. Optional lexical cover probes do not alter v_q^l. Each label list is padded to fixed length T_p with pseudorandom dummy labels, normalizing only the wire-level request length.

5.3 Cloud Search

The cloud resolves semantic probes and lexical postings, applies the quantized lexical cap, and groups candidates by path and compartment. Within each accessed pair (p, b), it ranks candidates by $\delta_{i,b}^{p\prime} = \|c_{q,b}^p - c_{i,b}^p\|_2$ and returns a fixed local top-t_p entries. No cross-compartment order is constructed. The padded outputs contain local aliases, encrypted path-native reranking payloads, and rerandomized join capsules. The two paths remain separate until client-side fusion.

5.4 Client Reranking and Fusion

The client decrypts the results and evaluates exact semantic scores $S_s(i) = \cos(v_q^s, v_i^s)$. For the lexical path, it reconstructs the native BM25 weights $x_{i,t}$ from the protected payload W_i and corpus statistics, and computes the exact candidate score $S_l(i) = y_q^\top x_i$. Join capsules J_i^p are decrypted for deduplication and client-side fusion. A standard RRF score is computed as $F(i) = \frac{1}{c+r_s(i)} + \frac{1}{c+r_l(i)}$, However, fixed RRF provides a reproducible fallback when target-domain relevance labels are unavailable, motivating a client-only lightweight gradient-boosted ranker *DuetRank*. It consumes rank transforms, sores, query statistics and posting costs to produce a query-adaptive fused ranking.

6 Security Analysis

We analyze security under the honest-but-curious cloud adversary defined in Section 4. DuetER provides cryptographic confidentiality for document payloads, exact client-side representations, and the wrapper identifiers. In contrast, CCADPE is an efficiency primitive that provides structured and empirically measured leakage reduction rather than embedding-coordinate confidentiality.

Definition 2 (Adaptive $\mathcal{L}$-simulation security) *Protocol Π is adaptively $\mathcal{L}$-simulation secure if, for every PPT $\mathcal{A}$, there exists a PPT $\mathcal{S}$ such that the real view $\mathsf{Real}_{\Pi,\mathcal{A}}$ and ideal view $\mathsf{Ideal}_{\mathcal{S},\mathcal{A}}$ are computationally indistinguishable.*

Theorem 3 (Leakage-relative wrapper simulation) *Let Π^{conc} denote* DuetER *instantiated with* CCADPE. *If HKDF/HMAC are secure PRFs, the two path aliases are independent secure PRPs, and AEAD satisfies multi-message IND-CPA security, then Π^{conc} is adaptively $\mathcal{L}^{\mathsf{conc}}$-simulation secure for the honest-but-curious adversary.*

Corollary 1 (Identifier-subview separation) *Consider the server-view restricted to path aliases, bucket labels, and join capsules, excluding DPE coordinates, access traces, and unpadded document fetch outputs. Under the assumptions of Theorem 3, no adversary can distinguish two executions with identical per-path leakage but distinct hidden semantic-lexical identifier mappings.*

Algorithm 2: Dual-path retrieval

Input: Query sets $\mathcal{Q}$, path keys sk_s, sk_l, result count k
Output: Decrypted evidence passages
foreach $q \in \mathcal{Q}$ **do**
  Encode q into normalized semantic and lexical query vectors;
  Generate semantic cell probes and budgeted lexical term probes;
  Construct a compartment-local query coordinate for every semantic probe and lexical partition;
  Cloud resolves the probes, generates path candidates, and groups them by compartment;
  Within each compartment, rank candidates by encrypted distance and return a fixed local top-t_p;
  Client decrypts the returned payloads and recomputes exact cosine and candidate-local BM25 scores;
  Client opens join capsules, deduplicates candidates, and applies DuetRank or RRF fusion;
  Fetch, authenticate, and decrypt the top-k evidence passages;
return the authorized evidence context;

This second theorem is the identifier-only projection of Theorem 3, not a global unlinkability claim in the presence of geometric side information. Appendix B.4 gives the concrete-view hybrid proof.

7 Evaluation Methodology

7.1 Research Questions

- **RQ1—Utility and scalability:** Can DuetER maintain retrieval quality and achieve effective dual-path fusion with constrained candidate computation on million-document complete corpora?
- **RQ2—Efficiency and design-space positioning:** What end-to-end overheads does DuetER incur at full-corpus scale, and how does it compare against SOTA private-RAG baselines?
- **RQ3—Empirical leakage validation:** Does the concrete server view exhibit the leakage predicted by our model, and which leakage channels are mitigated by path separation?

7.2 Experimental Settings

Datasets All retrieval and efficiency evaluations are conducted on the MS MARCO, Natural Questions (NQ), and HotpotQA. We use the 384-dimensional granite-embedding-small-english-r2 encoder (Awasthy et al. (2025)). Auxiliary corpora (20 Newsgroups, FiQA, TriviaQA) are used for external baseline, as detailed in Appendix C.

Baselines Plaintext baselines include BM25, dense cosine retrieval, and BM25-dense reciprocal rank fusion. Encrypted controls comprise semantic-only, lexical-only, and full dual-path DPE retrieval. Pisces (Liang et al. (2026)) and PRAG (Li et al. (2026)) serve as cost references.

Environments Complete-corpus experiments run on an Intel(R) Core(TM) i9-14900HX CPU paired with NVIDIA GeForce RTX 4060 Laptop GPU, whereas the comparison is performed in an Ubuntu 20.04.5 VMware guest with 16 vCPUs and 32 Gib RAM.

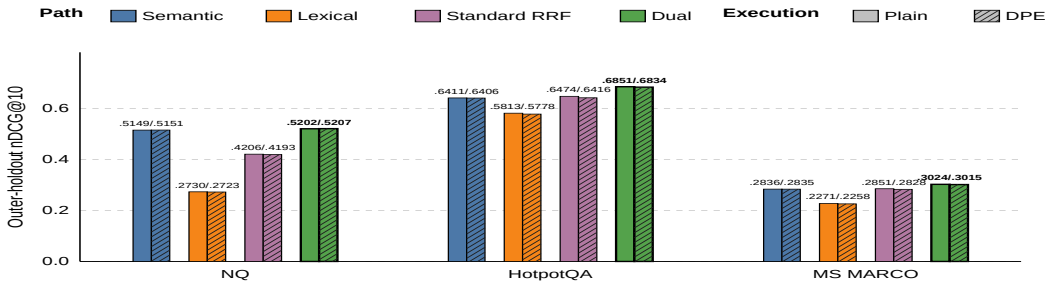

Figure 2: Complete-corpus path utility on outer holdouts with DuetRank.

7.3 RQ1: Complete-Corpus Utility and Scalability

Dual-path Utility We calculate nDCG@10 separately for the CCADPE semantic and lexical path and the plaintext semantic and lexical path. Figure 2 shows that DuetER exceeds both single paths and the CCADPE result retains at least 99.70% of its paired plaintext nDCG@10 on every corpus.

