

PRIME: Pre-Retrieval Index Modeling & Enhancement for Structured Document Retrieval

Cheuk Hei Chung^{*} Li Liu[†]

The Hong Kong University of Science and Technology (Guangzhou)

Abstract

High-precision retrieval over external knowledge bases is essential for LLM agents to operate effectively. However, existing Information Retrieval (IR) research primarily optimizes query-side retrievers (e.g., ColBERT, BGE) and generators (e.g., Self-RAG, GraphRAG), while reducing corpus-side representation to chunking. This leaves index-side quality an underexplored lever. We propose Pre-Retrieval Index Modeling & Enhancement (PRIME), an index-side optimization framework built on two complementary tracks: an LLM-driven Semantic Restructuring (SR) track that transforms long documents into structured, information-dense sections, and a deterministic Metadata Synthesis (MS) and Hierarchical Retrieval (HR) track that extracts semantic anchors and heading structure for coarse-to-fine section retrieval. Our evaluations show that deterministic metadata matches structure-aware chunking’s retrieval quality while cutting embedding computation by 45.8 $\times$ and index vector count by 22.4 $\times$. On top of these gains, Semantic Restructuring (SR) further improves Answer-F1 by 13.7% in full deployment, with characterized quality–fidelity–cost trade-offs. Across QASPER, BRIGHT, and NFCorpus, we further characterize three qualitative regimes that determine whether index-side optimization helps or degrades retrieval, while showing that these index-side effects are neutralized under a strong reranker.

1 Introduction

Efficient information retrieval (IR) is essential for LLM agents, as the quality of retrieved contexts serves as the foundational bottleneck for downstream generation (Guu et al. 2020; Lewis et al. 2020; Borgeaud et al. 2022; Izacard et al. 2023). While query-side retrievers and generators have been extensively refined—from sparse models (Robertson and Zaragoza 2009) to dense dual-encoders (Karpukhin et al. 2020), late-interaction frameworks (Khattab and Zaharia 2020), and cross-encoder rerankers—corpus-side representation has largely remained a fixed input. Conventional preprocessing merely alters text without enhancing its information density or logical structure.

Recent methods began to exploit document structure for indexing, and Singh, Reddy, and Chopra (2026) show that once a strong cross-encoder reranker is present, common retrieval enhancements yield no reliable additional gain. This raises the question of whether index-side representation can

^{*}Email: chchung634@connect.hkust-gz.edu.cn

[†]Corresponding author: Li Liu (avrillliu@hkust-gz.edu.cn)

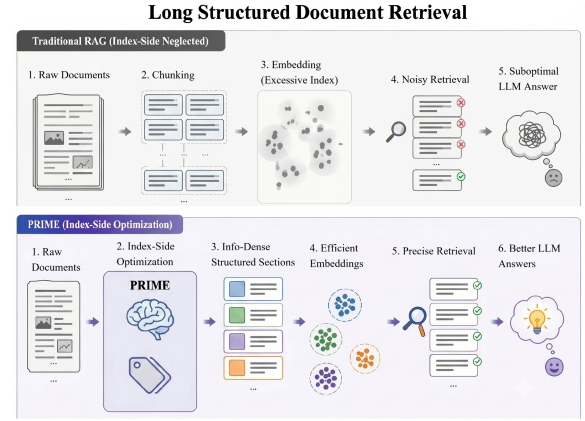


Figure 1: Overview of Index-Side Optimization. PRIME operates on the corpus representation prior to indexing, complementing existing query-side and generator-side methods without modifying the retriever.

be improved and whether its benefits are compensated by a strong pipeline.

We propose that retrieval quality is bounded by the information density of the indexed corpus. Once a strong retriever and reranker are in place, the remaining leverage lies in information structure and density at the index level. Guided by this principle, we present PRIME, an index-side quality optimization framework that operates on two complementary tracks without modifying the retriever. Semantic Restructuring (SR) employs an LLM-driven transformation to explicate implicit argument structure and compress descriptive prose within each section, increasing knowledge density for long structured documents. In parallel, Metadata Synthesis (MS) and Hierarchical Retrieval (HR) form a fully deterministic track: MS extracts semantic anchors and heading structure from the original document, which HR then consumes with lightweight high-density embeddings for coarse document filtering and a cross-encoder for fine-grained section reranking. Figure 1 situates PRIME in the retrieval landscape.

