

When Metadata Remembers: Ordered Provenance Enables Document-Level Embedding Inversion

Liwen Zheng¹, Qing Li^{2,✉}, Qingsong Zou¹, Yong Jiang¹

¹Tsinghua Shenzhen International Graduate School, Shenzhen, China

²Peng Cheng Laboratory, Shenzhen, China

¹{zhenglw25, zouqs24}@mails.tsinghua.edu.cn jiangy@sz.tsinghua.edu.cn

²andyliqing@gmail.com

Abstract

Text embeddings in RAG indexes, vector stores, and agent memories are rarely leaked as isolated vectors: they often appear with document IDs, chunk offsets, and preserved order. This paper shows that such ordered provenance is not passive metadata but an active privacy channel. We introduce MIRAGE, a public-shadow, zero-victim-pair attack that exploits ordered document-chunk leakage by turning facts recovered from earlier chunks into encoder-verified anchors for later reconstruction. MIRAGE builds public-shadow single-chunk candidates, extracts typed anchors, verifies them with counterfactual encoder-support checks, and uses verified memory to regenerate document-consistent chunks. Across five dense encoders and four decoders on ordered Enron and synthetic CanaryDocs, under a strict 512-query victim-time budget per chunk, MIRAGE outperforms strong fixed-bank reranking, improving macro privacy-unit recall by 5.65–6.10 points across encoders and 5.2–6.9 points across decoders; gains persist (+3.7–4.4 points) on two real, non-templated corpora. Its support check raises accepted-anchor precision from 43.6% to 69.4%, and 34.8% of newly recovered privacy units appear only in regenerated candidates absent from every fixed bank. These results reveal document grouping, chunk order, and offsets as sensitive attack surfaces that RAG and vector-store deployments must protect alongside embeddings themselves.

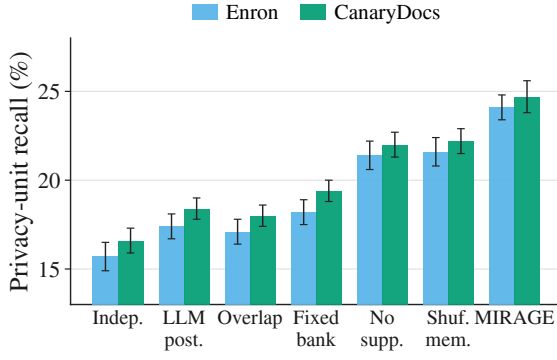
1 Introduction

Text embeddings are the default interface for vector search, retrieval-augmented generation (RAG), enterprise knowledge bases, and agent memory, and they are a privacy boundary: model-inversion and de-anonymization work shows that learned interfaces and side information can reveal private inputs (Fredrikson et al., 2015; Narayanan and

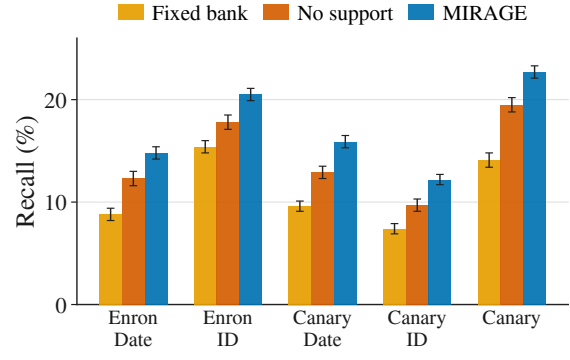
Shmatikov, 2008), while text-embedding attacks—reconstructing text from its embedding alone—recover words, phrases, named entities, and structured private information (Song and Raghunathan, 2020; Li et al., 2023; Morris et al., 2023; Huang et al., 2024; Chen et al., 2024b). Yet this work almost universally treats inversion as a single-vector problem. Real vector stores do not: vector databases pair dense vectors with metadata for hybrid filtering (Wang et al., 2021; Pan et al., 2024), and RAG/RALM pipelines keep provenance for grounding and source attribution (Tan et al., 2023; Ram et al., 2023). RAG caches, enterprise mail indexes, and document databases therefore often retain *ordered provenance metadata*—document identifiers, chunk offsets, and chunk order. When a backup, API log, or retrieval response leaks, the attacker often obtains (doc_id, chunk_idx, embedding) tuples together, not a shuffled bag of vectors.

Ordered provenance is an active privacy surface. Sensitive evidence is sequential: a name and email can be split across chunks; a date can disambiguate a later approval event; identifiers, URLs, and canary slots recur in structured positions. Document-level privacy risk therefore depends not only on independent chunk inversion, but on whether earlier recovered facts become encoder-verified constraints for later chunks.

We introduce MIRAGE, a two-level attack that operationalizes this insight beyond independent single-vector inversion. The local layer, MIRAGE-SC, trains a public-shadow candidate generator with budget-aligned span calibration, producing a strong single-chunk candidate bank. The document layer extracts typed anchors—entities, dates, identifiers, URLs, amounts, rare terms, and canaries—from earlier candidates, verifies them with counterfactual embedding-support checks, and feeds sup-



(a) Privacy-unit recall across deployable controls.



(b) Per-type recall (date, identifier, canary).

Figure 1: Main T-MC-Q results on the anchor configuration (e5-large-v2 + Qwen3-1.7B). Error bars are document-level bootstrap 95% CI half-widths.

ported anchors into memory-guided regeneration and document-level selection.

Figure 1 previews the core result: ordered provenance becomes a measurable structured leakage channel. Using the same public-shadow local bank as all compared methods and a strict $Q=512$ victim-time online encoder-query budget, MIRAGE raises parser-audited privacy-unit recall (PU-R) on the anchor configuration from 18.8 to 24.4, a paired gain of +5.65 [5.25, 6.16] points, recovering roughly 4.51k units versus 3.47k for fixed-bank reranking. Within the per-chunk controller interface in Section 3, LLM-postprocess, LLM-rerank, and no-support-check receive the same one-controller-call-per-chunk budget but do not close the gap. The lift persists across five encoders (+5.65–6.10 macro points) and four decoders (+5.2–6.9 points). Mechanism diagnostics attribute the gain to counterfactual encoder support and order-aware placement: 34.8% of newly correct units appear only in regenerated candidates absent from every fixed bank, and the effect transfers when BASEINVERT (our local inverter) is replaced by external zero-pair inverters.

Contributions. (i) We formulate ordered document-chunk leakage as a document-level inversion problem with zero victim plaintext pairs (T-MC-Q), and identify ordered provenance metadata as an under-studied privacy surface in RAG and vector stores. (ii) We introduce MIRAGE, which extracts typed anchors from earlier candidates, verifies them with counterfactual encoder support, and uses them to regenerate later chunks and resolve a document-level selection. (iii) We evaluate MIRAGE under controlled local banks within each comparison, a fixed victim-time encoder-query budget, and a ledgered per-chunk controller interface

across five encoders and four decoders, and additionally on two real, non-templated corpora, showing structured privacy-unit recall gains over strong fixed-bank reranking. With offline shadow costs and online controller calls explicitly accounted for, mechanism diagnostics attribute these gains to encoder-supported, order-aware anchor propagation.

2 Related Work

We position MIRAGE against four lines of work.

Embedding inversion with target-specific training. Model inversion and de-anonymization show that model interfaces and auxiliary structure can reveal private inputs (Fredrikson et al., 2015; Narayanan and Shmatikov, 2008). In text embeddings, work has progressed from lexical and attribute leakage (Song and Raghunathan, 2020) to generative sentence recovery (Li et al., 2023), Vec2Text-style paired training (Morris et al., 2023), multilingual inversion (Chen et al., 2024b), and conditional masked diffusion (Xiao, 2026). MIRAGE asks a different question: once local inversion is competent, how does ordered document context change the leakage surface?

Few-shot and zero-pair inversion. ALGEN aligns across models with few-shot leakage (Chen et al., 2025), TEIA studies transferable no-query inversion (Huang et al., 2024), and zero-shot inverters recover nontrivial text without victim pairs (Zhang et al., 2025a; Kim et al., 2026). MIRAGE also uses zero victim pairs, but separates local competence from document-level memory and shows that ordered evidence delivers structured recoveries beyond fixed-bank reranking; the memory layer

transfers when these inverters replace BASEINVERT (Appendix D.1).

RAG and vector-store metadata leakage.

Dense retrieval and vector databases store embeddings with document metadata, attributes, and hybrid-query support (Lewis et al., 2020; Karpukhin et al., 2020; Wang et al., 2021; Pan et al., 2024); RAG/RALM systems further use provenance for grounding and source attribution (Tan et al., 2023; Ram et al., 2023). Prior attacks leak retrieval datastores through retrieval-based LMs or black-box interaction (Huang et al., 2023; Zeng et al., 2024), extract RAG content via backdoors or feedback (Peng et al., 2024; Jiang et al., 2024), and exploit source transparency (Choi et al., 2025); surveys cover retrieval databases, access control, and answer leakage (Bodea et al., 2026; Zeng et al., 2025). MIRAGE identifies a complementary channel: document grouping and chunk order leaking with embeddings make ordered provenance itself exploitable.

Private representations and defenses. Prior defenses reduce representation leakage with adversarial or privacy-aware objectives (Coavoux et al., 2018), representation mixing (Huang et al., 2020), private/context-aware embeddings (Plant et al., 2021), and embedding-specific mutual-information optimization (Liu et al., 2026). MIRAGE shows that metadata minimization, order hiding, query control, and representation perturbation must be considered together when embeddings and provenance are exposed.

3 Threat Model and Protocols

Main problem (T-MC-Q). The attacker observes ordered chunk embeddings of a single document, $e_{1:K} = (e_1, \dots, e_K)$ with $e_i = E(x_i)$, where x_i is the document’s i -th chunk and E the target encoder, together with provenance metadata: encoder identity/version, document grouping, chunk order and approximate chunk size. This regime is common in vector-store backups, RAG caches, provenance-rich retrieval responses, and reranking logs. The attacker has no victim plaintext, victim text–embedding pairs, gradients, or closed weights, and trains no victim-pair inverter—the only encoder-specific component is the public-shadow adapter, trained offline on public non-victim documents (target-encoder-adapted, not victim-specific); they may use public

LMs, NER, regular expressions, public corpora, and public non-victim documents queried through the target encoder. All main online methods share a hard per-chunk victim-time encoder-query budget of $Q=512$; LLM-based methods additionally use one controller call per chunk, and final selection is always scored through the target encoder rather than by a global LLM. Public-shadow encoder calls and adapter training are reusable offline costs accounted separately. T-SC-Q is the single-chunk component of the same boundary.

Privacy units and PU-R. The output is a recovered document $\hat{x}_{1:K}$, scored by document-level privacy-unit recall

$$\text{PU-R} = \frac{\# \text{ correct units}}{\# \text{ gold units}},$$

plus type-specific recall (dates, identifiers, canaries) and token F1. Exact-recovery scoring is not new: exact word recovery is measured with precision and recall (Song and Raghunathan, 2020), and exact reconstruction of complete short inputs is reported by Morris et al. (2023). PU-R operationalizes this idea for ordered document-level inversion: gold units are extracted by the same typed parser used for anchors (NER spans, regex, template tags), normalized by type (case, whitespace, date format, URL path, email domain), and matched only under exact normalized match—no partial credit and no oracle selection. An auxiliary audit uses an LLM-Judge semantic-leakage probe.