Candidate Selectivity and Ranking Retention Table 1 places the two candidate mechanisms under a common complete-corpus view. On the semantic path, the index scans 13% of embedding vectors while preserving over 98% of the exact-dense nDCG@10. On the lexical path, the cumulative-posting planner improves relevant-candidate coverage by up to 0.3242. Within retrieved candidate pools, lexical DPE incurs negligible ranking distortion, shifting nDCG@10 by no more than 0.00340. Hyperparameter sensitivity ablations are deferred to Appendix D.

Table 1: Complete-corpus candidate selectivity and retrieval fidelity.

	Semantic residual-IVF						Lexical posting planner		
Data	Scan	RelCov	R/C nDCG@10	Ret. N	R/C R@100	Ret. R	2-term	Budget	$\lvert\Delta\rvert$
MS MARCO	12.99%	.8233	.2881/.2880	99.96%	.7819/.7620	97.45%	.4873	.8115	.00100
NQ	13.19%	.7296	.5212/.5181	99.42%	.9426/.9067	96.19%	.5827	.7975	.00121
HotpotQA	12.81%	.6900	.6468/.6391	98.81%	.7932/.7619	96.06%	.5182	.7677	.00340

Note. R/C denotes the path-specific reference and CCADPE result. Ret. N/R are nDCG@10 and Recall@100. RelCov is judged-relevant candidate coverage. Lexical 2-term/Budget denotes coverage under two-rarest-term and cumulative-posting probing, respectively. $|\Delta|$ is the absolute difference in final lexical nDCG@10 between the plaintext-proxy and CCADPE candidate pipelines.

Controlled Accuracy Comparison We compare DuetER against Pisces. Whereas Pisces constructs outputs via the unordered union of per-path top-10 candidates, DuetER performs learned fusion over larger candidate pools. We operate on the unordered output policy $U_{10} = S@10 \cup L@10$ to ensure fair comparison. Table 2 shows consistent recall gains for DuetER, which remain valid even when exactly matching Pisces' per-query output size. Besides, the efficiency difference is substantially larger and stable across queries. Overall, DuetER slightly improves union recall over Pisces, while reducing online latency by $192.7\times$ and communication by $637.4\times$, demonstrating comparable retrieval utility with substantially lower computational and communication overhead.

Table 2: Controlled comparison over 300 sampled FiQA test queries.

System	$R_s@10$	$R_l@10$	Mean $\lvert U_{10}\rvert$	$R@U_{10}$	Online time	Communication
DuetER	.4858	.3163	17.92	.5279	0.0837 s	2.925 MiB
Pisces	.4762	.3032	17.94	.5189	16.124 s	1864.34 MiB

Note. The unordered output policy is $U_{10} = S@10 \cup L@10$. $R_s/R_l@10$ is semantic and lexical recall@10. Mean $|U_{10}|$ is the number of mean candidates. $R@U_{10}$ is recall at mean candidates.

Summary The results show that DuetER retains strong plaintext fidelity on large corpora. Our indexing strategy delivers competitive performance with partial embedding scans, while lexical DPE maintains ranking quality over retrieved candidates. Furthermore, DuetER yields consistent improvements against Pisces under matched evidence budgets.

7.4 RQ2: Efficiency and Design-Space Positioning

Complete-Corpus System Cost We first characterize the end-to-end resource of DuetER. Our overall cost stems from three components: index-driven candidate retrieval, DPE-based ciphertext refinement, and lightweight client-side reranking and fusion. Table 3 summarizes storage, candidate volume, and per-query latency across workloads. Index-level pruning tightly bounds the number of candidates passed to expensive ciphertext-distance evaluation, keeping client-side overhead modest.

Comparison against Baselines We situate DuetER against Pisces and PRAG, two privacy-preserving RAG built on heavy cryptographic primitives. As shown in Figure 3, Pisces and PRAG incur substantial runtime and communication burdens. DuetER achieves multi-million-document retrieval with compact per-query network traffic and millisecond-scale server-side processing.

Table 3: Complete-corpus retrieval cost.

Data	Index (GB)	S/L Mean cand.	Semantic DPE (ms)	Semantic rerank (ms)	Lookup (ms)	Lexical DPE (ms)	Exact BM25 (ms)	DuetRank (ms)
NQ	3.63	1907/958	169.05	18.51	65.66	0.53	0.26	2.11
HotpotQA	6.62	1927/954	292.57	23.60	43.61	0.71	0.25	2.43
MS MARCO	11.29	3842/925	924.94	58.88	85.08	0.48	0.26	2.56

Note. S/L Mean cand: mean candidates of semantic and lexical path. Semantic DPE: semantic candidates ranking. Semantic rerank: semantic reranking time. Lookup: retrieval time. Lexical DPE: lexical candidates ranking. Exact BM25: BM25 reranking time. DuetRank: fusion time.

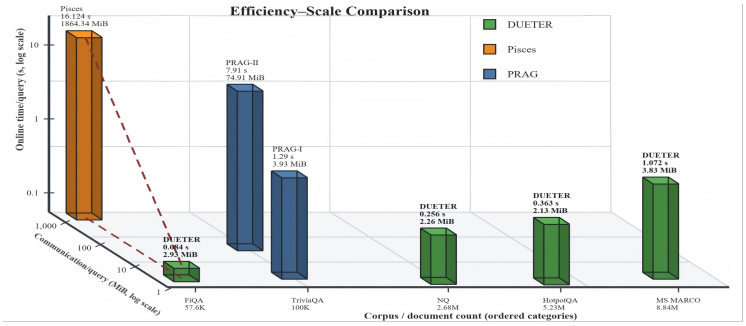

Figure 3: Online cost against other privacy-preserving RAG systems.

Summary The above results show that our indexing and DPE-based pipeline yields lightweight server-side computation and restrained communication. These observations confirm that DUETER occupies a distinct practical operating point compared with heavy cryptography-centric alternatives.

7.5 RQ3: Leakage Validation

Leakage across Different Views Table 4 compares leakage from plaintext embeddings, global DPE, and CCADPE. Global DPE retains a nearly isometric global metric space, leaving topic inference and geometry-based attacks largely unimpaired. In contrast, CCADPE reduces the evaluated attacks that rely on an immediately shared global frame, while independent aliases reduce direct equality-based matching between path identifiers.

Table 4: Leakage attacks from unprotected views to CCADPE.