Our key contributions are:

- Deterministic structural metadata matches structure-aware chunking’s retrieval quality while cutting embed-

PRIME: Pre-Retrieval Index Modeling & Enhancement

Index-side quality optimization through two complementary tracks

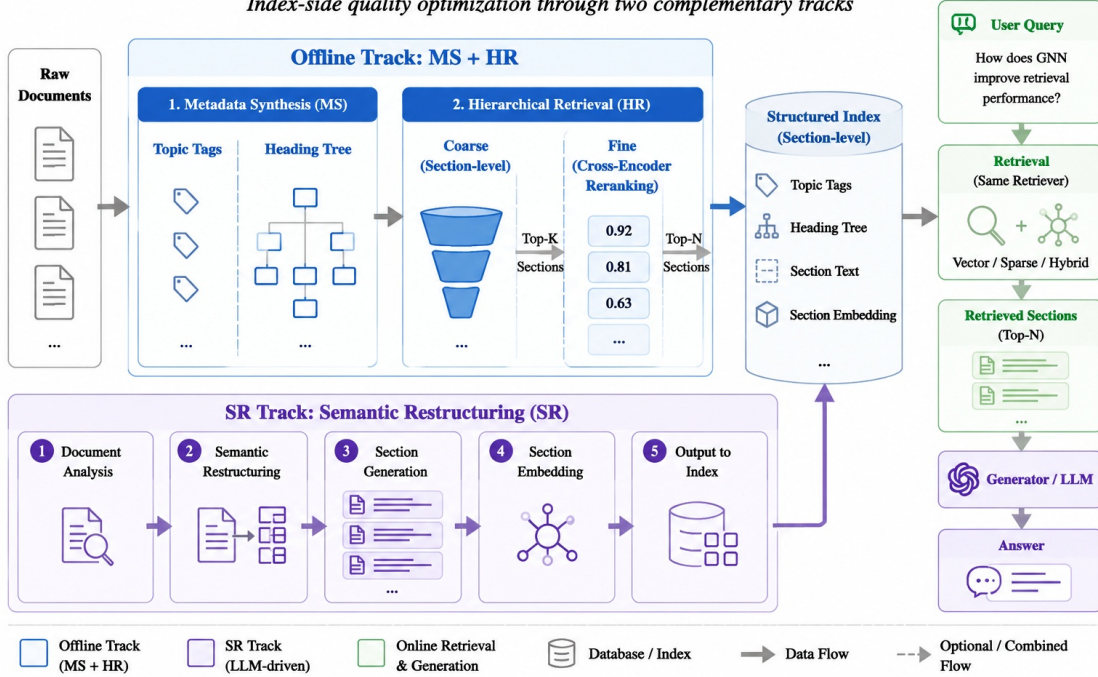


Figure 2: PRIME architecture. The framework operates on two complementary tracks: (Left) an LLM-driven Semantic Restructuring track transforms documents into knowledge-dense representations; (Right) a deterministic track uses Metadata Synthesis to extract structural anchors, which power a Hierarchical Retrieval pipeline with RRF fusion and cross-encoder reranking.

ding computation by 45.8× and index vector count by 22.4×, establishing a more efficient and lightweight alternative to LLM-mediated indexing.

- On top of these deterministic gains, Semantic Restructuring (SR) further yields an additional +13.7% Answer-F1 in full deployment, with characterized semantic quality, fidelity, and cost trade-offs.
- Through a 5×3 factorial ablation on QASPER, BRIGHT, and NFCorpus, we identify three corpus regimes, deficit-bridging, structural exploitation, and structural saturation, that determine whether index-side optimization helps or degrades retrieval, and show that a strong cross-encoder reranker neutralizes these gains.

2 Related Works

Query-side retrieval has been extensively optimized. From sparse retrieval (Robertson and Zaragoza 2009) to learned sparse representations such as SPLADE (Formal, Piwowarski, and Clinchant 2021). Dense retrieval progressed through dual-encoder architectures including DPR (Karpukhin et al. 2020), ANCE (Xiong et al. 2021) and Contriever (Izacard et al. 2021), multi-vector late-interaction models represented by ColBERT (Khattab and Zaharia 2020), and multilingual general-purpose encoders like BGE-M3 (Xiao et al. 2024). Multi-stage refinement adds cross-encoder rerankers (Nogueira and Cho 2019) and LLM-based listwise ranking approaches such as RankGPT (Sun et al.

2023). On the query formulation side, methods including HyDE (Gao et al. 2023) expand queries via hypothetical passages. However, these methods optimize how queries are encoded, matched, and ranked against a fixed document representation.

On Corpus-Side, several lines of work have modified the corpus before indexing. Early document preprocessing techniques (Manning 2008) adjust surface token forms. Early structured document retrieval (Wilkinson 1994) exploited pre-existing field markup for weighted term scoring but relied on author-provided structure. Extractive summarization for IR (Tombros and Sanderson 1998) selects a subset of existing sentences and trades recall for conciseness without generating new, denser text. Document expansion via query prediction such as Doc2Query (Nogueira et al. 2019), appends synthetic queries to broaden lexical coverage for term matching, yet leaves the document’s original knowledge content and information density unchanged. Chunking strategies (Bhat et al. 2025) determine where to place segment boundaries but preserve the token composition within each segment. Across these approaches, a common pattern emerges: each modifies how text is packaged, not what the text conveys.