Evaluation accepts one selected, length-constrained reconstruction per leaked chunk, not an unbounded list of guesses, so PU-R cannot be increased through enumeration. We interpret this parser-defined measure with token F1 and the LLM-Judge audit: it captures structured-unit recovery, not verbatim reconstruction of the entire document.

Scope. The attack itself requires three access conditions: (A1) ordered grouping leaks alongside the embeddings, (A2) the encoder is known/queryable, and (A3) public-shadow encoding is feasible. Two further conditions instantiate and measure the evaluation: (A4) chunk boundaries are approximately known, and (A5) sensitive formats are parser-visible. A4 is an evaluation convenience rather than an attack requirement: MIRAGE stays above the same-setting fixed-bank reranker under deliberate chunker mismatch, retaining 71–90% of its matched-boundary margin

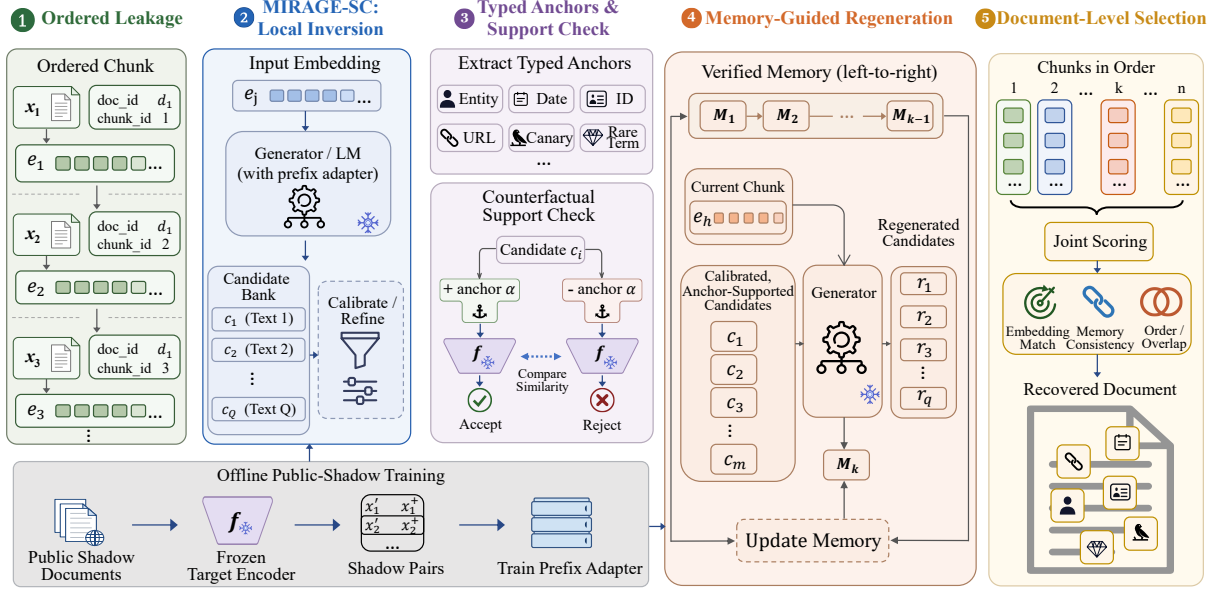


Figure 2: Overview of MIRAGE. Given an ordered leakage sequence $e_{1:K}$, MIRAGE-SC builds calibrated candidate banks, extracts encoder-verified typed anchors, and stores them in a left-to-right memory to guide subsequent regeneration. Final document selection combines embedding match, memory consistency, and order/overlap scores.

(Appendix G.4), though completely unknown segmentation remains outside that evidence. A5 determines which units the typed-anchor mechanism can identify and which units PU-R can measure; the non-parser-visible coverage gap is discussed in the limitations. Shuffled dumps, hidden-version proprietary APIs, sealed encoders, and unconstrained all-bank long-context controllers fall outside the main comparison. The latter expose a different interface: the controller sees every candidate bank and global document state at once, so we treat it as a separate attacker class rather than as a cost-matched baseline. MIRAGE uses zero victim plaintext pairs, an offline 80k-document public-shadow bank, and the $Q=512$ per-chunk online cap, distinguishing it from million-pair Vec2Text/CMDM training (Morris et al., 2023; Xiao, 2026) and few-shot ALGEN/TEIA transfer (Chen et al., 2025; Huang et al., 2024).

4 MIRAGE

Figure 2 presents MIRAGE as an ordered pipeline; we preview each step’s output here and detail it below. MIRAGE-SC first turns each leaked embedding e_i into a *candidate bank* of chunk reconstructions, calibrated by span-level edits. MIRAGE then extracts *typed anchors* (surface string, type, supporting chunks, confidence) from top candidates, and turns each into a *support decision*—accepted or banned—via counterfactual encoder

probes. Accepted anchors enter a left-to-right memory that conditions *regenerated text* for later chunks, and every candidate, banked or regenerated, is scored by target-encoder similarity, memory consistency, and order/overlap agreement before a final document-level selection. This separation keeps local inversion competence distinct from the document-level leakage created by ordered provenance.

4.1 Public-Shadow Local Base

For each e_i , the local inverter produces the first bank of candidate text reconstructions for chunk i , $C_i^0 = \text{BASEINVERT}(e_i, E, q_{\text{local}})$, ranked by $\cos(E(c), e_i)$. BASEINVERT is a frozen Qwen3-1.7B (Yang et al., 2025) prefix generator trained only on public-shadow chunks: a two-layer embedding-to-prefix adapter maps the target embedding into soft prefix tokens, and the backbone is trained with a next-token objective on public text paired with target-encoder embeddings. At test time the generator proposes text only. Every candidate is rescored through E , and final selection never uses generator likelihood (Appendix B.1).

MIRAGE-SC then applies span-level local calibration

$$C_i^{\text{rsc}} = \text{LOCALCALIBRATE}(e_i, C_i^0, E, q_{\text{cal}})$$

covering date normalization, numeric repair, short-span rewriting, entity-boundary correction, and

identifier editing. An edit is kept only when it increases target-embedding consistency. For single-chunk input, MIRAGE-SC returns the best calibrated candidate; for documents, the calibrated banks become the input to the memory pipeline.

4.2 Typed Cross-Chunk Memory

From the top- M local candidates, MIRAGE extracts *anchors*—typed facts carried forward from earlier chunks—recorded as $a = (s, t, \mathcal{I}, \rho)$, where s is a surface string, t is a type, $\mathcal{I}$ is the supporting-chunk set, and ρ is candidate support. Types include entities, dates, numbers, email/URL fragments, identifiers, rare n -grams, and canary strings.

Two channels populate memory. Repeated evidence covers entities, dates, URLs, and domain terms that recur across chunks or candidate variants; structural slots cover identifiers, account strings, case IDs, and canaries that may appear once but match strict templates. Thus canaries and identifiers need not repeat literally, but they require stronger type parsing and encoder support before propagation. The memory record stores type, support, and source-position fields rather than a free-form summary, so later regeneration can condition on specific checked facts. Type-specific extraction and support operations are described in [Appendix B](#).

4.3 Counterfactual Encoder Support

Surface plausibility alone is unsafe: hallucinated anchors contaminate later chunks. For each anchor a and candidate c_i , MIRAGE builds keep/insert and remove/replace counterfactuals,

$$\begin{aligned} c_i^{+a} &= \text{INSERTORKEEP}(c_i, a), \\ c_i^{-a} &= \text{REMOVEORREPLACE}(c_i, a), \end{aligned}$$

and measures

$$\Delta_i(a) = \cos(E(c_i^{+a}), e_i) - \cos(E(c_i^{-a}), e_i).$$

Repeated-evidence anchors must pass type-specific $\Delta_i(a)$ thresholds and cross-chunk support; structural-slot anchors may pass without repetition under stricter thresholds and parser constraints. Memory thus contains only encoder-supported typed anchors, not every plausible span. Anchors that fail support are recorded as banned inputs for regeneration, so later candidates are discouraged from copying unsupported private-looking strings.

4.4 Memory-Guided Regeneration

Regeneration is left-to-right: before processing chunk i , memory contains only anchors accepted from chunks $< i$:

$$C_i^1 = G(e_i, C_i^{\text{sc}}, \mathcal{M}_{< i}),$$

where G receives the embedding, calibrated local candidates, supported prior anchors, banned anchors, and length ranges. Fields are structured (surface, type, slot, support), not free-form summaries. Supported anchors therefore introduce new candidates rather than only reranking the fixed local bank; this is the path by which absent-bank units can be recovered.

Final document selection scores candidate x by

$$\begin{aligned} \text{score}_i(x) &= \alpha \cos(E(x), e_i) \\ &\quad + \beta m(x, \mathcal{M}_i) \\ &\quad + \gamma o(x, \hat{x}_{< i}, \tilde{C}_{> i}), \end{aligned}$$

with m measuring memory consistency and o penalizing overlap and order conflicts.

Inference schedule. MIRAGE first constructs calibrated local banks for every chunk, then scans left to right to accept supported anchors and regenerate candidates with $\mathcal{M}_{< i}$. A final document-level beam search resolves overlap across local and regenerated candidates. Regeneration remains causal: future chunks are not summarized into memory and future candidates are used only during final selection.

Fixed inference budget. All thresholds, weights, and candidate counts are frozen on a disjoint public development pool. [Figure 3](#) visualizes the per-chunk allocation: 224 calls on raw local proposals (top $K_0=64$), 32 on calibration, 128 on anchor support checks, 96 on regeneration proposals (top $K_1=24$), and 32 on final selection, totaling $Q=512$. The Qwen3-1.7B prefix adapter is trained offline on 80k public documents and frozen before victim-time inference. The no-support-check ablation keeps the same parser vocabulary, top- M candidates, and online cap, but skips verification and spends the 128-call slot on unchecked sampling. The per-step query ledger appears in [Table 8](#) ([Appendix](#)).

5 Experimental Setup

Data and protocols. We report both T-SC-Q and T-MC-Q under the same zero-victim-pair on-

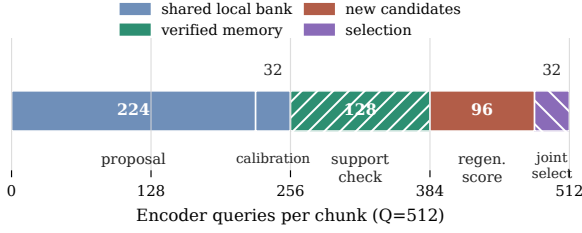


Figure 3: Per-chunk online encoder-query allocation for full MIRAGE under the fixed $Q=512$ budget. The support-check and regeneration segments are the only slots introducing verified memory and new candidates.

line family. T-SC-Q uses a held-out public single-chunk subset from the same public source pool, held out from the 80k shadow-training split. T-MC-Q uses Enron ordered documents (Klimt and Yang, 2004) and CanaryDocs, chunked into ordered 128-token blocks with 32-token overlap. Test sets contain 600 Enron documents (3,841 chunks, 9,619 normalized privacy units) and 600 CanaryDocs documents (3,564 chunks, 8,878 units across 24 template families). Shadow, development, and test documents are split by mailbox user, synthetic value pool, and seed; the 80k public-shadow pool is partitioned 60k/20k between Enron-style and CanaryDocs shadows. Beyond these, we evaluate on two real, non-templated corpora—the Ubuntu Dialogue Corpus (Lowe et al., 2015) and a customer-support corpus reconstructed from Twitter reply threads (Axelbrooke, 2017)—under the same ordered chunking protocol, with corpus-specific unit vocabularies (masked placeholders and brand handles excluded from the gold set); details in Appendix A.1. Full split card for Enron and CanaryDocs in Appendix A.