Attack	Metric	Plaintext	Global DPE	CCADPE
Topic inference	20-class accuracy	.6473	.6440	.5933
Neighborhood reconstruction	Top-10 neighbor recall	1.0000	.9630	.4172
Known-record alignment	Top-1 accuracy with 4 anchors	1.0000	.9993	.0120
Known-query recovery	Normalized gain over mean predictor	1.0000	.9187	.0015
Direct alias linkage	Top-1 linkage accuracy	1.0000	.0304	.0210

Note. Higher values indicate greater adversarial success. The normalized recovery gain is $(c_{\mathrm{attack}} - c_{\mathrm{base}})/(1 - c_{\mathrm{base}})$ under 384 known queries.

Summary The results confirm that compartment-local keying substantially reduces the success of the evaluated global-frame attacks, and that independent aliases suppress direct identifier matching, while leaving the declared local-coordinate and operational leakage intact.

8 Conclusion

We presented DUETER, an efficient dual-path retrieval architecture for outsourced RAG. DUETER maps semantic similarity and lexical proxy to a common Euclidean interface, combines path-specific keyed indexes with independent aliases, and performs cross-path fusion at the client. CCADPE enables efficient cloud-side ranking while limiting global geometric exposure. Across complete corpora databases, DUETER scans only 13% of semantic vectors while retaining over 98% of exact-dense nDCG@10. Furthermore, extensive experiments demonstrate that DUETER achieves significant performance advantages over SOTA solutions in retrieval accuracy and communication and computation overhead. Future work will investigate dynamic updates.

AI USE STATEMENT

In this work, we used generative AI tools solely for grammatical correction, wording polishing, and sentence fluency optimization of the manuscript text. We have not used generative AI tools for research idea formulation, literature review, experimental design, data generation, data analysis, result interpretation, code development, or any other core research tasks. All other disclosure items specified in the ICLR 2027 AI Policy for Authors are not applicable to this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. 2023. URL https://arxiv.org/abs/2303.08774.

Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, Vignesh P, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, Ken Barker, Martin Franz, Madison Lee, Todd Ward, Salim Roukos, David Cox, Luis Lastras, Jaydeep Sen, and Radu Florian. Granite embedding r2 models, 2025. URL https://arxiv.org/abs/2508.21085.

Alexander Bienstock, Sarvar Patel, Joon Young Seo, and Kevin Yeo. Batch PIR and labeled PSI with oblivious ciphertext compression. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 5949–5966, 2024. URL https://www.usenix.org/conference/usenixsecurity24/presentation/bienstock.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2206–2240, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Yihang Cheng, Lan Zhang, Junyang Wang, Mu Yuan, and Yunhao Yao. RemoteRAG: A privacy-preserving LLM cloud RAG service. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3820–3837, 2025. URL https://aclanthology.org/2025.findings-acl.197/.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pp. 253–262, 2004. URL https://doi.org/10.1145/997817.997857.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, pp. 211–407, 2014. URL https://doi.org/10.1561/0400000042.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2288–2292, 2021. URL https://doi.org/10.1145/3404835.3463098.

Michael J. Freedman, Yuval Ishai, Benny Pinkas, and Omer Reingold. Keyword search and oblivious pseudorandom functions. In *Theory of Cryptography*, pp. 303–324, 2005.

Georg Fuchsbauer, Riddhi Ghosal, Nathan Hauke, and Adam O'Neill. Approximate distance-comparison-preserving symmetric encryption. In *Security and Cryptography for Networks*, pp. 117–144, 2022.

Gayathri Garimella, Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. Oblivious key-value stores and amplification for private set intersection. In *Advances in Cryptology – CRYPTO 2021*, volume 12826, pp. 395–425, 2021.

Nicolas Grislain. RAG with differential privacy. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pp. 847–852, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3929–3938, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. Privacy implications of retrieval-based language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14887–14902, 2023. URL https://aclanthology.org/2023.emnlp-main.921/.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021. URL https://aclanthology.org/2021.eacl-main.74/.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 117–128, 2011.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, pp. 1–38, 2023. URL https://doi.org/10.1145/3571730.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, pp. 535–547, 2021.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017. URL https://aclanthology.org/P17-1147/.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. URL https://aclanthology.org/2020.emnlp-main.550/.

Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach, 2020. URL https://arxiv.org/abs/2010.01195.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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 Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, pp. 452–466, 2019. URL https://aclanthology.org/Q19-1026/.

Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, pp. 331–339, 1995. URL https://www.sciencedirect.com/science/article/pii/B9781558603776500487.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Zhijun Li, Minghui Xu, Huayi Qi, Wenxuan Yu, Tingchuang Zhang, Qiao Zhang, GuangYong Shang, Zhen Ma, and Xiuzhen Cheng. PRAG: End-to-end privacy-preserving retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2604.26525.

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*, pp. 61885–61904, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6529ad2d7e2cf7bbc76785684b794d57-Paper-Conference.pdf.

Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. pp. 1941–1942, 2018.

Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 824–836, 2020.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human-generated MAchine reading COmprehension dataset, 2017. URL https://openreview.net/forum?id=Hk1iOLcle.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, pp. 333–389, 2009. URL https://doi.org/10.1561/1500000019.

Anshumali Shrivastava and Ping Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In *Advances in Neural Information Processing Systems*, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/c98e7c4b8f20d384e3ad857d0ee226cc-Paper.pdf.

Nandan Thakur, Nils Reimers, Andreas Rucklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of NeurIPS Datasets and Benchmarks*, 2021.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1113–1120, 2009. URL https://doi.org/10.1145/1553374.1553516.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. URL https://aclanthology.org/D18-1259.

Dixi Yao and Tian Li. Differentially private retrieval augmented generation with random projection. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=5DfhoxRPXh.

Huanyi Ye, Jiale Guo, Ziyao Liu, and Kwok-Yan Lam. Efficient privacy-preserving retrieval augmented generation with distance-preserving encryption. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pp. 668–676, 2025.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG), 2024. URL https://aclanthology.org/2024.findings-acl.267.

A Detailed Preliminaries

A.1 Conditional Approximate Distance-Preserving Encryption

We instantiate each path with an independent conditional scale-and-perturb construction. The concrete algorithm of conditional approximate distance encryption consists of key generation, bounded noise generation, database encryption, query encryption, and decryption.