Index-Side Restructuring: Recent Progress and Remaining Gap. Recent methods have begun exploiting document structure for indexing. Multi-view Content-aware Indexing (Dong et al. 2024), DISRetrieval (Chen et al. 2026b), and PaperRegister (Wang et al. 2026) organize documents by

content structure, discourse, or register but derive representations from original tokens without transformation. Summary-Augmented Chunking (Reuter et al. 2025) injects document-level context into chunks, improving retrieval without rewriting the indexed text. RAPTOR (Sarthi et al. 2024) generates recursive LLM summaries but relies on embedding-driven clustering. GraphRAG (Edge et al. 2024) and follow-up graph-based methods such as PathRAG (Chen et al. 2026a) and LeanRAG (Zhang et al. 2026) reorganize at the graph level without modifying indexed text. Across these methods, structure is exploited but text is not transformed; text is enriched but its information density is not increased. Singh, Reddy, and Chopra (2026) add a critical constraint: once a strong cross-encoder reranker is present, retrieval enhancements yield no Holm-corrected gain. PRIME occupies the resulting gap; generative text restructuring that increases information density (SR), combined with deterministic metadata extraction (MS) and structure-aware hierarchical retrieval (HR), with characterized boundary conditions.

3 Method

3.1 Preliminary

Let $\mathcal{C} = \{D_1, \dots, D_N\}$ denote the corpus where each document D_i consists of second-level headings (H2). PRIME processes documents through two tracks, illustrated in Figure 2. **Semantic Restructuring (SR)** employs LLM-driven analysis to transform D_i into a knowledge-dense representation D_i^* that preserves the original H2 structure. **Metadata Synthesis (MS) and Hierarchical Retrieval (HR)** form a deterministic track where MS extracts top-5 discriminative terms $\mathbf{t}$ and a heading tree $\mathbf{h}$ from the corpus (D_i or D_i^*), and HR consumes these metadata through a two-stage pipeline: coarse document filtering on high-density metadata embeddings followed by section-level cross-encoder reranking.

PRIME supports two deployment modes. **Offline mode** runs MS and HR directly on the original corpus without SR, producing metadata $M_i = \text{MS}(D_i)$. **Full mode** applies SR to produce D_i^* , then extracts metadata from the restructured text ($M_i^* = \text{MS}(D_i^*)$) and runs HR; D_i^* also replaces D_i as input to BM25 and the cross-encoder reranker. For component ablation, we also evaluate the SR body text by itself without MS metadata (denoted SR (full-text) in Table 2).

Formally, the two modes are:

$$\begin{aligned}
 \text{Offline: } & D_i \xrightarrow{\text{MS}} M_i \\
 & I_{\text{vec}} = \text{Index}(M_i) \\
 & I_{\text{bm25}} = \text{Index}(D_i) \\
 \\
 \text{Full: } & D_i \xrightarrow{\text{SR}} D_i^* \\
 & D_i^* \xrightarrow{\text{MS}} M_i^* \\
 & I_{\text{vec}} = \text{Index}(M_i^*) \\
 & I_{\text{bm25}} = \text{Index}(D_i^*)
 \end{aligned} \tag{1}$$

where $M_i = (\mathbf{t}_i, \mathbf{h}_i)$ denotes the top-5 discriminative terms and heading tree.

3.2 Semantic Document Restructuring

SR transforms each document D_i into a knowledge-dense representation D_i^* while preserving the original H2 structure. It operates along two semantic axes: **Explication** re-constructs implicit argumentative structures within each H2 section into declarative statements, and **Compression** eliminates descriptive prose while retaining core contributions, conclusions, key terms, and numerical results. By removing low-IDF filler tokens, SR shifts D_i^* toward high-IDF discriminative terms, so the cross-encoder’s pairwise attention operates on a denser semantic signal per token and the LLM generator consumes fewer tokens for the same answer-relevant content.

To ensure complete coverage, SR invokes the LLM per H2 section following the document’s heading tree, mitigating lost-in-the-middle effects (Liu et al. 2024). The computational cost is $c(D) = \sum_{h \in H2(D)} (\alpha \cdot \text{len}(h) + \beta)$, where α captures per-token inference cost and β the per-section prompt and API overhead; output tokens add a symmetric cost factor (full break-even analysis in the Supplementary Material, Appendix D.6).

3.3 Deterministic Metadata Synthesis

The Metadata Synthesis (MS) module generates two types of metadata for the HR pipeline (Hierarchical Retrieval): discriminative terms for document-level separation and a heading tree for section-level navigation.

Discriminative terms are selected as the top-5 candidates using a TF-IDF scoring function (Salton and Buckley 1988):

$$\begin{aligned}
 \text{score}(t) &= \text{TF}(t) \times \text{IDF}_{\text{corpus}}(t) \\
 &\times (1 - \alpha)^{\text{first_pos}/n} \times (1 + \beta \cdot \mathbb{1}_{\text{in_heading}}(t)). \tag{2}
 \end{aligned}$$

where $\text{TF}(t)$ and $\text{IDF}_{\text{corpus}}(t)$ denote the term frequency and corpus-level inverse document frequency of t , first_pos is the token position of t ’s first occurrence in D_i , $n = |D_i|$, $\alpha = 0.3$ controls the positional decay rate, and $\beta = 2.0$ weights the heading bonus. Terms are whitespace-delimited tokens after lowercasing and punctuation removal. $\text{IDF}_{\text{corpus}}$ uses the smoothed form $\log \frac{N+1}{\text{df}(t)+1} + 1$ to avoid negative values when $\text{df}(t) = N$, and is computed from the evaluation corpus itself rather than from an external reference, since domain-specific IDF better captures distinctiveness across structurally diverse corpora (evaluated in Experiments).