Encoders, decoders, and resources. MIRAGE is evaluated across five open dense encoders: e5-large-v2 (anchor, dim=1024) (Wang et al., 2022), bge-m3 (Chen et al., 2024a), jina-embeddings-v3 (Sturua et al., 2024), Qwen3-Embedding-0.6B, and Qwen3-Embedding-4B (Zhang et al., 2025b). For decoder scaling, the encoder is fixed to e5-large-v2 and we test Qwen3-1.7B, Qwen3-4B, Qwen3-8B (Yang et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Target encoders are queried only as a black box; victim-time online queries are capped at $Q=512$ per chunk. Public-shadow encoding and prefix-adaptor training are reusable offline costs. Model and configuration details in Appendix C.

Baselines and controls. For T-SC-Q we compare ZSInvert (Zhang et al., 2025a), Zero2Text (Kim et al., 2026), and MIRAGE-SC. For T-MC-Q, all main rows share the same MIRAGE-SC local bank and $Q=512$ cap; LLM-based rows use one controller call per chunk, deterministic rows none. These controls isolate explanations within this ledgered interface: *Independent* MIRAGE-SC isolates the local bank; *LLM postprocess* controls for generic cleanup; *LLM-rerank* for controller-cost-matched selection (controller restricted to picking the best existing candidate); *overlap stitching* and *fixed-bank reranking* for deterministic structure; *no-support-check* and *shuffled-memory* remove encoder support and ordered placement. Unconstrained long-context controllers that see all candidate banks at once are a separate attacker class under Section 3. For the cross-encoder sweep we rebuild the local bank per encoder (decoder fixed to Qwen3-1.7B); for the decoder sweep we hold the encoder fixed and rebuild only the regeneration controller (Appendix B).

Metrics. T-SC-Q reports BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), ChrF++ (Popović, 2017), and token F1. T-MC-Q reports PU-R, Date-R, ID-R, Canary-R, and token F1, with an LLM-Judge probe as an auxiliary audit. Main results use seeds {41, 42, 43}. Error bars and $\pm$ values are evaluation-unit bootstrap 95% CI half-widths for point estimates (chunks for T-SC-Q, documents for T-MC-Q); bracketed intervals report paired document-bootstrap 95% CIs for T-MC-Q method differences.

6 Results and Analysis

6.1 Local Base Check and Ceiling

Table 1 reports the T-SC-Q local-base check. MIRAGE-SC, trained on 80k public-shadow documents, provides a strong local base (Tok-F1 46.4 vs. ZSInvert 38.9 and Zero2Text 41.8 on the public subset). Later checks show that the document-level memory mechanism transfers to weaker external proposers (Table 4) and remains useful when the base is strengthened (Appendix G.2), suggesting that the document-level gains are not only repairs to a weak local base.

6.2 Ordered Multi-Chunk Recovery

Figure 1 reports the main T-MC-Q result on the anchor configuration (e5-large-v2 + Qwen3-

Table 1: Single-chunk T-SC-Q local-base check on the public subset.

Metric	ZSInvert	Zero2Text	MIRAGE-SC
BLEU-4	30.6 $\pm$ 0.8	35.9 $\pm$ 0.7	42.7 $\pm$ 0.6
ROUGE-L	55.2 $\pm$ 0.7	59.8 $\pm$ 0.6	64.9 $\pm$ 0.5
BERTScore	89.8 $\pm$ 0.2	91.1 $\pm$ 0.2	93.0 $\pm$ 0.1
ChrF++	34.9 $\pm$ 0.6	37.6 $\pm$ 0.5	42.1 $\pm$ 0.5
Tok-F1	38.9 $\pm$ 0.7	41.8 $\pm$ 0.6	46.4 $\pm$ 0.5

1.7B), with per-dataset values in Table 2. Full MIRAGE raises macro-averaged PU-R over fixed-bank reranking from 18.8 to 24.4, a paired gain of +5.65 [5.25, 6.16] points, recovering 4.51k normalized privacy units out of 18.5k versus 3.47k under the shared $Q=512$ cap. MIRAGE also retains a +2.7 [1.6, 3.5] gain over the no-support-check ablation with token F1 within 2 points.

Within the ledgered per-chunk interface, the controls separate the memory mechanism from simpler explanations. LLM postprocessing and overlap stitching provide only modest gains, so the improvement is not generic cleanup or overlap repair. LLM-rerank reaches 19.9 macro PU-R when the controller can only select existing candidates, below no-support-check at 21.7 and full MIRAGE at 24.4. The same-corpus margin persists beyond these datasets: on two real, non-templated corpora (UDC and a Twitter customer-support corpus), MIRAGE improves PU-R over fixed-bank reranking by +3.7 to +4.4 points in all four corpus–encoder pairs, with the support-check and regeneration markers moving in the same direction (Appendix F.3). External-base transfer, parser-independent audits, and robustness checks are summarized in later main-text analyses, with details left to the appendix.

For both corpora, the 80k-document adapter and candidate generator are reused without corpus-specific retraining; only held-out parser rules are specialized. The document-level margin therefore survives an out-of-domain local bank, separating it from template familiarity or an unusually strong in-domain proposer.

6.3 Cross-Encoder Generalization

We sweep five open dense encoders while fixing the decoder to Qwen3-1.7B. All five produce strong single-chunk local banks (Tok-F1 > 43; Appendix D.2), making weak-base repair alone an unlikely explanation for the document-level gains. Table 3 reports macro-averages over Enron and

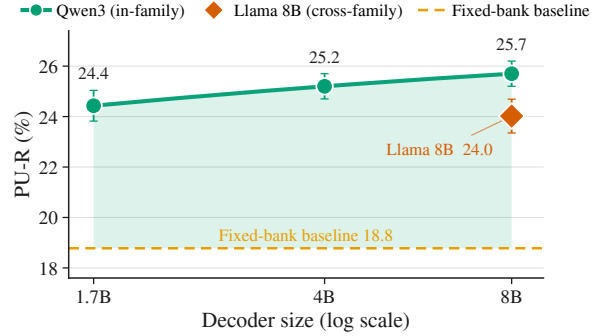


Figure 4: Decoder scaling with e5-large-v2 fixed. Shading marks the gain over fixed-bank (18.8); the Llama diamond tests cross-family transfer.

CanaryDocs. Macro gains range from +5.65 to +6.10 PU-R points, and the per-dataset cells span +5.35 to +6.31 across the 10 encoder×dataset pairs (Appendix F). ID-R gains span +4.80 to +5.55 macro points. Gains hold across distinct training paradigms: bge-m3 and jina-embeddings-v3 sit outside the e5 family yet show the same relative lift, so the effect generalizes beyond same-family encoder–decoder alignment. Per-dataset breakdowns and additional metrics (Canary-R and Tok-F1) are in Appendix F.

6.4 Decoder Scaling and Cross-Family Robustness

We fix the encoder to e5-large-v2 and sweep four instruction-tuned decoders spanning Qwen3 scale and Llama transfer. Fixed-bank reranking is decoder-independent (18.8 PU-R). Figure 4 shows that larger Qwen3 decoders raise absolute MIRAGE PU-R, and Llama-3.1-8B-Instruct still gives +5.15 points, so the memory-regeneration gain is not confined to a Qwen-family controller; per-dataset values are in Table 12.

Because the local candidate bank and fixed-bank comparator remain unchanged, this sweep isolates the regeneration controller. Qwen3 scaling adds only 0.8 and then 0.5 macro PU-R points; its total 1.3-point scale gain is much smaller than the +5.65 full-vs.-fixed gain at 1.7B. Thus capacity changes the attainable level but does not explain the ordered-memory benefit.

A pipeline-level Llama BASEINVERT check, a Qwen3-encoder + Llama-decoder check, and decoder-side no-support ablations are in Appendices G.2, F.2, and G.1.

Table 2: Per-dataset T-MC-Q results (anchor configuration, $Q=512$). The last three rows of each block generate new candidates; shuffled-memory and full additionally apply the support check. Dashes: Canary-R not applicable; $\pm$: document-bootstrap 95% CI half-width.

Dataset	Method	PU-R	Date-R	ID-R	Canary-R	Tok-F1
Enron	Independent MIRAGE-SC	15.7 $\pm$ 0.8	6.4 $\pm$ 0.5	15.4 $\pm$ 0.6	–	35.7 $\pm$ 0.8
	Fixed-bank rerank	18.2 $\pm$ 0.7	8.8 $\pm$ 0.6	15.4 $\pm$ 0.6	–	34.0 $\pm$ 0.8
	LLM-rerank	19.8 $\pm$ 0.7	9.9 $\pm$ 0.6	16.3 $\pm$ 0.6	–	34.6 $\pm$ 0.8
	MIRAGE no-support-check	21.4 $\pm$ 0.8	12.3 $\pm$ 0.7	17.8 $\pm$ 0.7	–	37.7 $\pm$ 0.7
	MIRAGE shuffled-memory	21.6 $\pm$ 0.8	12.0 $\pm$ 0.6	17.8 $\pm$ 0.7	–	37.7 $\pm$ 0.7
	MIRAGE full	24.1$\pm$0.7	14.8$\pm$0.6	20.5$\pm$0.6	–	39.3$\pm$0.6
CanaryDocs	Independent MIRAGE-SC	16.6 $\pm$ 0.7	7.2 $\pm$ 0.5	7.7 $\pm$ 0.5	13.9 $\pm$ 0.6	36.8 $\pm$ 0.7
	Fixed-bank rerank	19.4 $\pm$ 0.7	9.6 $\pm$ 0.5	7.4 $\pm$ 0.5	14.1 $\pm$ 0.7	34.4 $\pm$ 0.8
	LLM-rerank	20.0 $\pm$ 0.6	10.4 $\pm$ 0.5	8.1 $\pm$ 0.5	14.9 $\pm$ 0.7	34.8 $\pm$ 0.8
	MIRAGE no-support-check	22.0 $\pm$ 0.7	12.9 $\pm$ 0.6	9.7 $\pm$ 0.6	19.5 $\pm$ 0.7	38.4 $\pm$ 0.7
	MIRAGE shuffled-memory	22.2 $\pm$ 0.7	12.6 $\pm$ 0.6	9.7 $\pm$ 0.6	19.0 $\pm$ 0.7	38.5 $\pm$ 0.7
	MIRAGE full	24.7$\pm$0.9	15.9$\pm$0.6	12.2$\pm$0.5	22.7$\pm$0.6	40.2$\pm$0.6

Table 3: Cross-encoder generalization with Qwen3-1.7B fixed. Macro averages over Enron and CanaryDocs; bracketed intervals are paired document-bootstrap 95% CIs for Δ PU-R.