1. Key Generation $\mathsf{KeyGen}(1^\lambda, d_p, \beta_p)$. For path p, it samples a master key K_p and derives $(K_p^R, K_p^{db}, K_p^q, K_p^{bucket}, K_p^{alias}, K_p^{vec}) \leftarrow \mathsf{HKDF}(K_p)$. It also samples a positive scale s_p and creates a keyed orthogonal transform $R_p = P_p D_{p,2} H D_{p,1}$, where $D_{p,1}, D_{p,2}$ are random sign diagonals, H is a normalized Walsh-Hadamard transform, and P_p is a permutation.
2. Bounded Noise Generation $\mathsf{Ball}(K, r, \rho, d)$. Given key K, nonce r, radius ρ, and dimension d, it uses $\mathsf{PRF}(K, r)$ to sample $g \sim \mathcal{N}(0, I_d)$ and $u \sim U(0,1)$, then returns $n = \rho u^{1/d} \frac{g}{\|g\|_2}$. Thus n is distributed in a d-dimensional ball and $\|n\|_2 \le \rho$.
3. Database Encryption $\mathsf{DBEnc}(sk_p, v_i^p)$. For $\|v_i^p\|_2 \le 1$, it samples $r_i^p \leftarrow \{0,1\}^\lambda$ and computes $n_i^p = \mathsf{Ball}\left(K_p^{db}, r_i^p, \frac{3 s_p \beta_p}{8}, d_p\right)$ and $c_i^p = s_p R_p v_i^p + n_i^p$. The encrypted database is output as $C_i^p = (c_i^p, r_i^p)$.
4. Query Encryption $\mathsf{QEnc}(sk_p, v_q^p)$. For the query vector v_q^p, it independently samples r_q^p and computes $\eta_q^p = \mathsf{Ball}\left(K_p^q, r_q^p, \frac{s_p \beta_p}{8}, d_p\right)$ and $c_q^p = s_p R_p v_q^p + \eta_q^p$. The encrypted query is only output as c_q^p.
5. Decryption $\mathsf{Dec}(sk_p, (c, r))$. Given (c_i^p, r_i^p), it recomputes n_i^p and recovers the database vector $v_i^p = R_p^\top \frac{c_i^p - n_i^p}{s_p}$.

Algorithm 3: Conditional distance encryption for path p

Input: Security parameter 1^λ, dimension d_p, tolerance β_p
Output: DPE key sk_p, auxiliary path keys aux_p, ciphertext C, query token T, or recovered vector v

Function `KeyGen` $(1^\lambda, d_p, \beta_p)$ **:**
- $K_p \xleftarrow{\$} \{0,1\}^\lambda$;
- $(K_p^R, K_p^{db}, K_p^q, K_p^{bucket}, K_p^{alias}, K_p^{vec}) \leftarrow \mathsf{HKDF}(K_p)$;
- Derive random-sign diagonals $D_{p,1}, D_{p,2}$ and permutation P_p from K_p^R;
- $R_p \leftarrow P_p D_{p,2} H D_{p,1}$;
- sample $s_p \in \mathbb{R}_{>0}$;
- $sk_p \leftarrow (s_p, \beta_p, R_p, K_p^{db}, K_p^q)$;
- $aux_p \leftarrow (K_p^{bucket}, K_p^{alias}, K_p^{vec})$;
- **return** (sk_p, aux_p);

Function `Ball` (K, r, ρ, d) **:**
- Use $\mathsf{PRF}(K, r)$ to sample $g \sim \mathcal{N}(0, I_d)$ and $u \sim U(0,1)$;
- **return** $\rho u^{1/d} g / \|g\|_2$;

Function `DBEnc` (sk_p, v) **:**
- Require $\|v\|_2 \le 1$; $r \xleftarrow{\$} \{0,1\}^\lambda$;
- $n \leftarrow \mathsf{Ball}(K_p^{db}, r, 3 s_p \beta_p / 8, d_p)$;
- $c \leftarrow s_p R_p v + n$;
- **return** $C = (c, r)$;

Function `QEnc` (sk_p, v) **:**
- $r \xleftarrow{\$} \{0,1\}^\lambda$;
- $\eta \leftarrow \mathsf{Ball}(K_p^q, r, s_p \beta_p / 8, d_p)$;
- **return** $T = s_p R_p v + \eta$;

Function `Dec` $(sk_p, (c, r))$ **:**
- $n \leftarrow \mathsf{Ball}(K_p^{db}, r, 3 s_p \beta_p / 8, d_p)$;
- **return** $R_p^\top (c - n) / s_p$;

A.2 Compartment Conditional Approximate Distance-Preserving Encryption

Let $u_i^p = \mathsf{DBEnc}(sk_p, v_i^p)$ and $u_q^p = \mathsf{QEnc}(sk_p, v_q^p)$ denote path-level owner and client intermediates that are never released. For every path compartment b, HKDF independently samples r_c coordinates uniformly without replacement to form $S_{p,b} \in \{0,1\}^{r_c \times d_p}$, and derives a sign diagonal $D_{p,b} \in \{-1, +1\}^{r_c \times r_c}$, a translation $t_{p,b}$, and a local alias key. With $A_{p,b} = \sqrt{d_p / r_c} D_{p,b} S_{p,b}$, the

released coordinates are

$$\mathsf{CmpDBEnc}_{p,b}(v_i) = A_{p,b}u_i^p + t_{p,b}, \quad \mathsf{CmpQEnc}_{p,b}(v_q) = A_{p,b}u_q^p + t_{p,b}.$$

The common translation cancels only inside (p, b), giving $\|c_{q,b}^p - c_{i,b}^p\|_2 = \|A_{p,b}(u_q^p - u_i^p)\|_2$. The sign diagonal does not alter Euclidean norms, whereas coordinate selection introduces additional projection distortion.

Algorithm 4: Dual-path Compartment-DPE refinement

Input: Path candidate compartments $\{\mathcal{H}_{p,b}\}_{p,b}$, vectors $\{v_i^p\}$, local dimension r_c
Output: Cell-local ciphertexts and fixed-shape local rankings

foreach *path* p *and nonempty compartment* b **do**
  derive $(A_{p,b}, t_{p,b}, K_{p,b}^{alias}, K_{p,b}^{vec})$;
  foreach $i \in \mathcal{H}_{p,b}$ **do**
    Derive $a_{i,b}^p \leftarrow \mathsf{PRP}_{K_{p,b}^{alias}}(i\|b)$;
    Compute $c_{i,b}^p \leftarrow \mathsf{CmpDBEnc}_{p,b}(v_i^p)$;
    Rerandomize path-native reranking-payload and join-capsule ciphertexts;
foreach *accessed* (p, b) *for query* q **do**
  $c_{q,b}^p \leftarrow \mathsf{CmpQEnc}_{p,b}(v_q^p)$;
  rank only $\mathcal{H}_{p,b}$ by $\|c_{q,b}^p - c_{i,b}^p\|_2$ and emit fixed local top t_p;
Client union local outputs and recomputes native cosine/BM25 scores;

A.3 Dual-Path Representations

The semantic and lexical paths rely on dense cosine similarity and sparse BM25 lexical relevance as fundamentally scoring functions, respectively. Therefore, they cannot directly share a single encryption-and-ranking mechanism. Our objective is to map both heterogeneous retrieval signals onto a common Euclidean-distance ranking interface, so that CCADPE can be instantiated per path and reused for cloud-side candidate ranking. For each path, we derive unit-norm query and document representations whose Euclidean distance is monotonically related to the native relevance score of that path. Superscripts s and l denote the semantic and lexical paths, respectively.