The heading tree, parsed from H2 headings of D_i , together with the document title and discriminative terms form a high-density metadata embedding for coarse filtering and provide section boundaries for fine-grained reranking.

3.4 Structure-Aware Hierarchical Retrieval

The Hierarchical Retrieval (HR) module consumes MS-generated metadata through a two-stage pipeline: coarse document filtering followed by section-level reranking. This design reflects the observation that document-level relevance can be assessed via coarse semantic signals, while section-level relevance requires fine-grained token interaction.

Stage 1 — High-Density Document Filtering Stage 1 retrieves candidate documents using two parallel pathways. The BM25 pathway indexes the full document text for exact keyword matches. The vector pathway indexes the high-density metadata (title + discriminative terms + heading tree) encoded via BGE-small-en-v1.5 (384-dim; Xiao et al., 2024). Results from both pathways are merged via Reciprocal Rank Fusion (RRF, $k = 60$) (Cormack, Clarke, and Buettcher 2009) to form a top-30 candidate set. The heading tree is invariant to whether SR is applied (H2 structure is preserved). In PRIME (offline), discriminative terms are extracted from D_i . In full PRIME, terms are extracted from D_i^* , reflecting SR’s vocabulary shift; the resulting impact is quantified in the Ablation.

Stage 2 — Section-Level Evidence Localization Stage 2 scores each H2 section of candidate documents using a cross-encoder (ms-marco-MiniLM-L-6-v2; Reimers & Gurevych, 2019), fine-tuned on MS MARCO (Bajaj et al. 2016). The top-5 sections by CE score are forwarded to the LLM generator. Document-level rankings in Experiments are derived via best-section aggregation: $\text{score}(D) = \max_{s \in D} \text{CE}(q, s)$. The lightweight cross-encoder eliminates query-time LLM overhead typical of listwise rerankers. Full model configuration and hyperparameters are provided in Appendix B.

4 Experiments

Datasets We evaluate on three structurally diverse benchmarks: **QASPER** (Dasigi et al. 2021) serves as the target scenario for PRIME’s full pipeline: we use 416 papers from the official test set, producing 1,451 queries with document-level qrels (one relevant document per query, 415 negatives). Queries are natural language questions without title or metadata. **BRIGHT** (SU et al. 2025) tests cross-domain generalization of PRIME’s offline mode across mixed-structure sources. We evaluate all twelve subsets despite four (AoPS, LeetCode, TheoremQA-Q, and TheoremQA-Theorems) requiring reasoning which fall outside PRIME’s design scope. **NFCorpus** (Boteva et al. 2016), from the BEIR benchmark (Thakur et al. 2021), defines the boundary condition where structural metadata provides no discriminative signal.

Evaluation Metrics We evaluate PRIME across four complementary dimensions: retrieval quality, downstream QA utility, text density, and operational cost. This evaluation framework establishes a complete evidence chain from index-side representation quality to end-to-end task performance.

Retrieval Quality. All three datasets use NDCG@5, P@5, R@5, P@1, R@1, and MRR. QASPER uses document-level qrels (one relevant doc per query), so $P@1 \equiv R@1$ and $R@5 = 5 \times P@5$, and NDCG@5 reduces to a rank-based metric driven entirely by the position of the single relevant document. NFCorpus uses graded relevance (0–2); BRIGHT uses evidence-based binary qrels.

Consumption-Side QA Quality. QA is evaluated via QASPER’s official Answer-F1 (Dasigi et al. 2021), computed as the best match over the three reference answers per question.

Text Density — Corpus-Level Transformation. We quantify SR’s effect on document representation via four metrics. Three capture orthogonal dimensions: **stopword ratio** (filler compression), **technical term density** (domain terms per 100 tokens), and **mean IDF** (average token distinctiveness) (Robertson 2004). The fourth, **semantic density**, integrates them into a composite signal-to-noise ratio. We define semantic density as:

$$\text{Semantic Density} = \frac{\text{content_tokens} \times \text{mean_idf}}{\text{stop_tokens} + 1}$$

The numerator estimates total information carried by a document, content-bearing tokens weighted by their average distinctiveness across the corpus. The denominator estimates noise, filler words that contribute surface length without semantic content. A full derivation appears in Appendix D.7.