Encoder	Fix. PU-R	MIR. PU-R	Δ PU-R	MIR. ID-R
e5-large-v2	18.8	24.4	+5.65 [5.25,6.16]	16.3
bge-m3	17.3	23.0	+5.74 [5.24,6.21]	15.0
jina-v3	19.3	25.3	+5.95 [5.37,6.56]	17.5
Qwen3-Emb-0.6B	16.5	22.6	+6.10 [5.73,6.56]	13.7
Qwen3-Emb-4B	21.1	27.1	+5.94 [5.40,6.45]	18.3

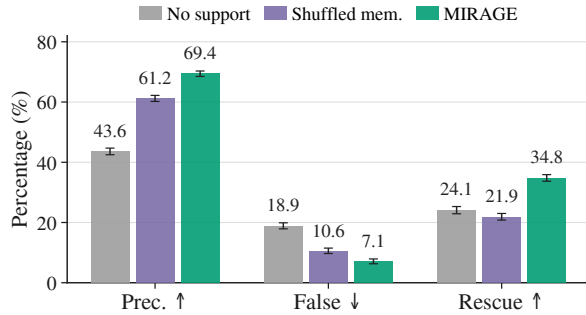


Figure 5: Mechanism diagnostics under the shared candidate pool and $Q=512$.

6.5 Mechanism Evidence

Figure 5 links the gain to support-checked memory rather than generic fluency. Support checks raise accepted-anchor precision from 43.6% to 69.4%, cut false-anchor rate from 18.9% to 7.1%, and recover 34.8% of newly correct units from candidates absent from every fixed bank. Shuffled memory keeps cleaner anchors but lowers rescue, showing that order-aware placement is load-bearing; relative to their own fixed-bank baselines the anchor-sensitive types gain most (dates +67%, canaries

Table 4: External base inverters with and without MIRAGE memory. Macro Enron+CanaryDocs; details in Appendix D.1.

Base inverter	+ Fixed	+ MIRAGE mem.	Δ_{mem}	Anchor prec.
ZSInvert	15.5	21.3	+5.8	61.0%
Zero2Text	14.2	20.3	+6.1	55.0%
BASEINVERT (ours)	18.8	24.4	+5.65	69.4%

+61%, identifiers +56%, rare terms +44%), with absolute points in Figure 6 and Appendix E.1 (cross-encoder diagnostics in Appendix H). A typical trace is an earlier chunk supporting an identifier that fixed-bank candidates omit later, but MIRAGE regenerates from memory in the appropriate slot.

The ablations support a staged interpretation: fixed-bank to no-support-check isolates generation beyond the local bank; no-support-check to full MIRAGE reflects anchor filtering; and shuffled-memory to full MIRAGE tests position-aware propagation. The absent-bank statistic is diagnostic because selection alone cannot recover a unit missing from every local bank.

Transfer across base inverters. Table 4 swaps only the local proposer for ZSInvert or Zero2Text (Zhang et al., 2025a; Kim et al., 2026). External bases gain +5.8 to +6.1 macro PU-R from MIRAGE memory, while full MIRAGE remains highest because BASEINVERT is the strongest local proposer.

6.6 Ablations and Robustness Checks

Table 5 shows that weakening the local base, removing support checks, shuffling memory, or

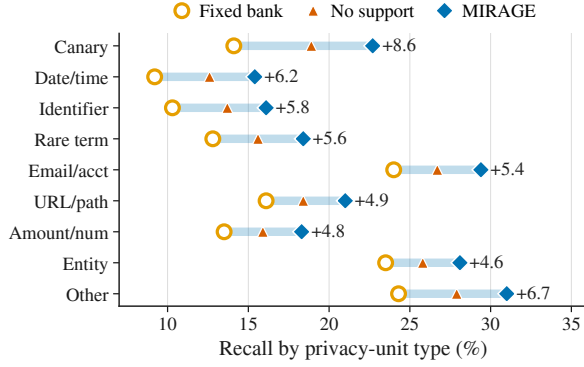


Figure 6: Per-type recall gains over fixed bank. Anchor-sensitive types gain most; full values are in Appendix E.1.

Table 5: Core ablations on the anchor configuration. Macro Enron+CanaryDocs except Canary-R.

Variant	T-SC-Q Tok-F1	PU-R	ID-R	Canary-R
MIRAGE full	46.4	24.4	16.3	22.7
Weaker SC base	42.0	21.3	13.4	19.4
No support check	46.4	21.7	13.7	19.5
Shuffled memory	46.4	21.9	13.8	19.0
Fixed bank	46.4	18.8	11.4	14.1

using fixed-bank reranking all reduce T-MC-Q performance; cross-encoder ablations are in Appendix G.1.

Five alternative explanations are checked within the stated interface. Gains persist under stronger local bases ($\Delta +7.5$ – 7.7 over the corresponding Independent MIRAGE-SC base), a Llama BASEINVERT (+5.6/+5.7), held-out CanaryDocs templates (Canary-R 18.6, $\Delta +4.8$), and $Q \in \{256, 512, 1024\}$ (+5.3/5.6/6.2); under deliberate boundary mismatch MIRAGE retains 71–90% of the matched-boundary margin and stays above the same-setting fixed-bank reranker (Appendices G.2–G.4). The parser-independent LLM-Judge audit moves in the same direction (Appendix E.3).

The budget sweep also argues against an unused-query explanation. Even with 896 proposal queries at $Q=2048$, fixed-bank reranking reaches 21.0 PU-R, below full MIRAGE at $Q=512$ (24.3 in the sweep). Thus the gain is not explained merely by additional local proposals and remains positive across $Q=64$ – 2048 .

6.7 Representative Defenses

We evaluate provenance-focused defenses on Enron and CanaryDocs, and two further stress tests on Enron only; these are stress tests rather than a full defense benchmark; the rate-limited row is a

Table 6: Representative defense stress tests (residual PU-R). Rate-limiting targets query access and PQ-8bit the embedding representation; the remaining rows act on provenance metadata. –: not evaluated on CanaryDocs.

Setting	Measured stress test	Enron	Canary Docs
No defense	full MIRAGE	24.1	24.7
Strip metadata + coarsen order	order-hidden	8.6	10.4
Adaptive attacker	dev-calibrated	11.3	12.8
Rate-limit $Q=128$	MIRAGE at $Q=128$	13.5	–
PQ-8bit quantization	fixed-bank	10.4	–

separate Enron-only defense measurement, distinct from the proportional budget sweep in Table 16. Table 6 reports residual PU-R for provenance metadata, query access, and embedding representation; the no-defense row uses per-dataset full-MIRAGE values rather than the macro headline (24.4). Stripping metadata and coarsening order transfers to CanaryDocs, and an adaptive attacker re-calibrated on held-out development data recovers only part of the gap (Appendix I).

Hiding grouping/order removes the channel in Section 3 and gives the largest drop (24.1→8.6). PQ-8bit quantization perturbs the local bank, reducing the fixed-bank comparator from 18.2 to 10.4, and $Q=128$ rate-limiting shrinks support-check/regeneration opportunities, but both leave residual leakage. Canary monitoring is useful but policy- and evasion-dependent rather than a residual-PU-R row.

7 Conclusion

We recast text embedding inversion as ordered document-level leakage and show that provenance metadata is an exploitable privacy channel. MIRAGE turns this channel into an attack: public-shadow local inversion proposes candidates, typed memory carries reusable evidence, counterfactual encoder support filters it, and regeneration recovers structured units absent from fixed banks. Under the ledgered per-chunk interface, gains persist across five encoders and four decoders, and additionally on two real, non-templated corpora. On the anchor configuration, 34.8% of newly correct privacy units are absent from every fixed bank. The implication is direct: document grouping, chunk order, and offsets should be protected alongside embeddings.

Limitations

Three scope conditions bound the present evidence. (i) *Deployment variants*. Evaluation covers open-weight dense encoders with public-shadow query access. Proprietary APIs, hybrid sparse–dense indexes, multi-vector storage, broader quantized deployments, and unordered leakage remain natural extensions. (ii) *Metric vocabulary*. PU-R audits a fixed type vocabulary shared with anchor extraction. Token F1 and an LLM-Judge leakage probe move with PU-R, but information not visible to the typed parser remains a coverage gap: relational facts, paraphrased secrets, inferred attributes, and confidential business logic are only partially reached by these audits. (iii) *Defenses*. The defense results are point-estimate stress tests rather than a defense benchmark (rate-limit and quantization rows are Enron-only): metadata stripping and order coarsening transfer across datasets, and an initial adaptive probe recovers part of the leakage (Appendix I), but defenses evaluated jointly with utility tasks, and a full adaptive, multi-defense evaluation, remain future work. Taken together, these limits delineate the deployment conditions under which the channel is most exploitable.

Ethics Statement

This work studies privacy risks from leaked embeddings and provenance metadata under a controlled offline evaluation. We did not probe, attack, or collect data from any deployed RAG system, vector database, API, service, or user account. All experiments use only public datasets and a synthetic dataset constructed by us for measurement; the synthetic canaries are artificial strings rather than real secrets.

The goal is defensive: to show that document grouping, chunk order, and offsets can increase reconstruction risk when exposed alongside embeddings. The harms at stake are concrete: if identifiers, correspondence, or private records are reconstructed from leaked stores, affected individuals and organizations face exposure of personal and confidential information. We report aggregate reconstruction metrics and representative stress tests to help practitioners reason about metadata minimization, access control, query monitoring, and canary-based auditing; Appendix I.1 expands these into an operational checklist for RAG and vector-store operators.

To reduce dual-use risk, released artifacts should

support auditing rather than direct deployment of the attack. We will release the synthetic Canary-Docs benchmark¹ together with the evaluation and defense harnesses, split manifests, parser rules, prompts, and query ledgers needed to reproduce the reported measurements.²

Acknowledgments

This work was supported by the Project of the Basic and Frontier Research Project of PCL under Grant No. 2025QYA001, the Open Project of National Engineering Laboratory for Technology of Internet Domain Name under Grant No. KF202512, the Guangdong S&T Programme under Grant No. 2024A0101010001, and the Shenzhen R&D Program under Grant No. KJZD20230923114059020. It was also supported in part by the Shenzhen Key Laboratory of Software Defined Networking under Grant No. ZDSYS20140509172959989.

References

- Stuart Axelbrooke. 2017. [Customer support on Twitter](#). Kaggle. Published on Kaggle under the Thought Vector organization; dataset metadata lists Stuart Axelbrooke as creator. Accessed August 2026.
- Andreea-Elena Bodea, Stephen Meisenbacher, Alexandra Klymenko, and Florian Matthes. 2026. [SoK: Privacy risks and mitigations in retrieval-augmented generation systems](#). *Preprint*, arXiv:2601.03979.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. [The secret sharer: Evaluating and testing unintended memorization in neural networks](#). In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, USA. USENIX Association.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. [M3-embedding: Multi-linguality, 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. Association for Computational Linguistics.
- Yiyi Chen, Heather Lent, and Johannes Bjerva. 2024b. [Text embedding inversion security for multilingual language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7808–7827, Bangkok, Thailand. Association for Computational Linguistics.