A.3.1 Semantic Path

Let E be a text encoder. For a passage d_i and a query q, we encode both inputs and normalize the resulting embeddings to unit norm:

$$v_i^s = \frac{E(d_i)}{\|E(d_i)\|_2}, \qquad v_q^s = \frac{E(q)}{\|E(q)\|_2}.$$

For any pair of unit vectors, the squared Euclidean distance and the cosine similarity are related by the identity

$$\|v_q^s - v_i^s\|_2^2 = 2 - 2\cos(E(q), E(d_i)),$$

Thus, for a fixed query q, ranking documents by ascending $\|v_q^s - v_i^s\|_2$ exactly reproduces the ranking by descending cosine similarity. This equivalence allows the cloud to rank DPE ciphertexts by Euclidean distance, while exact cosine scoring is deferred to client-side reranking.

A.3.2 Lexical Path

For vocabulary term t, define the document coordinate

$$x_{i,t} = \mathrm{IDF}(t)\frac{f(t, d_i)(k_1 + 1)}{f(t, d_i) + k_1(1 - b + b|d_i|/\,\mathrm{avgdl})}.$$

$\mathrm{IDF}(t)$ is inverse document frequency of the term t. $f(t, d_i)$ is the term frequency of t in document d_i. k_1, b are BM25 hyperparameters. $|d_i|$ and $avgdl$ are document and mean document lengths. Let

$x_i = (x_{i,t})_{t\in V}$ and let $y_q = (y_{q,t})_{t\in V}$ be the unique-term query indicator, where $y_{q,t} = \mathbf{1}[t \in q]$. The native BM25 score is

$$\mathrm{BM25}(q, d_i) = \sum_{t\in V} y_{q,t}x_{i,t} = \sum_{t\in q} x_{i,t} = y_q^\top x_i.$$

However, encrypting the raw $|V|$-dimensional sparse vectors x_i and y_q directly would require cryptographic operations over the full vocabulary and is computationally prohibitive at practical corpus scales. To reduce the dimension, we compress the sparse representations with signed feature hashing. Let $S \in \{-1, 0, +1\}^{m_l\times|V|}$ be a signed feature-hashing matrix in which each vocabulary column contains exactly one nonzero entry, placed by a bucket hash $h : V \to \{1, \dots, m_l\}$ and signed by a sign hash $\xi : V \to \{-1, +1\}$. Formally, for all $j \neq h(t)$, we have $S_{h(t),t} = \xi(t)$ and $S_{j,t} = 0$. The compressed vectors are

$$\hat{x}_i = Sx_i, \qquad \hat{y}_q = Sy_q,$$

which reduce the ambient dimension from $|V|$ to $m_l \ll |V|$. Then, lexical ranking becomes a maximum-inner-product search (MIPS) over the sketched vectors. For a fixed query, one maximizes $\hat{y}_q^\top \hat{x}_i$. To express this objective through the Euclidean-distance interface, we apply the asymmetric MIPS-to-Euclidean reduction. Let $M \geq \max_i \|\hat{x}_i\|_2$ be an upper bound on the sketched document norms, and define the asymmetric projections

$$P(\hat{x}_i) = \left[\frac{\hat{x}_i}{M}; \sqrt{1 - \left\|\frac{\hat{x}_i}{M}\right\|_2^2}\right], \quad Q(\hat{y}_q) = \left[\frac{\hat{y}_q}{\|\hat{y}_q\|_2}; 0\right],$$

where $[\cdot;\cdot]$ denotes vector concatenation. Both $P(\hat{x}_i)$ and $Q(\hat{y}_q)$ lie in $\mathbb{R}^{m_l+1}$ and have unit norm. Expanding the squared Euclidean distance yields

$$\|Q(\hat{y}_q) - P(\hat{x}_i)\|_2^2 = 2 - \frac{2\hat{y}_q^\top \hat{x}_i}{M\|\hat{y}_q\|_2}. \tag{1}$$

For a fixed query, minimizing Eq.1 maximizes the sketched BM25 score. We set $v_i^l = P(\hat{x}_i)$ and $v_q^l = Q(\hat{y}_q)$ for CCADPE refinement. Separately, the lexical payload stores a keyed token-ID sequence with multiplicities and $|d_i|$ under AEAD, together with client-held N, avgdl, and document frequencies. It reconstructs $x_{i,t}$ and therefore $y_q^\top x_i$ exactly for every returned candidate.

B Proof Details

B.1 Proof of Conditional Comparison Preservation

Preliminary projection-distortion bounds for CCADPE. For any nonzero vector $z \in \mathbb{R}^{d_p}$, we define coordinate coherence by

$$\mu(z) = \frac{d_p\|z\|_\infty^2}{\|z\|_2^2} \in [1, d_p].$$

Let $\kappa(\varepsilon) = 2\varepsilon - \varepsilon^2$. By the Bernstein-Serfling concentration inequality for sampling without replacement, for $0 < \varepsilon < 1$ we have

$$\Pr\left[\left|\|A_{p,b}z\|_2 - \|z\|_2\right| > \varepsilon\|z\|_2\right] \leq 2\exp\left(-\frac{r_c\kappa(\varepsilon)^2}{2\mu(z)(1 + \kappa(\varepsilon)/3)}\right).$$

Let $Z_{p,b}$ be the set of intermediate query–document and document–document differences evaluated inside compartment (p, b), and let $\mu^*_{p,b} = \max_{z\in Z_{p,b}} \mu(z)$. Applying a union bound across all elements in $Z_{p,b}$ yields the projection failure probability

$$\delta_{p,b}^{\text{proj}} \leq 2M_{p,b}\exp\left(-\frac{r_c\kappa(\varepsilon_{p,b}^{\text{proj}})^2}{2\mu^*_{p,b}(1 + \kappa(\varepsilon_{p,b}^{\text{proj}})/3)}\right). \tag{2}$$

A sufficient dimension for target projection failure $\delta_{p,b}^{proj}$ is

$$r_c \geq \frac{2\mu^*_{p,b}(1 + \kappa/3)}{\kappa^2}\log\frac{2M_{p,b}}{\delta_{p,b}^{proj}}, \quad \kappa = 2\varepsilon_{p,b}^{proj} - \left(\varepsilon_{p,b}^{proj}\right)^2.$$

Randomized Hadamard flattening bounds the maximal coherence over the workload:

$$\Pr\left[\mu^*_{p,b} > 2\log\frac{2d_pM_{p,b}}{\delta^{flat}_{p,b}}\right] \le \delta^{flat}_{p,b}.$$

Define the total compartment failure probability as

$$\delta^{cc}_{p,b} := \delta^{flat}_{p,b} + \delta^{proj}_{p,b}.$$

When coherence is measured empirically for a fixed workload, we set $\delta^{flat}_{p,b} = 0$, hence $\delta^{cc}_{p,b} = \delta^{proj}_{p,b}$.