Implementation All PRIME variants share a unified backend: BGE-small-en-v1.5 (384-dim) (Xiao et al. 2024), hnswlib configured with $M=32$, $ef=200$, and cosine similarity (Malkov and Yashunin 2018), and SQLite FTS5 with porter stemming. All query-side components (embedding, BM25, vector search, and cross-encoder (ms-marco-MiniLM-L-6-v2) (Reimers and Gurevych 2019)) run on a single NVIDIA RTX 5090 GPU. SR uses DeepSeek-V4-pro (Xu et al. 2026) at temp=0; QA generation in Context Quality uses DeepSeek-V4-Flash. All hyperparameters were fixed before evaluation based on literature defaults with no tuning on test queries. Full configuration and hyperparameter tables are in Appendix D.

4.1 Main Results

Retrieval Quality

QASPER Index-side optimization improves retrieval, but its measurable advantage narrows as downstream pipeline components increasingly compensate for weaker embeddings. We establish this through two complementary analyses: a shared-pipeline ceiling that reveals convergence under a strong reranker (Table 1), and a factorial decomposition that isolates index representation from pipeline effects (Table 2).

All six methods converge within $\text{NDCG}@5 = 0.389\text{--}0.398$, echoing the reranker-saturation effect reported by Singh, Reddy, and Chopra (2026): once a strong cross-encoder is in the loop, corpus-side differences become nearly inseparable. This apparent equivalence conflates index quality with pipeline strength. We therefore evaluate a 5×3 factorial matrix varying only the indexed text representation across three pipelines (Dense-only, Hybrid BM25+RRF, Hybrid+CE), disentangling the two. 95% paired bootstrap CIs (10,000 resamples) are reported.

As shown in Table 2, **Index representation matters in isolation**: in the Dense-only column, all four structured representations outperform chunk-256 ($p < 0.001$). The convergence in table 1 reflects that BM25 and the CE reranker mask these underlying embedding differences: the advantage of structured over raw narrows from +15–22% (Dense-only) to +7–9% (Hybrid) to <0.005 (Hybrid+CE).

Table 1: Retrieval quality on QASPER 416 (1,451 queries). *BGE-M3 uses chunk-256 max-pooling; Summary-Augmented prepends section summaries to chunks; RAPTOR-lite does section+window retrieval without LLM summarization. Structure-chunks is a simpler H2-aware chunking baseline (no MS metadata, no HR pipeline). PRIME (Full) is excluded because SR alters the BM25-indexed text, breaking the shared-pipeline condition.*

Method	NDCG5	NDCG10	Recall5	Precision5	P1	MRR
Raw (chunk-256)	0.391	0.405	0.462	0.092	0.306	0.373
BGE-M3 (1024-dim)	0.389	0.398	0.453	0.091	0.306	0.371
Summary-Augmented	0.398	0.411	0.471	0.094	0.309	0.379
RAPTOR-lite	0.391	0.408	0.461	0.092	0.307	0.375
PRIME (offline) (ours)	0.396	0.409	0.469	0.094	0.309	0.377
Structure-chunks	0.396	0.411	0.466	0.093	0.310	0.379

Table 2: Factorial ablation: NDCG@5 on QASPER 416. *95% paired bootstrap CIs in brackets. All configurations share identical retrieval pipelines; only the indexed text varies between rows. The five index representations and three retrieval pipelines are defined in Preliminary.*

Index Representation	Dense-only	Hybrid (BM25+Vec+RRF)	Hybrid+CE
Raw (chunk-256)	0.175 [.164,.186]	0.311 [.296,.327]	0.391 [.370,.412]
Structure-chunks	0.201 [.187,.216]	0.338 [.321,.354]	0.396 [.375,.416]
PRIME (offline)	0.201 [.186,.217]	0.332 [.315,.349]	0.396 [.375,.416]
SR (full-text)	0.208 [.196,.221]	0.245 [.232,.258]	0.328 [.309,.347]
PRIME (Full)	0.214 [.199,.229]	0.241 [.227,.256]	0.325 [.307,.345]

Within this narrowing gap, structured representations exhibit distinct efficiency trade-offs: **Deterministic metadata matches structure-aware chunking at a fraction of the cost.** Structure-chunks and PRIME (offline) remain indistinguishable (overlapping 95% CIs). Yet, PRIME (offline) embeds only metadata while capturing the same document-filtering signal as full-text chunking, at 45.8× less embedding computation and 22.4× fewer indexed vectors (Operational Cost). In contrast, both SR-based representations **trade lexical signal for semantic density.** PRIME (Full) achieves the highest Dense-only NDCG@5 (0.214), confirming that the compact metadata embedding (title, topic tags, and heading tree extracted from SR text, ~200 tokens) already captures stronger discriminative signal than raw chunk-256. However, BM25—common to both SR (full-text) and PRIME (Full) via the same SR body index—drops substantially below Raw in the Hybrid pipeline (0.241 and 0.245 vs. 0.311), because rewriting shifts text toward high-IDF terms at the expense of surface keyword overlap. CE reranking partially recovers this loss (Hybrid+CE: 0.325 and 0.328), though this retrieval-side cost is ultimately offset by improved context quality during generation (Context Quality).

BRIGHT We evaluate PRIME (offline) on all twelve BRIGHT tasks. Table 3a reports the eight retrieval tasks; TheoremQA-Theorems is excluded from the average and noted in the caption. The other three reasoning tasks are reported separately in Appendix E of Supplementary Material.