¹<https://huggingface.co/datasets/LevenKoko/MIRAGE-CanaryDocs>

²<https://github.com/LevenKoko/MIRAGE>

- Yiyi Chen, Qionghai Xu, and Johannes Bjerva. 2025. [ALGEN: Few-shot inversion attacks on textual embeddings via cross-model alignment and generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24330–24348, Vienna, Austria. Association for Computational Linguistics.
- Chanwoo Choi, Jinsoo Kim, Sukmin Cho, Soyeong Jeong, and Buru Chang. 2025. [The RAG paradox: A black-box attack exploiting unintentional vulnerabilities in retrieval-augmented generation systems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23723–23744, Suzhou, China. Association for Computational Linguistics.
- Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. [Privacy-preserving neural representations of text](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.
- Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. [Model inversion attacks that exploit confidence information and basic countermeasures](#). In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, New York, NY, USA. Association for Computing Machinery.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, et al. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. 2023. [Privacy implications of retrieval-based language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14887–14902, Singapore. Association for Computational Linguistics.
- Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. 2020. [TextHide: Tackling data privacy in language understanding tasks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1368–1382, Online. Association for Computational Linguistics.
- Yu-Hsiang Huang, Yuche Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. 2024. [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, Bangkok, Thailand. Association for Computational Linguistics.
- Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, Yang Chen, and Min Yang. 2024. [Feedback-guided extraction of knowledge base from retrieval-augmented LLM applications](#). *Preprint*, arXiv:2411.14110.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Doohyun Kim, Donghwa Kang, Kyungjae Lee, Hyeongbo Baek, and Brent Byunghoon Kang. 2026. [Zero2text: Zero-training cross-domain inversion attacks on textual embeddings](#). *Preprint*, arXiv:2602.01757.
- Bryan Klimt and Yiming Yang. 2004. [The Enron corpus: A new dataset for email classification research](#). In *Machine Learning: ECML 2004*, pages 217–226. Springer.
- Patrick 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. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Haoran Li, Mingshi Xu, and Yangqiu Song. 2023. [Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040, Toronto, Canada. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tiantian Liu, Hongwei Yao, Feng Lin, Tong Wu, Zhan Qin, and Kui Ren. 2026. [Eguard: Defending LLM embeddings against inversion attacks via text mutual information optimization](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35644–35652.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. [The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems](#). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.
- John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. [Text embeddings reveal \(almost\) as much as text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

- Arvind Narayanan and Vitaly Shmatikov. 2008. [Robust de-anonymization of large sparse datasets](#). In *2008 IEEE Symposium on Security and Privacy*, pages 111–125. IEEE Computer Society.
- James Jie Pan, Jianguo Wang, and Guoliang Li. 2024. [Survey of vector database management systems](#). *The VLDB Journal*, 33(5):1591–1615.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Yuefeng Peng, Junda Wang, Hong Yu, and Amir Houmansadr. 2024. [Data extraction attacks in retrieval-augmented generation via backdoors](#). *Preprint*, arXiv:2411.01705.
- Richard Plant, Dimitra Gkatzia, and Valerio Giuffrida. 2021. [CAPE: Context-aware private embeddings for private language learning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7970–7978, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: Words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlga, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. [In-context retrieval-augmented language models](#). *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- Congzheng Song and Ananth Raghunathan. 2020. [Information leakage in embedding models](#). In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 377–390, New York, NY, USA. Association for Computing Machinery.
- Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. [jina-embeddings-v3: Multilingual embeddings with task LoRA](#). *Preprint*, arXiv:2409.10173.
- Wang-Chiew Tan, Yuliang Li, Pedro Rodriguez, Richard James, Xi Victoria Lin, Alon Halevy, and Scott Yih. 2023. [Reimagining retrieval augmented language models for answering queries](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6131–6146, Toronto, Canada. Association for Computational Linguistics.
- Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, and 3 others. 2021. [Milvus: A purpose-built vector data management system](#). In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627, New York, NY, USA. Association for Computing Machinery.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *Preprint*, arXiv:2212.03533.
- Han Xiao. 2026. [Embedding inversion via conditional masked diffusion language models](#). *Preprint*, arXiv:2602.11047.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. 2025. [Mitigating the privacy issues in retrieval-augmented generation \(RAG\) via pure synthetic data](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24527–24558, Suzhou, China. Association for Computational Linguistics.
- Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. 2024. [The good and the bad: Exploring privacy issues in retrieval-augmented generation \(RAG\)](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4505–4524, Bangkok, Thailand. Association for Computational Linguistics.
- Collin Zhang, John X. Morris, and Vitaly Shmatikov. 2025a. [Universal zero-shot embedding inversion](#). *Preprint*, arXiv:2504.00147.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*. OpenReview.net.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 Embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.

A Dataset, Splits, and Auditability

Default appendix setup. Unless otherwise stated, every appendix experiment uses the same default configuration: per-chunk online encoder-query budget $Q=512$; anchor encoder e5-large-v2; online regeneration controller / BASEINVERT backbone Qwen3-1.7B (frozen prefix generator trained on 80k public-shadow documents); seeds {41, 42, 43}; parser-audited normalized type matching; evaluation-unit bootstrap 95% CI half-widths denoted by $\pm$ (chunks for T-SC-Q checks, documents for T-MC-Q experiments). The real-corpus and boundary-mismatch tables report point estimates without bootstrap intervals. Each subsection only states differences from this default. Cross-references that span subsections retain the labels referenced from the main paper.

A.1 Data, splits, and protocols

Table 7: Data split and protocol summary. The public-shadow training source contains 80k documents; the public single-chunk development/test subsets contain 1k/2k held-out chunks and are used only for the T-SC-Q local-base check.

Dataset	Unit	Train/dev/test	Test chunks	Gold PUs
Public shadow / SC	doc/chunk	80k/1k/2k	2k	–
Enron docs	doc	60k/100/600	3,841	9,619
CanaryDocs	doc	20k/100/600	3,564	8,878

Chunking: public 128 tok/no ovlp.; Enron 128 tok/32 ovlp., 6.4 ch/doc.
CanaryDocs: 128 tok/32 ovlp., 5.9 ch/doc; 24 templates.

CanaryDocs construction. CanaryDocs follows canary-based leakage measurement practice (Carlini et al., 2019) but uses synthetic document templates rather than training-data memorization prompts. The 24 template families cover account, calendar, approval, URL/path, invoice, and project-code slots (example safe values: ACCT-7Q3-LIME, CASE-42-BLUE, /vault/april/route-19); canaries are inserted at typed slots in otherwise natural synthetic documents with *disjoint value pools* across public-shadow, development, and test splits, so Canary-R measures recovery of disjoint secret values under known or related slot schemas, not arbitrary unseen secret formats. The test set contains 1,181 normalized canary units across 600 documents (about 1.97 units per document), matched by the same exact normalized type rule as other privacy units. Residual template memorization is probed by holding 12 template families out of shadow training (Table 15).

Real-corpus setup. Two real, non-templated corpora supplement the main benchmark: the Ubuntu Dialogue Corpus (UDC) (Lowe et al., 2015), a corpus of technical-support dialogues, and a customer-support corpus reconstructed from Twitter reply threads (TWCS) (Axelbrooke, 2017). One dialogue thread forms one document; turns are concatenated in temporal order and chunked with the same ordered 128-token/32-overlap protocol, with chunk order following the timeline. Splits are held out by dialogue thread and user. The 80k-document public-shadow adapter is reused without retraining, so the local bank is out-of-domain by construction and the quantity of interest is the same-corpus margin over fixed-bank reranking rather than absolute levels. Privacy units follow the same parser framework and normalization protocol as the main evaluation, with corpus-specific rules frozen on held-out development data: for UDC, paths, versions, package names, URLs, hostnames, usernames, and error codes; for TWCS, anonymized user handles/IDs, URLs, date expressions, and names, where emails are masked as `__email__` and masked placeholders and public brand handles are excluded from the gold set.

A.2 Auditability and scope boundaries

To support audit and reproduction we freeze, before any test-time evaluation, the split manifest (document IDs, canary value pools, seeds), chunk manifest (boundaries, overlap, order), public-shadow ID list, adapter checkpoint hashes, candidate-bank JSONL outputs of MIRAGE-SC, the typed anchor-parser rules, prompt skeleton, and the per-seed run/query ledger that verifies the $Q=512$ cap and bootstrap intervals. Split separation is enforced by mailbox user, synthetic value pool, and seed; within each reported comparison/configuration, methods share the same MIRAGE-SC local bank and equal- Q online budget, and LLM-based controls share the one-controller-call-per-chunk interface. Any artifact release will follow the licenses and terms of the original datasets and models: we will not redistribute sensitive Enron plaintext, complete URL paths, full synthetic canary value lists, or other non-redistributable content; released materials will be limited to synthetic or redacted data, scripts, manifests, hashes, and evaluation code consistent with the applicable source terms. Scope conditions (A1–A5) are in Section 3; release commitments are summarized in the Ethics Statement.

B Inference Procedure and Baseline Configurations

Table 8: Fixed MIRAGE inference ledger ($Q=512$).

#	Operation	Q	Output
1	Local proposal	224	Top- $K_0=64$ unique BASEINVERT candidates
2	Local calibration	32	Date/number/boundary/ID edits; keep only similarity-improving
3	Anchor extraction	0	Typed anchors from top- $M=16$ (offline parse)
4	Counterfactual support	128	Keep/insert and remove/replace probes for $\Delta_i(a)$
5	Memory-conditioned regeneration	96	Top- $K_1=24$ regenerated candidates
6	Joint selection	32	Beam search (size 8) over similarity, memory match, overlap

The LLM/controller only proposes candidates; final scoring is always done by the target encoder. Public dev pool: 20 Enron-style public mails, 20 synthetic canary docs, 10 procedural docs. Frozen thresholds: $\tau_{\text{rep}}=0.018$, $\tau_{\text{slot}}=0.031$, anchors/chunk ≤ 12 , joint weights $\alpha:\beta:\gamma=1:0.28:0.12$. Offline cost: 80k public shadow documents encoded once; Qwen3-1.7B prefix adapter trained offline and frozen at victim time.

B.1 Local generator and BASEINVERT configuration

BASEINVERT implementation. The shared MIRAGE-SC local bank uses a frozen Qwen3-1.7B prefix generator with a two-layer embedding-to-prefix adapter (32 soft tokens), trained on chunks derived from 80k public-shadow documents and paired with target-encoder embeddings, with no victim text. At inference it proposes 224 candidates per chunk; candidates are target-encoder scored, deduplicated, and truncated to top $K_0=64$. Fixed calibration edits for dates, numbers, entity boundaries, and identifiers are retained only when similarity improves. Optimizer, schedule, seed, checkpoint hash, hardware, wall-clock time, split IDs, parser rules, prompt skeleton, candidate banks, and bootstrap scripts are frozen before test-time evaluation (see [Appendix A.2](#)).