Combining CADPE bounded noise and CCADPE projection error. At the unreleased path-DPE layer, ciphertext intermediates are

$$u_q = sRv_q + \eta, \qquad u_i = sRv_i + n_i,$$

where R is orthogonal and the noise terms satisfy

$$\|\eta\|_2 \le \frac{s\beta}{8}, \qquad \|n_i\|_2 \le \frac{3s\beta}{8}.$$

Let $\delta_i = \|v_q - v_i\|_2$ and $z_i = u_q - u_i$. By reverse triangle inequality and orthogonality, we have

$$\left|\frac{\|z_i\|_2}{s} - \delta_i\right| \le \frac{\beta}{2}.$$

The coordinate actually released to the server is $c_{q,b} = A_bu_q + t_b$ and analogously for documents, so the local distance is $\delta'_{i,b} = \|A_bz_i\|_2$. Conditioned on the event $\mathcal{E}_b(\varepsilon^{\mathrm{proj}}_b)$ from Eq. 2,

$$\left|\frac{\delta'_{i,b}}{s} - \delta_i\right| \le \frac{|\|A_bz_i\|_2 - \|z_i\|_2|}{s} + \left|\frac{\|z_i\|_2}{s} - \delta_i\right| \le \varepsilon^{\mathrm{proj}}_b\left(\delta_i + \frac{\beta}{2}\right) + \frac{\beta}{2}.$$

Since all path vectors are unit norm, $\delta_i \le 2$, and hence

$$\left|\frac{\delta'_{i,b}}{s} - \delta_i\right| \le \Gamma^q_b := \frac{\beta}{2} + \varepsilon^{\mathrm{proj}}_b\left(2 + \frac{\beta}{2}\right).$$

For two documents i, j in the same compartment, if $\delta_i < \delta_j - 2\Gamma^q_b$, then

$$\frac{\delta'_{i,b}}{s} \le \delta_i + \Gamma^q_b < \delta_j - \Gamma^q_b \le \frac{\delta'_{j,b}}{s}.$$

This yields $\delta'_{i,b} < \delta'_{j,b}$ on $\mathcal{E}_b$. By union bound, the simultaneous statement fails with probability at most $\delta^{\mathrm{cc}}_{p,b}$. This completes the proofs of Theorem 1.

B.2 Proof of Stored-Geometry Distortion

For two stored documents in compartment b, let $z_{ij} = u_i - u_j = sR(v_i - v_j) + (n_i - n_j)$, $d_{ij} = \|v_i - v_j\|_2$, and $d'_{ij,b} = \|A_bz_{ij}\|_2$. From CADPE document noise bounds and orthogonality:

$$\left|\frac{\|z_{ij}\|_2}{s} - d_{ij}\right| \le \frac{3\beta}{4}.$$

Conditioned on $\mathcal{E}_b(\varepsilon^{\mathrm{proj}}_b)$ and using $d_{ij} \le 2$,

$$\left|\frac{d'_{ij,b}}{s} - d_{ij}\right| \le \frac{3\beta}{4} + \varepsilon^{\mathrm{proj}}_b\left(d_{ij} + \frac{3\beta}{4}\right) \le \Gamma^{db}_b := \frac{3\beta}{4} + \varepsilon^{\mathrm{proj}}_b\left(2 + \frac{3\beta}{4}\right).$$

Thus, for two stored pairs represented in the same compartment, if $d_{ij} < d_{k\ell} - 2\Gamma^{db}_b$, then $d'_{ij,b} < d'_{k\ell,b}$. The simultaneous statement fails with probability at most $\delta^{\mathrm{cc}}_{p,b}$. This completes the proofs of Theorem 2.

B.3 Lexical Error Propagation and Ranking Correctness

The lexical path introduces approximation at three distinct stages: a quantized posting cap, feature-hashed representation, and CCADPE refinement. We derive the corresponding error terms below and then state a sufficient ranking condition. The semantic path has no hashing or lexical-quantization error, but remains subject to the probabilistic CCADPE comparison condition in Theorem 1 and to candidate truncation.

Signed feature-hashing sketch error. Signed feature hashing compresses high-dimensional sparse BM25 vectors into lower-dimensional representations, but introduces statistical sketch error. It is critical to distinguish probabilistic bounds over the hash-generation randomness at setup time from the deterministic behaviour of the fixed keyed hash matrix deployed during system inference. The following lemma characterizes the statistical properties of the inner-product estimate produced by signed feature hashing.

Lemma 1 (Sketch error) *Let the bucket hash be pairwise-uniform and the Rademacher sign hash be four-wise independent. For arbitrary fixed real-valued vectors $\boldsymbol{x}, \boldsymbol{y}$, define the sketched inner-product estimate $Z = (S\boldsymbol{y})^\top (S\boldsymbol{x})$. Then Z is an unbiased estimator of the true inner product $\boldsymbol{y}^\top \boldsymbol{x}$, with variance*

$$\mathrm{Var}[Z] = \frac{1}{m_l}\left(\|\boldsymbol{x}\|_2^2 \|\boldsymbol{y}\|_2^2 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 - 2 \sum_t x_t^2 y_t^2 \right).$$

By the Chebyshev inequality, for any $\nu \in (0, 1)$,

$$\Pr\left[|Z - \boldsymbol{y}^\top \boldsymbol{x}| \geq \sqrt{\mathrm{Var}[Z]/\nu} \right] \leq \nu.$$

Proof. Let h_t denote the bucket hash and $\xi_t \in \{-1, +1\}$ the Rademacher sign hash for vocabulary entry t. By construction of the signed feature-hashing matrix S, we have $Z - \boldsymbol{y}^\top \boldsymbol{x} = \sum_{a<b} \xi_a \xi_b \mathbf{1}[h_a = h_b] (x_a y_b + x_b y_a)$. Four-wise independence of sign random variables cancels cross-terms for distinct unordered pairs, while pairwise-uniform bucket hashing yields collision probability $1/m_l$. Direct expansion yields the stated variance expression. Unbiasedness $\mathbb{E}[Z] = \boldsymbol{y}^\top \boldsymbol{x}$ follows immediately. Application of Chebyshev's inequality completes the argument.

This probability bound holds over the random choice of hash functions at system setup. It is not a uniform simultaneous guarantee over the full document corpus for a single fixed deployed hash seed; a naive union-bound over all documents would produce overly loose bounds. We denote the maximum sketch error for a query-document pair as ϵ_{sk}.