The Δ distribution across the eight structured tasks falls into three qualitative regimes, each predicted by corpus properties that can be assessed before retrieval. At one extreme, **deficit-bridging** (pony, +175.6%): dense API reference documentation is poorly represented by BGE-small’s pretraining distribution, so the discriminative terms by MS serve as a natural-language bridge for the encoder. Typical structured

documents earn modest gains from **structure-exploiting** (biology to economics, +4.6% to +22.3%), where structured H2 sections with [edit] wiki markers anchor CE section reranking. At the opposite end, **structure-absent or saturated** domains (sustainable_living to earth_science, -3.1% to -23.1%) show that prose already well-embedded by BGE-small, with sparse or empty heading trees, receives no additional discriminative signal from metadata. TheoremQA-Theorems (-59.8%, excluded from the average) marks a distinct fourth mode rather than saturation: its LaTeX blocks are parsed as H2 sections but fragment mathematical proofs and requires reasoning, so the CE scores an isolated definition against a query that requires a complete inferential chain.

NFCorpus NFCorpus (Boteva et al. 2016) defines the opposite boundary. This dataset features short documents without H2 headings, PRIME extracts only a pseudo-title with five TF-IDF topic tags per document. We also report PRIME (Full-Text) where the same topic tags are injected alongside the original full-text embedding to isolate the metadata contribution.

As shown in Table 4, PRIME alone drops 19.5% NDCG@5 below BGE-small: the five topic tags are insufficient context for short, high-density abstracts. In contrast, PRIME (Full-Text) recovers to within 2.1% of BGE-small (NDCG@5 0.366 vs 0.374), slightly exceeding in MRR, confirming that the tags introduce a mild perturbation without adding discriminative signal, the boundary condition where structural metadata ceases to help and begins to interfere.

Context Quality Retrieval Quality showed that SR degrades retrieval under Hybrid+CE (NDCG@5 0.325 vs. 0.391 for Raw). We find that SR’s downstream gains in context quality partially compensate for this retrieval cost. Table 5 traces the causal chain from text density through evidence localization to answer quality.

Table 3: Cross-domain retrieval on the eight BRIGHT structured tasks. $\Delta = (\text{PRIME-Baseline})/\text{Baseline}$; 95% CIs from 10,000-sample paired bootstrap (BCa) over per-query NDCG@5. Per-task CIs span zero (65-103 queries/task), so individual task deltas are not independently significant; the +6.9% macro-average reflects a distributional shift across domains rather than a per-task claim. TheoremQA-Theorems (PRIME NDCG@5 = 0.012, $\Delta = -59.8\%$) is excluded from this average.

Tasks	BGE-small (baseline)			PRIME (offline)			Δ NDCG@5	95% CI (Δ)
	NDCG@5	P@5	R@5	NDCG@5	P@5	R@5		
Pony	0.067	0.073	0.048	0.185	0.154	0.126	+175.6%	[−0.063, +0.299]
Biology	0.379	0.128	0.491	0.464	0.155	0.598	+22.3%	[−0.188, +0.357]
Stackoverflow	0.234	0.075	0.359	0.272	0.079	0.363	+16.2%	[−0.203, +0.279]
Robotics	0.181	0.052	0.253	0.200	0.059	0.282	+10.5%	[−0.250, +0.290]
Economics	0.247	0.070	0.333	0.258	0.074	0.354	+4.6%	[−0.242, +0.265]
Sustainable_living	0.250	0.080	0.342	0.242	0.082	0.343	−3.1%	[−0.259, +0.244]
Psychology	0.217	0.069	0.304	0.199	0.061	0.259	−8.2%	[−0.277, +0.242]
Earth_science	0.458	0.157	0.537	0.352	0.128	0.419	−23.1%	[−0.404, +0.193]
Avg (8 tasks)	0.254	0.088	0.333	0.272	0.099	0.343	+6.9%	—

Table 4: Retrieval quality on the full NFCorpus. Graded relevance 0–2 binarized at ≥ 1 . PRIME(FT) denotes PRIME (Full-Text), Δ_1 denotes PRIME vs BGE-small and Δ_2 denotes PRIME (Full-Text) vs BGE-small.