B.2 Baseline grouping and query ledger

Baselines. All deployable T-MC-Q controls share the MIRAGE-SC local bank and the same $Q=512$ online encoder-query cap; external single-chunk baselines share the same per-chunk budget but no public-shadow generator.

Single-chunk external baselines (T-SC-Q): **ZS-Invert** performs zero-shadow online inversion under a 512-query budget for local proposal scoring/refinement; **Zero2Text** is a zero-training cross-domain inverter under the same cap and no victim pairs.

Deployable T-MC-Q controls: **Independent MIRAGE-SC** takes per-chunk top-1 from the calibrated local bank (no document-level memory). **LLM postprocess** consumes the top local candidates and local metadata via one controller call per chunk with no support-checked memory and the same regeneration budget as MIRAGE. **Overlap stitching** is a deterministic beam over fixed local candidates with overlap rewards/penalties and no new candidates. **Fixed-bank rerank** runs a document-level beam over the fixed bank using encoder similarity plus overlap with no regeneration. **LLM-rerank** uses one controller call per chunk to pick an index in the fixed bank (no regeneration, no memory); parse failures fall back to encoder top-1.

Mechanism ablations: **No support check** reuses the parser vocabulary, top- M candidates, and 12-anchor cap of full MIRAGE but accepts anchors without verification, spending the same 128-call slot on unchecked anchor variants. **Shuffled memory** keeps support-checked anchors but permutes their positions before regeneration.

B.3 Typed memory operations

Typed memory operations. Typed memory operations share the parse-then-support recipe in [Section 4](#): candidate anchors are accepted only when delete, replace, or near-neighbor probes favor the original form under the target encoder. Entity anchors use repeated NER spans; date/time and numeric anchors use regex parses plus value/unit edits; identifier, rare-term, and canary anchors use finite-state templates or rare n -grams with minimal-edit probes. Accepted anchors carry type, support, and source-position fields into regeneration; rejected anchors are banned from later chunks.

C Models, Prompts, and Reproducibility Details

Encoders. All five encoders run in dense retrieval mode with L2-normalized embeddings, and every evaluated chunk is ≤ 128 tokens. We report native dimensions because dimensionality is controlled later: e5-large-v2 (1024-d) is the anchor encoder; bge-m3 and jina-embeddings-v3 provide non-e5

1024-d families; Qwen3-Embedding-0.6B (1024-d) and Qwen3-Embedding-4B (2560-d) cover Qwen3 scale, with a PCA-1024 control for the 4B model in [Appendix F](#).

Decoders. Qwen3-1.7B is the BASEINVERT backbone and main regeneration controller; Qwen3-4B, Qwen3-8B, and Llama-3.1-8B-Instruct are used only for decoder scaling. The regeneration controller produces up to 96 raw regenerated proposals via diverse beam search; candidates are target-encoder scored, deduplicated, and the top $K_1=24$ are retained. Ranking and selection are deterministic given seeds $\{41, 42, 43\}$. Frozen candidate sizes ($K_0=64$, $K_1=24$, $M=16$), beam size 8 (joint selection), support thresholds, and query-budget allocation are fixed before test time.

Regeneration/postprocess controller prompt schema. LLM postprocess, no-support-check, shuffled-memory, and full MIRAGE share one frozen JSON-style schema; variants differ only in accepted-memory and banned-anchor slots. The schema contains the chunk index, top-16 local candidates, accepted/rejected anchors, length/style constraints, and an output-only instruction. LLM-rerank is separate because it selects an existing candidate index rather than regenerating text.

D External Baselines and Single-Chunk Prerequisite

D.1 External zero-pair baselines on TMCQ

Table 9 extends the TMCQ comparison to two strong external zero-victim-pair inverters from prior work, ZSInvert ([Zhang et al., 2025a](#)) and Zero2Text ([Kim et al., 2026](#)), on the anchor configuration. Each external method is applied per chunk under the same $Q=512$ victim-time online encoder-query cap; iterative refinement steps that re-query the target encoder count toward this cap. The “+ fixed-bank” row wraps the per-chunk reconstruction with shared-bank rerank; “+ MIRAGE memory” replaces only the local proposer in MIRAGE with the external inverter while keeping support-checked anchor memory, counterfactual support, and order-aware regeneration. Intermediate “+ overlap stitching” rows track the “+ fixed-bank” row within bootstrap noise, as deterministic stitching offers no new candidates. The LLM-Judge Leakage column uses GPT-4.1 (gpt-4.1, temperature 0) with the same prompt as [Zhang et al. \(2025a\)](#); absolute numbers are not a direct repro-

duction of the unchunked Enron setting in that paper because we operate under chunked TMCQ, the $Q=512$ per-chunk cap, ordered splits, and parser-audited privacy units. The anchor-quality side of this comparison (accepted-anchor precision, false-anchor rate) is reported in the inline “Mechanism diagnostic across base inverters” paragraph below. PU-R / Date-R / ID-R / Canary-R / Tok-F1 in the bottom two rows are reproduced from [Table 2](#); the LLM-Judge column is newly measured.

Mechanism diagnostic across base inverters.

All variants use the support-checked memory layer; **accepted-anchor precision** (fraction of accepted propagated anchors matching a gold privacy unit of the same type, macro across Enron and Canary-Docs) is $55\pm1.2\%$ for Zero2Text + MIRAGE memory, $61\pm1.1\%$ for ZSInvert + MIRAGE memory, and $69.4\pm0.9\%$ for full MIRAGE with its own BASEINVERT; the corresponding **false-anchor rates** are $14.3\pm1.0\%$, $11.5\pm1.0\%$, and $7.1\pm0.8\%$. Replacing BASEINVERT with a noisier external inverter lowers anchor quality even when the support check is active—BASEINVERT and the support check operate on complementary axes (candidate quality vs. acceptance gating).

D.2 Single-chunk prerequisite across encoders

For each encoder in the generalization sweep, we check that the local MIRAGE-SC bank is strong enough that document-level gains are unlikely to be only repairs to a weak base. Per-encoder T-SC-Q Tok-F1 (same protocol as [Table 1](#); 80k public-shadow docs, frozen Qwen3-1.7B backbone, $Q=512$ per chunk; adapter re-trained per encoder) is 46.4 ± 0.5 for e5-large-v2, 45.1 ± 0.6 for bge-m3, 47.3 ± 0.4 for jina-embeddings-v3, 43.8 ± 0.7 for Qwen3-Emb-0.6B, and 48.9 ± 0.4 for Qwen3-Emb-4B; corresponding BLEU-4/ROUGE-L/BERTScore/ChrF++ values follow the same rank order with CI half-widths ≤ 0.8 throughout. All five encoders exceed Tok-F1 43, and the rank Qwen3-Emb-4B > jina-v3 > e5 > bge-m3 > Qwen3-Emb-0.6B is close to the fixed-bank PU-R order in [Table 3](#).

E Typed Breakdown, Type-Wise Precision, and Auxiliary Metrics

E.1 Typed breakdown of recall

The typed breakdown figure is in the main text ([Figure 6](#)). In absolute points the largest full-minus-fixed gains are canary strings (+8.6) and

Table 9: External zero-pair baselines on Enron and CanaryDocs ($Q=512$).

Dataset	Method	PU-R	Date-R	ID-R	Canary-R	Tok-F1	LLM-Judge
Enron	ZSInvert raw	12.8 $\pm$ 0.8	4.6 $\pm$ 0.5	5.1 $\pm$ 0.6	–	30.4 $\pm$ 0.8	68 $\pm$ 2
	+ fixed-bank rerank	15.2 $\pm$ 0.7	6.0 $\pm$ 0.5	6.8 $\pm$ 0.6	–	31.0 $\pm$ 0.7	73 $\pm$ 2
	+ MIRAGE memory	21.0 $\pm$ 0.7	10.5 $\pm$ 0.6	12.5 $\pm$ 0.6	–	34.8 $\pm$ 0.6	79 $\pm$ 2
	Zero2Text raw	11.5 $\pm$ 0.8	4.0 $\pm$ 0.5	4.4 $\pm$ 0.6	–	28.6 $\pm$ 0.9	65 $\pm$ 2
	+ fixed-bank rerank	13.9 $\pm$ 0.7	5.1 $\pm$ 0.5	5.8 $\pm$ 0.6	–	29.4 $\pm$ 0.8	70 $\pm$ 2
	+ MIRAGE memory	20.0 $\pm$ 0.7	9.8 $\pm$ 0.6	11.5 $\pm$ 0.6	–	33.8 $\pm$ 0.6	76 $\pm$ 2
	Independent MIRAGE-SC	15.7 $\pm$ 0.8	6.4 $\pm$ 0.5	15.4 $\pm$ 0.6	–	35.7 $\pm$ 0.8	75 $\pm$ 2
	MIRAGE full	24.1$\pm$0.7	14.8$\pm$0.6	20.5$\pm$0.6	–	39.3$\pm$0.6	88$\pm$2
CanaryDocs	ZSInvert raw	13.4 $\pm$ 0.7	5.0 $\pm$ 0.5	5.5 $\pm$ 0.5	8.2 $\pm$ 0.7	31.0 $\pm$ 0.8	67 $\pm$ 2
	+ fixed-bank rerank	15.8 $\pm$ 0.6	6.0 $\pm$ 0.5	6.7 $\pm$ 0.5	10.0 $\pm$ 0.7	31.6 $\pm$ 0.7	72 $\pm$ 2
	+ MIRAGE memory	21.5 $\pm$ 0.6	10.9 $\pm$ 0.6	12.8 $\pm$ 0.5	17.0 $\pm$ 0.6	35.5 $\pm$ 0.6	80 $\pm$ 2
	Zero2Text raw	12.2 $\pm$ 0.7	4.4 $\pm$ 0.5	4.8 $\pm$ 0.5	7.5 $\pm$ 0.7	29.3 $\pm$ 0.8	64 $\pm$ 2
	+ fixed-bank rerank	14.5 $\pm$ 0.7	5.4 $\pm$ 0.5	6.0 $\pm$ 0.5	9.4 $\pm$ 0.7	30.0 $\pm$ 0.7	69 $\pm$ 2
	+ MIRAGE memory	20.5 $\pm$ 0.7	10.0 $\pm$ 0.6	11.8 $\pm$ 0.5	15.5 $\pm$ 0.6	34.5 $\pm$ 0.6	76 $\pm$ 2
	Independent MIRAGE-SC	16.6 $\pm$ 0.7	7.2 $\pm$ 0.5	7.7 $\pm$ 0.5	13.9 $\pm$ 0.6	36.8 $\pm$ 0.7	75 $\pm$ 2
	MIRAGE full	24.7$\pm$0.9	15.9$\pm$0.6	12.2$\pm$0.5	22.7$\pm$0.6	40.2$\pm$0.6	89$\pm$2

the *Other/non-anchorable* residual (+6.7), ahead of dates (+6.2), identifiers (+5.8), and rare domain terms (+5.6). The residual ranks second in absolute terms because it also starts from the highest fixed-bank recall (24.3, versus 9.2–14.1 for dates, identifiers, and canaries); measured against their own fixed-bank baselines, the anchor-sensitive types gain most (dates +67%, canaries +61%, identifiers +56%, rare domain terms +44%), while the non-anchorable residual gains +28% and ranks seventh of nine types. Memory-guided regeneration therefore improves generic prose as well—consistent with the token-F1 movement in Table 2—but the categories that cross-chunk anchor binding targets are the ones that improve most in proportion to what fixed-bank reranking already recovers. Per-type recall values are tabulated together with type-wise precision in Table 10 below.