Candidate-index quantization error. The posting cap uses uniformly quantized BM25 coordinates. Let $e_{i,t} = \mathrm{Quant}_{\Delta_q}(x_{i,t}) - x_{i,t}$ denote the quantization residual for term t in document i. In our implementation, we use a bounded uniform quantizer without clipping, so each per-term residual satisfies $|e_{i,t}| \leq \Delta_q/2$, where Δ_q is the quantization step size. For a given query with term-weight vector $\boldsymbol{y}_q$, the score perturbation for document i satisfies

$$\left| \sum_t y_{q,t}\, e_{i,t} \right| \leq \frac{\Delta_q}{2} \sum_t |y_{q,t}| = \frac{\Delta_q}{2} \|\boldsymbol{y}_q\|_1.$$

Thus the pairwise gap error is at most $\epsilon_{\mathrm{quant},l} = \Delta_q \|\boldsymbol{y}_q\|_1$. Consider two documents i, j, if $\Delta_{ij}^l = y_q^\top x_i - y_q^\top x_j > \epsilon_{\mathrm{quant},l}$, quantization alone cannot reverse i and j at the posting-cap gate. This condition does not by itself guarantee inclusion: the selected postings and the cap C_l must also contain i.

CCADPE refinement error. We now incorporate CCADPE refinement error with both CADPE noise and compartment-projection distortion. For two candidates admitted to the same compartment, suppose $|\hat{y}_q^\top \hat{x}_u - y_q^\top x_u| \leq \epsilon_{\mathrm{sk}}$ for $u \in \{i, j\}$ and let $\Delta_{ij}^l = y_q^\top x_i - y_q^\top x_j > 0$. For the unit-norm lexical-path vectors in compartment b, write $\delta_i = \|v_q^l - v_i^l\|_2$ and $\rho_{i,b} = \|c_{q,b}^l - c_{i,b}^l\|_2 / s_l$. Appendix B.1 gives, on $\mathcal{E}_{l,b}$,

$$|\rho_{i,b} - \delta_i| \leq \Gamma_{l,b}^q, \qquad \Gamma_{l,b}^q = \frac{\beta_l}{2} + \varepsilon_{l,b}^{\mathrm{proj}} \left(2 + \frac{\beta_l}{2} \right),$$

and therefore $|\rho_{i,b}^2 - \delta_i^2| \le \Gamma_{l,b}^q(4 + \Gamma_{l,b}^q)$. Mapping this distortion back into BM25 score units via Appendix A.3, the CCADPE-induced score-gap distortion is bounded by

$$\epsilon_{\mathrm{CCADPE},l} \le M\|\hat{\boldsymbol{y}}_q\|_2\Gamma_{l,b}^q(4 + \Gamma_{l,b}^q).$$

Combining sketch and CCADPE distortion, conditioned on $\mathcal{E}_{l,b}$, their local order is preserved if

$$\Delta_{ij}^l > 2\epsilon_{\mathrm{sk}} + \epsilon_{\mathrm{CCADPE},l}.$$

For a workload covered by $\mathcal{Z}_{l,b}$ these conditions hold simultaneously with probability at least $1-\delta_{l,b}^{\mathrm{cc}}$. Survival requires two sequential events: admission by the posting planner/cap, governed in part by $\epsilon_{\mathrm{quant},l}$, and survival among the local top-t_l, governed by $2\epsilon_{\mathrm{sk}} + \epsilon_{\mathrm{CCADPE},l}$. The errors are not added because they act on different gates. Once i and j are returned, the client evaluates $y_q^\top x_i$ and $y_q^\top x_j$ directly, so their final pairwise BM25 order has no quantization, sketch, or CCADPE score error. Candidate-local exactness does not restore any document discarded by either gate.

B.4 Hybrid Proof of Concrete-View Wrapper Simulation

This section proves Theorem 3 for the protocol instantiated with the concrete DPE construction. For both paths, the concrete view is $\mathsf{G}_p^{\mathsf{cmp}}$. Let $\mathsf{Real}_{\Pi^{\mathsf{conc}},\mathcal{A}}(1^\lambda, \mathcal{D}, \mathbf{q})$ denote the view of the real adversary for a polynomial-length adaptive query sequence $\mathbf{q} = (q_1, \ldots, q_T)$. In the ideal experiment, simulator $\mathcal{S}$ receives $\mathcal{L}^{\mathsf{conc}}(\mathcal{D}, \mathbf{q})$. We use hybrids H_0, H_1, H_2, H_3. We argue computational indistinguishability between successive hybrids.

1. H_0**: Real execution.** This hybrid runs the actual protocol Π^{conc}. The adversary obtains stored document coordinates, each adaptive query coordinate, index and access traces, returned aliases, and ciphertext payloads and capsules.
2. H_1**: Replace keyed labels.** Replace HKDF/HMAC-derived bucket labels and auxiliary values with consistent random-function outputs. The simulator maintains one value per equality class exposed by $\mathsf{SearchEq}$ and preserves the incidence, size, and access relations in $\mathcal{L}^{\mathsf{conc}}$. Any distinguisher between H_0 and H_1 directly yields a PPT distinguisher against the PRF security. Hence we have $H_0 \overset{c}{\approx} H_1$.
3. H_2**: Replace path aliases.** Replace every semantic and lexical compartment-local PRP alias with an independent random permutation. A semantic document replicated across cells is represented by unrelated abstract handles, while each lexical document is represented by the abstract handle of its keyed partition. The simulator assigns aliases consistently to the abstract handles appearing in $\mathsf{PostInc}$, Acc, and Cand, but creates no semantic-lexical join map. Distinguishing H_1 from H_2 breaks PRP security. Thus, we have $H_1 \overset{c}{\approx} H_2$.
4. H_3**: Replace protected plaintexts.** Replace every document payload, path-native reranking payload, and join-capsule plaintext by an all-zero string of the same leaked length and use fresh nonces. Multi-message IND-CPA of AEAD gives $H_2 \overset{c}{\approx} H_3$. The independently randomized capsules introduce no explicit cross-path equality signal.

At H_3, the simulator generates every non-DPE value from $\mathcal{L}^{\mathsf{conc}}$: it samples consistent bucket labels and path-local aliases, instantiates AEAD under sampled keys on equal-length zero strings, and reproduces the declared posting, candidate, access, response, and timing relations. The simulator receives $(\mathsf{G}_s^{\mathsf{cmp}}, \mathsf{G}_l^{\mathsf{cmp}})$ as explicit leakage and derives the corresponding within-compartment DistTr values, which are redundant given the released coordinate views. For adaptive queries, it extends its label and alias tables only when a newly leaked equality class or handle appears, then replays the newly supplied query coordinate and operational trace. Thus H_3 has the ideal distribution generated from $\mathcal{L}^{\mathsf{conc}}$. By transitivity, $\mathsf{Real}_{\Pi^{\mathsf{conc}},\mathcal{A}} \overset{c}{\approx} \mathsf{Ideal}_{\mathcal{S},\mathcal{A}}(1^\lambda, \mathcal{L}^{\mathsf{conc}}(\mathcal{D}, \mathbf{q}))$. This proof establishes that the wrapper does not amplify the declared leakage.