Method	NDCG@5	P@5	R@5	MRR
BM25	0.350	0.295	0.119	0.530
BGE-small	0.374	0.324	0.121	0.538
BGE-base	0.387	0.336	0.126	0.561
PRIME	0.301	0.264	0.084	0.460
PRIME (FT)	0.366	0.316	0.118	0.540
Δ_1	−19.5%	−18.5%	−30.6%	−14.5%
Δ_2	−2.1%	−2.5%	−2.5%	+0.4%

Table 5 quantifies SR’s effect at two intermediate stages. Panel A reveals that SR increases the knowledge density of the context substantially: filler tokens drop by 36.3%, mean IDF rises 19.7%, and semantic density increased by 127.3%. Panel B shows that this signal propagates to all evidence retrieval metrics on the 708 queries with section-level evidence in both corpora: SR+CE outperforms Raw+CE on section-level NDCG@5 (+45.7%), Recall@5 (+59.9%), Precision@5 (+7.9%), MAP (+69.4%), MRR (+15.4%), and Token Recall@5 (+14.1%), while indexing 29.0% fewer evidence tokens per query, confirming that the CE better localizes relevant content within fewer tokens when the underlying text is denser. Across QASPER’s question categories, SR improves section-level Recall@5 for extractive (+63%), abstractive (+56%), and unanswerable (+61%) queries, indicating the gain stems from text density rather than answer format. Yes/no queries are excluded (2 queries with evidence-bearing sections). Full derivation of density metrics in Appendix D.5, full evidence retrieval metrics in Appendix D.7.

Finally, we quantify SR’s end-to-end QA impact, denoting configurations as retrieval corpus to generation corpus. All configurations share the same full-body hybrid retrieval pipeline, differing only in whether the raw or SR corpus being indexed and read. Raw to Raw yields a baseline Answer-F1 of 0.0842. Raw to SR adds +2.9% (0.0866), isolating SR’s value as denser reading material. Full to Full reaches

Table 5: SR causal chain: text density and evidence localization (QASPER 416). Panel A uses fixed raw-corpus reference IDF. Panel B uses section qrels from gold evidence spans mapped to H2 sections, averaged over the 708 shared queries of Raw and SR section qrels to avoid survivor bias.

	Raw	SR	Δ %
Panel A: Corpus-level density (416 docs)			
Stopword ratio	0.365	0.233	−36.3%
Mean IDF	1.947	2.331	+19.7%
Tech term density (per 100 toks)	46.3	53.8	+16.1%
Semantic density (composite)	3.42	7.76	+127.3%
Panel B: Section-level evidence retrieval (CE reranker)			
NDCG@5	0.127	0.185	+45.7%
Recall@5	0.128	0.204	+59.9%
Precision@5	0.089	0.096	+7.9%
MAP	0.095	0.161	+69.4%
MRR	0.198	0.229	+15.4%
Token Recall@5	0.442	0.505	+14.1%
Gold evidence toks per query	346.9	246.3	−29.0%

0.0957 (+13.7%), despite its document-level retrieval being substantially weaker (demonstrated in Table 2). The gap between Raw to SR and Full to Full shows that SR benefits QA primarily through improved cross-encoder evidence localization on denser sections, which more than compensates for the document-level retrieval drop.

SR Fidelity SR increases text density but risks discarding answer-critical content or introducing factual errors; Table 6 summarizes three audit dimensions across evidence retention, entity preservation, and NLI entailment on QASPER (416 documents, 5,589 raw to SR section pairs, 2,231 gold evidence spans).

As shown in Table 6, 63.2% of evidence spans survive restructuring; losses are almost exclusively through compression rather than hallucination (only 1.3% of spans appear in SR but not raw text). The 25.7% of documents with retention below 50% share a recognizable profile: multi-step procedural reasoning or cross-section comparisons whose intermediate steps SR’s declarative reformulation flattens into short

Table 6: SR fidelity audit on QASPER. Panel A: a span is retained if its character-level content appears verbatim in the SR block. Panel B uses spaCy en_core_web_sm. Panel C uses roberta-large-mnli.

Metrics	Value
Panel A: Evidence retention (2,231 gold spans, 373 docs)	
Overall span retention	63.2%
Median per-document retention	66.7% (p25: 44.4%, p75: 100%)
Perfect retention (100% of spans)	26.8% of docs
High retention (80–99%)	35.1% of docs
Low retention (<50%)	25.7% of docs
Panel B: Entity & signal preservation (raw→SR, paired)	
Numerical values preserved	91.0%
Acronyms preserved	92.5%
Named entities preserved	72.7%
Extraneous spans (hallucinated)	1.3%
Panel C: NLI entailment (roberta-large-mnli, 5,589 pairs)	
Entailment	4,061 (72.7%)
Neutral	1,286 (23.0%)
Contradiction	242 (4.3%)

summaries. Panel B shows strong preservation of atomic signals (numerical values 91.0% and acronyms 92.5%); named entity rates drop to 72.7%, reflecting repeated mentions collapsing under compression. Panel C reveals a 4.3% contradiction rate (95.7% NLI-consistent); 181 of 416 documents have at least one contradiction, concentrated in formal/mathematical content where declarative reformulation flattens conditional or quantified statements.

Taken together, these three audits define the quality-fidelity profile of SR. For roughly 60% of documents, SR delivers substantially denser text at negligible factual cost. For the remaining documents, the primary risk is omission under compression, not fabrication. Therefore, SR is best suited to declarative, evidence-dense documents while domains rich in procedural reasoning or formal logic require caution.