E.2 Type-wise precision calibration

Table 10 reports the precision side of the support-weighted type breakdown shown in Figure 6. The rows use the same fixed parser vocabulary as PU-R and therefore remain parser-audited rather than human-annotated. Type-wise precision under full MIRAGE parallels the per-type recall pattern (canaries, identifiers, and dates show the largest precision gains over fixed-bank and over no-support-check), providing a check against parser-specific recall inflation and indicating that the support-

checked memory layer improves both axes per type rather than trading one for the other.

E.3 Auxiliary audit metrics

The LLM-Judge semantic-leakage probe (Zhang et al., 2025a) rises alongside PU-R and Tok-F1 when MIRAGE memory replaces fixed-bank reranking for both ZSInvert and Zero2Text on Enron and CanaryDocs; with MIRAGE-SC as the proposer, full MIRAGE likewise exceeds Independent MIRAGE-SC on all three measures in Table 9. Together, LLM-Judge and Tok-F1 provide model-judged and token-overlap checks independent of the typed parser.

F Generalization Across Encoders, Decoders, and Corpora

F.1 Per-dataset cross-encoder generalization

Table 11 reports per-dataset PU-R, ID-R, Canary-R (CanaryDocs only), and Tok-F1 for each of the five encoders, with Fixed-bank and full MIRAGE paired so the encoder-conditional Δ PU-R is read directly. All rows use the fixed Qwen3-1.7B decoder, identical chunking, candidate counts, query budget, and seeds. Canary-R is reported only on CanaryDocs (synthetic slots; “–” for Enron indicates not applicable).

Embedding-dimension control for Qwen3-Embedding-4B. Qwen3-Embedding-4B

Table 10: Type-wise recall/precision calibration.

Privacy-unit type	Fixed-R	Fixed-P	NoSupp-P	MIRAGE-R	MIRAGE-P
Entity mention	23.5	27.2	28.6	28.1	31.4
Email/account	24.0	28.1	29.0	29.4	33.1
Date/time	9.2	12.4	15.3	15.4	18.9
Amount/numeric	13.5	16.2	18.0	18.3	20.8
URL/path token	16.1	18.8	20.1	21.0	24.2
Identifier/case ID	10.3	13.1	16.0	16.1	19.5
Canary string	14.1	17.5	22.0	22.7	27.8
Rare domain n -gram	12.8	15.5	18.2	18.4	21.7
Other/non-anchorable	24.3	25.8	29.3	31.0	32.7

Table 11: Per-dataset cross-encoder generalization ($Q=512$).

Dataset	Encoder	Fixed PU-R	MIRAGE PU-R	Δ PU-R	Δ ID-R	MIRAGE Canary-R	Tok-F1
Enron	e5-large-v2	18.18 $\pm$ 0.86	24.13 $\pm$ 0.77	+5.95	+5.07	–	39.26 $\pm$ 0.64
	bge-m3	16.92 $\pm$ 0.92	22.54 $\pm$ 0.69	+5.62	+6.58	–	38.43 $\pm$ 0.43
	jina-embeddings-v3	19.04 $\pm$ 0.90	25.00 $\pm$ 0.79	+5.96	+5.25	–	40.02 $\pm$ 0.45
	Qwen3-Emb-0.6B	16.01 $\pm$ 0.72	21.91 $\pm$ 0.75	+5.90	+5.18	–	37.41 $\pm$ 0.52
	Qwen3-Emb-4B	20.84 $\pm$ 0.77	26.53 $\pm$ 0.88	+5.69	+5.28	–	41.16 $\pm$ 0.55
CanaryDocs	e5-large-v2	19.38 $\pm$ 0.71	24.73 $\pm$ 0.94	+5.35	+4.74	22.67 $\pm$ 0.74	40.08 $\pm$ 0.53
	bge-m3	17.58 $\pm$ 0.74	23.45 $\pm$ 0.76	+5.87	+4.51	21.37 $\pm$ 0.67	39.62 $\pm$ 0.69
	jina-embeddings-v3	19.65 $\pm$ 1.01	25.58 $\pm$ 0.73	+5.93	+5.19	23.49 $\pm$ 0.69	41.00 $\pm$ 0.51
	Qwen3-Emb-0.6B	16.94 $\pm$ 0.82	23.25 $\pm$ 0.82	+6.31	+5.92	20.32 $\pm$ 0.59	38.66 $\pm$ 0.49
	Qwen3-Emb-4B	21.42 $\pm$ 0.98	27.61 $\pm$ 0.89	+6.19	+4.31	23.98 $\pm$ 0.73	42.08 $\pm$ 0.47

produces 2560-dimensional embeddings natively, unlike the other four 1024-d encoders. To check for a pure-dimensionality artifact, we PCA-project Qwen3-Embedding-4B to 1024-d and rerun the anchor-configuration inversion. The PCA-1024 variant attains Fixed-bank PU-R 19.32 $\pm$ 0.70 and full MIRAGE PU-R 25.10 $\pm$ 0.71 (macro), vs. native-2560 Fixed 21.13 / full 27.07 and e5-large-v2 (native 1024-d) Fixed 18.78 / full 24.43. The Δ PU-R stays in the +5.65–5.94 band across the three configurations, consistent with the MIRAGE mechanism contributing beyond raw dimensionality. PCA-projected Qwen3-Embedding-4B still outperforms three of the four native 1024-d encoders under full MIRAGE, indicating embedding quality also contributes; we report native dimensions in the main table (matching deployment) and treat this as a supplementary control.

F.2 Per-dataset decoder scaling

Table 12 reports the per-dataset decoder sweep at fixed e5-large-v2 encoder. All decoders share the same local MIRAGE-SC bank; only the online regeneration controller varies, so the Fixed-bank baselines are decoder-independent.

Cross-family pair and configuration coverage.

A 2×2 encoder-family $\times$ decoder-family check (e5 vs. Qwen3-Emb-0.6B, paired with Qwen3-1.7B vs. Llama-3.1-8B) provides a check against a same-family artifact: Δ PU-R over fixed-bank is +5.65/ + 5.15/ + 6.10/ + 5.50 across the four cells (macro Enron+CanaryDocs; paired document-bootstrap CI half-widths ≤ 0.86), with the Qwen3-encoder + Llama-decoder cell preserving a +5.50 gain. Counting the encoder sweep (Table 11), decoder sweep (Table 12), and this 2×2 , we cover 9 of the 20 cells in the 5 encoders $\times$ 4 decoders grid; the remaining 11 are natural follow-ups.

F.3 Real-Corpus Generalization

To probe external validity beyond Enron and CanaryDocs, we additionally evaluate T-MC-Q on two real, non-templated corpora; corpora, splits, and parser rules follow Appendix A.1.

Table 13 shows same-corpus margins of +3.7 to +4.4 PU-R over fixed-bank reranking in all four corpus–encoder pairs. The mechanism markers move together with the main results: accepted-anchor precision rises from 42.1% and 44.8% without support checks to 58.6% and 61.2% with them (e5-large-v2), and 21–25% of newly correct units

Table 12: Per-dataset decoder scaling ($Q=512$).

Dataset	Decoder	Fixed PU-R	MIRAGE PU-R	Δ PU-R	Δ ID-R	MIRAGE Canary-R	Tok-F1
Enron	Qwen3-1.7B	18.2 $\pm$ 0.7	24.1 $\pm$ 0.7	+5.9	+5.1	–	39.3 $\pm$ 0.6
	Qwen3-4B	18.2 $\pm$ 0.7	24.7 $\pm$ 0.6	+6.5	+6.1	–	40.0 $\pm$ 0.5
	Qwen3-8B	18.2 $\pm$ 0.7	25.1 $\pm$ 0.6	+6.9	+6.8	–	40.6 $\pm$ 0.5
	Llama-3.1-8B	18.2 $\pm$ 0.7	23.5 $\pm$ 0.7	+5.3	+4.7	–	38.8 $\pm$ 0.7
CanaryDocs	Qwen3-1.7B	19.4 $\pm$ 0.7	24.7 $\pm$ 0.9	+5.4	+4.7	22.7 $\pm$ 0.6	40.2 $\pm$ 0.6
	Qwen3-4B	19.4 $\pm$ 0.7	25.7 $\pm$ 0.6	+6.3	+4.8	23.2 $\pm$ 0.6	40.7 $\pm$ 0.5
	Qwen3-8B	19.4 $\pm$ 0.7	26.3 $\pm$ 0.5	+6.9	+5.0	23.7 $\pm$ 0.5	41.2 $\pm$ 0.4
	Llama-3.1-8B	19.4 $\pm$ 0.7	24.4 $\pm$ 0.6	+5.0	+4.0	22.0 $\pm$ 0.7	39.6 $\pm$ 0.6

Table 13: Real-corpus generalization (PU-R; Qwen3-1.7B decoder). Anchor-prec.: no-support $\rightarrow$ full (–: no no-support-check variant on this encoder); Regen-only: newly correct units absent from every fixed bank; Δ computed from unrounded values.

Corpus	Encoder	Fixed-bank	MIRAGE	Δ PU-R	Anchor-prec.	Regen-only
UDC	e5-large-v2	15.2	19.4	+4.1	42.1 $\rightarrow$ 58.6%	22.7%
	bge-m3	14.8	18.6	+3.7	–	21.4%
TWCS	e5-large-v2	16.4	20.8	+4.4	44.8 $\rightarrow$ 61.2%	24.9%
	bge-m3	15.9	20.0	+4.1	–	23.6%

appear only in regenerated candidates absent from every fixed bank. The reused adapter yields a weaker out-of-domain local bank—-independent MIRAGE-SC attains 12.4 PU-R on UDC and 13.1 on TWCS, below same-setting fixed-bank reranking (15.2 and 16.4) despite comparable Tok-F1 (31.6 vs. 31.0 on UDC)—yet the document layer still produces net gains, so the margin does not ride on a stronger local base. Absolute levels are lower than in-domain, as expected under an out-of-domain bank; the margins and the mechanism markers jointly indicate that the ordered-memory mechanism operates beyond templated slots.

G Ablations and Robustness Stress Tests

G.1 Cross-encoder ablation generalization

Table 14 reports that the ablation pattern from the main text (Table 5) is also observed beyond e5-large-v2 across three representative encoders—e5-large-v2 (E5, 1024-d), bge-m3 (BGE, 1024-d), and Qwen3-Embedding-4B (Qwen3, 2560-d). For all three, removing support checks, shuffling memory order, or falling back to fixed-bank reranking lowers PU-R relative to full MIRAGE, while the single-chunk Tok-F1 is unchanged (the local bank is unaffected by document-level ablations). Table 14 includes bootstrap half-widths for macro PU-R.