C Dataset Details

Table 5 records the corpus sizes and query workloads used in the retained evaluation.

Table 5: Evaluation datasets.

Dataset	Documents	Evaluated queries
MS MARCO (Nguyen et al. (2017))	8,841,823	43 test; 6,980 dev
NQ (Kwiatkowski et al. (2019))	2,681,468	3,452 test
HotpotQA (Yang et al. (2018))	5,233,329	7,405 test
20 Newsgroups (Lang (1995))	4,500	3,000 train; 1,500 test
FiQA (Maia et al. (2018))	57,638	300 test
TriviaQA (Joshi et al. (2017))	100,000	10000 test

D Ablation Studies

We collect hyper-parameter sensitivity ablations for DuetER, including compartment-local return depth, semantic IVF probe count, and the amount of target-domain data used to calibrate DuetRank.

Local Return Depth Sensitivity. We study the trade-off between retrieval accuracy and number of candidates by varying the number of local depth, with results summarized in Table 6. Increasing t_p beyond the selected point provides diminishing returns. For the semantic path, the larger setting nearly doubles the candidate union but changes nDCG@10 by at most .00053. Lexical unions approach the 1,000-candidate planner cap on NQ and MS MARCO, while the additional quality gains remain limited. Although Recall@100 increases with the increase of t_p, the improvement is limited when t_p exceeds the parameter we chose.

Table 6: Compartment local-depth sensitivity.

		Smaller depth			Selected depth			Larger depth		
Data	Path	t_p	nDCG@10/R@100	Mean union	t_p	nDCG@10/R@100	Mean union	t_p	nDCG@10/R@100	Mean union
NQ	Semantic	8	.51723/.86618	962	16	.51812/.90670	1,907	32	.51782/.92539	3,777
NQ	Lexical	32	.23140/.49585	510	64	.27435/.65585	958	128	.27551/.66517	996
HotpotQA	Semantic	8	.63836/.74600	968	16	.63909/.76194	1,927	32	.63962/.77322	3,830
HotpotQA	Lexical	32	.48767/.55820	508	64	.57348/.69379	954	128	.57672/.70101	991
MS MARCO	Semantic	16	.28790/.72250	1,934	32	.28799/.76197	3,842	64	.28805/.77721	7,622
MS MARCO	Lexical	32	.19874/.50213	495	64	.22262/.63145	925	128	.22373/.63703	961

Note. Mean union is the number of candidates reranked by the client.

IVF probe-count sensitivity. We further ablate the number of lookup probes for the residual spherical-IVF. As shown in Table 7, larger probe settings boost nDCG@10 and Recall@100 at the cost of scanning more embedding vectors. The 128-probe configuration adopted in our main experiments represents a natural quality-cost knee. Further increasing probes to 256 brings only marginal nDCG@10 and Recall@100 improvement while substantially expanding candidate volume.

Table 7: Final-system semantic probe-count sensitivity.

Data	n_{probe}	Scan	Mean union	RefCov	nDCG@10	Recall@100
NQ	64	6.86%	948	.6822	.51596	.89871
	128	13.19%	1,907	.7295	.51833	.90967
	256	25.59%	3,828	.7608	.51984	.91657
HotpotQA	64	6.64%	960	.6391	.62950	.74571
	128	12.81%	1,927	.6900	.63925	.76178
	256	25.03%	3,866	.7214	.64365	.76982
MS MARCO	64	6.69%	1,913	.7888	.28449	.74801
	128	12.99%	3,842	.8233	.28799	.76197
	256	25.38%	7,701	.8308	.28865	.77012

Note. Scan counts examined, possibly replicated, cell entries relative to corpus size; RefCov is reference Top-100 coverage in the returned compartment union. Local depths are 16 for NQ/HotpotQA and 32 for MS MARCO.

DuetRank Calibration Robustness DuetRank uses client-side gradient boosted for semantic-lexical fusion with RRF fallback. We test sensitivity to calibration-set size across ten deterministic seeds. Table 8 refits DuetRank on nested deterministic subsets of the original calibration half while holding model capacity and the semantic-backoff coefficient fixed. The

results show that all corpora significantly outperform their semantic baselines at 50%, and full calibration produces the highest raw nDCG@10.

Table 8: DuetRank calibration robustness on final compartment rankings.

Data	Calibration	Queries	DuetRank nDCG@10
NQ	25%	428	.52107
	50%	856	.51997
	100%	1,713	**.52167**
HotpotQA	25%	916	.68147
	50%	1,832	.68169
	100%	3,665	**.68217**
MS MARCO	25%	864	.29408
	50%	1,727	.30205
	100%	3,454	**.30260**

Summary. The above ablations illustrate inherent trade-offs between computational/communication resource consumption and retrieval effectiveness. For DuetRank, sufficient in-domain calibration queries are required to guarantee stable improvements, and the built-in RRF fallback mitigates risk under limited calibration data. These observations confirm that corpus-specific calibration is required to select appropriate hyper-parameter values.

E Parameter Selection and Calibration

This appendix explains the selection of the system parameters and distinguishes representation dimensions, compartment parameters, candidate budgets, and reproducibility settings, as detailed in Table 9.

Table 9: Parameters used in the complete-corpus experiments.

Path	Parameter	Final value	Role
Semantic	Encoder/output dimension	$384/512$	384-dimensional encoder, zero-padded to 512-dimension
Semantic	Local dimension $r_{c,s}$	256	Coordinates released per semantic compartment
Semantic	IVF cells	2,048	Residual spherical-IVF partition count
Semantic	n_{probe}	128	Semantic cells accessed per query
Semantic	Local return depth t_s	16 for NQ/HotpotQA; 32 for MS MARCO	Per-cell semantic return depth
Lexical	Signed-hash dimension m_l	1,024	Dimension of the lexical BM25 proxy
Lexical	CCADPE work dimension d_l	8,192	Zero-padded dimension used by the keyed selectors
Lexical	Local dimension $r_{c,l}$	32	Coordinates released per lexical compartment
Lexical	Compartments K_l^{cmp}	16	Keyed lexical partitions addressed by every query
Lexical	Local return depth t_l	64	Per-partition encrypted-ranking depth
Lexical	Planner cap C_l	1,000	Maximum candidates entering CCADPE refinement
Lexical	Exact-reranking cap	300	Maximum candidates reranked by client-side BM25
Lexical	Posting budget B_l	50,000 for NQ/HotpotQA; 250,000 for MS MARCO	Maximum cumulative posting volume per query
Both	CADPE tolerance β_p	0.1	Bounded-noise parameter
Both	CADPE scale s_p	3	Common coordinate scale
Both	Coordinate representation	FP16	Server-side storage and transmission format
Both	Experiment seed	20260917	Reproducible transforms, noise, and selectors