Table 14: Cross-encoder ablation generalization (macro PU-R, $Q=512$).

Variant	e5-large-v2	bge-m3	Qwen3-Emb-4B
MIRAGE full	24.4$\pm$0.61	23.0$\pm$0.51	27.1$\pm$0.63
No support check	21.7 $\pm$ 0.7	20.5 $\pm$ 0.7	24.3 $\pm$ 0.6
Shuffled memory	21.9 $\pm$ 0.7	19.5 $\pm$ 0.8	23.2 $\pm$ 0.7
Fixed bank	18.8 $\pm$ 0.56	17.3 $\pm$ 0.59	21.1 $\pm$ 0.62

Decoder no-support-check ablation. Repeating the no-support-check ablation on the largest and most diverse decoders tests whether the support check remains useful across decoder scale and family. With a fixed e5-large-v2 encoder, removing support checks drops macro PU-R by 2.5 points on Qwen3-8B (25.7 $\pm$ 0.5 $\rightarrow$ 23.2 $\pm$ 0.6; ID-R 17.4 $\rightarrow$ 15.1) and 2.9 points on Llama-3.1-8B-Instruct (24.0 $\pm$ 0.6 $\rightarrow$ 21.1 $\pm$ 0.7; ID-R 15.8 $\rightarrow$ 13.1), comparable to the +2.7 drop reported in Table 5 for the anchor configuration. See Table 14 for the full encoder-side ablation.

G.2 Strong-base, cross-family BaseInvert, and template-family stress tests

We bundle three robustness stress tests into one summary table (Table 15) and report dataset-level Canary-R for the template-family rows where Canary-R is the load-bearing metric; per-dataset values for the strong-base and cross-family-base

Table 15: Stress-test summary (macro unless noted). Strong-base deltas compare full MIRAGE with the corresponding Independent MIRAGE-SC base; cross-family BaseInvert and template-family deltas compare with fixed-bank. The Strong-K row expands K_0 from 64 to 256 and scales the local proposal/scoring budget proportionally; all other rows use the default $Q=512$.

Setting	Comparator PU-R	Full PU-R	Δ PU-R	Canary-R (Full)
<i>Strong-base ablation (e5-large-v2; macro)</i>				
Standard ($K_0=64$, Qwen3-1.7B BaseInvert)	16.2 $\pm$ 0.6	24.4 $\pm$ 0.6	+8.2	22.7
Strong-K ($K_0=256$)	18.0 $\pm$ 0.7	25.7 $\pm$ 0.6	+7.7	23.5
Strong-Backbone (Qwen3-4B)	18.8 $\pm$ 0.6	26.3 $\pm$ 0.6	+7.5	24.0
<i>Cross-family BaseInvert (macro over 2 encoders $\times$ 2 decoders)</i>				
Qwen3-1.7B BaseInvert	17.63 $\pm$ 0.76	23.23 $\pm$ 0.72	+5.60	21.94
Llama-3.2-1B BaseInvert (same shadow set)	15.61 $\pm$ 0.87	21.31 $\pm$ 0.85	+5.70	20.58
<i>Template-family generalization (CanaryDocs)</i>				
Seen (12 templates; Fixed/NoSupp/Full)	20.4 / 23.5 / 26.2 $\pm$ 0.7	—	+5.8	17.2 / 23.1 / 26.8
Held-out (12 templates; Fixed/NoSupp/Full)	18.0 / 20.6 / 22.8 $\pm$ 0.7	—	+4.8	11.0 / 15.4 / 18.6
Slot-only R (Seen / Held-out, Full)	—	—	—	29.3 / 20.5
Full-string R (Seen / Held-out, Full)	—	—	—	24.9 / 16.0

rows will be released in the artifact bundle (see the Ethics Statement) and reported here as Enron/CanaryDocs macro-averages. The three stress tests address: (i) base repair—stronger local banks via $K_0 \times 4$ or a Qwen3-4B backbone; the memory Δ stays at +7.5–7.7 relative to the stronger base, down from +8.2 at the standard base; (ii) pipeline family bias—a Llama-3.2-1B BASEINVERT backbone trained on the same 80k shadow documents; the gain stays at +5.60 (Qwen3) / +5.70 (Llama), with overlapping bootstrap CIs; (iii) template memorization—12 CanaryDocs template families held out from shadow training; Canary-R drops from 26.8 (seen) to 18.6 (held-out) but the memory gain over fixed-bank stays at +4.8 PU-R, so the mechanism operates beyond template memorization. Slot-only R (matching only the canary value regardless of surrounding template) exceeds Full-string R, as expected. Table 16 covers the fourth stress test (single-budget artifact) under proportional slot scaling; the fifth alternative, boundary mismatch, is in Appendix G.4.

G.3 Query-budget sensitivity

Table 16 sweeps the per-chunk online encoder-query budget $Q \in \{64, 128, 256, 512, 1024, 2048\}$ with proportional slot scaling (proposal:calibration:support:regeneration:final = 7:1:4:3:1, macro Enron+CanaryDocs on the anchor configuration). This is a separate budget-sensitivity run; its $Q=512$ row is reported as the sweep estimate rather than replacing

Table 16: Query-budget sensitivity (macro PU-R; proportional slot scaling). This sweep is a separate budget-sensitivity run; its $Q=512$ row can differ slightly (≤ 0.1 PU-R) from the headline anchor-configuration estimate in Table 2.

Q	Fixed	NoSupp	MIRAGE	Δ	Tok-F1
64	11.1 $\pm$ 1.06	12.6 $\pm$ 1.23	13.9 $\pm$ 1.16	+2.80	29.5
128	14.5 $\pm$ 0.87	16.5 $\pm$ 0.93	18.2 $\pm$ 0.85	+3.71	33.4
256	16.7 $\pm$ 0.74	19.4 $\pm$ 0.84	22.0 $\pm$ 0.84	+5.31	37.3
512	18.8 $\pm$ 0.60	21.8 $\pm$ 0.70	24.3 $\pm$ 0.64	+5.56	39.9
1024	20.2 $\pm$ 0.59	23.5 $\pm$ 0.52	26.4 $\pm$ 0.57	+6.20	41.2
2048	21.0 $\pm$ 0.53	24.4 $\pm$ 0.48	27.5 $\pm$ 0.42	+6.48	41.7

the headline anchor-configuration estimate in Table 2. MIRAGE’s advantage grows with Q and is most cost-effective at $Q=256$ –512; fixed-bank reranking shows diminishing returns above $Q=512$, and even with a larger proportional proposal pool (896 proposal queries at $Q=2048$) remains below MIRAGE at $Q=512$, consistent with memory-conditioned regeneration rather than unused budget slots.

G.4 Boundary-Mismatch Robustness

The protocol assumes approximate boundary knowledge (A4 in Section 3). Table 17 tests how much boundary alignment the attack actually requires: the attacker’s assumed segmentation is deliberately misaligned with the victim chunker (ordered 128-token/32-overlap blocks), with offset jitter of ± 4 to ± 32 tokens (± 32 equals the overlap size), chunk sizes of 96 and 160 tokens, and

Table 17: Boundary-mismatch robustness (Enron PU-R, anchor configuration); both methods use the same mismatched setting.

Boundary condition	MIRAGE	Fixed-bank	Δ	Retained
Matched (victim chunker)	24.1	18.2	+5.9	100%
Offset jitter ± 4	23.4	18.1	+5.3	90%
Offset jitter ± 8	23.0	18.0	+5.0	85%
Offset jitter ± 16	22.1	17.4	+4.7	80%
Offset jitter ± 32 (= overlap)	21.0	16.8	+4.2	71%
Chunk size 96 / 160	21.9 / 21.4	17.1 / 16.9	+4.8 / +4.5	81% / 76%
Sentence-based chunker	21.2	16.6	+4.6	78%

Table 18: Mechanism evidence across three encoders (fixed Qwen3-1.7B decoder, macro Enron+CanaryDocs); definitions follow Figure 5.

Encoder	Variant	Acc.-prec.	False-rate	Rescue
e5	Full	69.4 $\pm$ 0.9	7.1 $\pm$ 0.8	34.8 $\pm$ 1.1
	NoSupp	43.6 $\pm$ 1.1	18.9 $\pm$ 1.0	24.1 $\pm$ 1.2
bge-m3	Full	67.8 $\pm$ 1.0	8.0 $\pm$ 0.9	33.5 $\pm$ 1.2
	NoSupp	42.1 $\pm$ 1.2	20.3 $\pm$ 1.1	23.0 $\pm$ 1.3
Qwen3-Emb-4B	Full	72.5 $\pm$ 0.8	5.8 $\pm$ 0.7	36.9 $\pm$ 1.0
	NoSupp	46.4 $\pm$ 1.0	16.5 $\pm$ 1.0	26.8 $\pm$ 1.2

a sentence-based chunker in place of fixed-length blocks. Both MIRAGE and fixed-bank reranking run under the same mismatched setting, on Enron with the anchor configuration.

MIRAGE stays above the same-setting fixed-bank reranker under every tested mismatch, retaining 71–90% of the matched-boundary margin (+5.9 down to +4.2), with gradual degradation and no cliff at the overlap scale. Exact victim-chunker alignment is therefore not required across these settings; completely unknown segmentation remains outside this evidence. This supports treating boundary information as an evaluation convenience (A4) rather than a requirement of the attack.

H Cross-encoder Mechanism Evidence

Table 18 extends Figure 5 to bge-m3 and Qwen3-Embedding-4B; accepted-anchor precision remains above 65% and false-anchor rate below 10% under full MIRAGE across all three encoders, while removing support checks lowers precision by 25–26 points, doubles or triples the false-anchor rate, and reduces absent-bank rescue by about 10 points—evidence for the role of the support-check mechanism in the observed gains.

I Cross-Dataset and Adaptive Defense Probes

Table 6 reports the defense stress tests, extending the strongest of them—stripping provenance metadata and coarsening chunk order—to CanaryDocs; the adaptive attacker is specified here: against the coarsened-order defense, the attacker relaxes the order-penalty weight and re-fits the support threshold on held-out development data, keeping the $Q=512$ budget. The same qualitative reduction appears on CanaryDocs as on Enron, and the adaptive attacker recovers part of the leakage while remaining well below the no-defense point estimate. These are initial probes rather than a defense benchmark; a full adaptive, multi-defense evaluation remains future work.

I.1 Deployment Checklist

For RAG and vector-store operators (Table 6 quantifies the measured effects):

- **Metadata minimization.** Treat document grouping, chunk order, and chunk offsets as sensitive and strip or coarsen them whenever embeddings leave the trust boundary.
- **Access control.** Restrict read access to provenance fields independently of vector access.
- **Query monitoring.** Budget and monitor encoder-query access; rate limits reduce reconstruction under the same local bank.
- **Canary auditing.** Insert synthetic canary documents and audit stores, caches, and logs for their recovery.

An adaptive attacker recovers part of the leakage, so these measures should be layered rather than relied on singly.