Operational Cost PRIME (offline) consumes zero query-time LLM tokens and matches structure-aware chunking in retrieval quality (Retrieval Quality, overlapping 95% CIs). SR adds a one-time preprocessing cost of 56K tokens/doc including per-section prompt overhead (Semantic Restructuring). Against LLM pointwise reranking (~1,926 tokens/query per 20-candidate batch), SR breaks even at **12,096 queries**, roughly 8.3× QASPER’s test set, after which it too operates at zero per-query LLM cost. Full resource accounting tables, per-dataset cost breakdowns, and break-even analysis are in Appendix D.6.

4.2 Ablation

The 5×3 factorial design of Table 2 constitutes a component-wise ablation, isolating index representations under fixed retrieval pipelines. Under Dense-only, PRIME (Full) leads at 0.214 (+22.2% over raw, $p < 0.001$); PRIME (offline) matches structure-chunks (0.201 vs. 0.201). Under Hybrid+CE, SR-based variants drop below raw (0.325–0.328 vs. 0.391, $p < 0.001$), confirming that CE reranking on unmodified H2 sections saturates document-level signal—index-side gains primarily manifest at the context level (Context Quality).

5 Discussion

Our results reveal that index-side optimization is effective at the context level. However, the fact that deterministic metadata matches structured chunking while SR’s QA gains flow primarily through improved retrieval rather than context substitution points to a broader issue: retrieval metrics alone are an incomplete lens on corpus-side preprocessing. Additionally, the three-regime pattern across BRIGHT points to a broader implication. Index-side interventions are not universally beneficial, their utility is predictable from corpus properties measurable before retrieval. This shifts the evaluation focus from absolute improvement to corpus-conditioned utility, precisely the dimension that standard IR benchmarks, averaged over heterogeneous corpora, cannot resolve. Four open problems follow. SR’s quality-fidelity trade-off raises the question of when rewriting can be trusted without verification; its break-even cost (Operational Cost) motivates lighter restructuring strategies without losing information; the interference on heading-free corpora (Retrieval Quality, NFCorpus) suggests corpus-adaptive strategies; and our text-only pipeline is inherently limited to textual evidence — table and figure content in QASPER is not indexed. These call for a principled theory of index-side optimization conditioned on document structure, rather than empirical heuristics.

6 Conclusion

We presented PRIME, an index-side optimization framework combining LLM-driven Semantic Restructuring with deterministic Metadata Synthesis and Hierarchical Retrieval. Deterministic metadata matches structure-aware chunking at 45.8× less embedding computation and 22.4× fewer indexed vectors. On top of these gains, Semantic Restructuring further improves Answer-F1 by 13.7% in full deployment, with characterized quality–fidelity–cost trade-offs. Through extensive evaluation, we identified the operating conditions and cost profiles of this framework. Future work includes improving SR fidelity under text restructuring in order to fulfill its retrieval potential.

References

- Bajaj, P.; Campos, D.; Craswell, N.; Deng, L.; Gao, J.; Liu, X.; Majumder, R.; McNamara, A.; Mitra, B.; Nguyen, T.; et al. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.
- Bhat, S. R.; Rudat, M.; Spiekermann, J.; and Flores-Herr, N. 2025. Rethinking chunk size for long-document retrieval: A multi-dataset analysis. *arXiv preprint arXiv:2505.21700*.
- Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.
- Boteva, V.; Gholipour, D.; Sokolov, A.; and Riezler, S. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, 716–722. Springer.
- Chen, B.; Guo, Z.; Yang, Z.; Chen, Y.; Chen, J.; Liu, Z.; Shi, C.; and Yang, C. 2026a. PathRAG: Pruning Graph-based Retrieval Augmented Generation with Relational Paths. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 30183–30191.
- Chen, H.; Yang, Y.; Li, Y.; Zhang, M.; Hu, B.; and Zhang, M. 2026b. Beyond Chunking: Discourse-Aware Hierarchical Retrieval for Long Document Question Answering. In Liakata, M.; Moreira, V. P.; Zhang, J.; and Jurgens, D., eds., *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 18176–18198. San Diego, California, United States: Association for Computational Linguistics. ISBN 979-8-89176-390-6.
- Cormack, G. V.; Clarke, C. L.; and Buettcher, S. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 758–759.
- Dasigi, P.; Lo, K.; Beltagy, I.; Cohan, A.; Smith, N. A.; and Gardner, M. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4599–4610.
- Dong, K.; Deik, D. G. X.; Lee, Y. Q.; Zhang, H.; Li, X.; Zhang, C.; and Liu, Y. 2024. Multi-view content-aware indexing for long document retrieval. *arXiv preprint arXiv:2404.15103*.
- Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Formal, T.; Piwowarski, B.; and Clinchant, S. 2021. 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*, 2288–2292.
- Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1762–1777.
- Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval Augmented Language Model Pre-Training. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 3929–3938. PMLR.
- Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; and Grave, E. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43.
- Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 6769–6781.
- Khattab, O.; and Zaharia, M. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 39–48.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.
- Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.
- Malkov, Y. A.; and Yashunin, D. A. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4): 824–836.
- Manning, C. D. 2008. *Introduction to information retrieval*. Syngress Publishing,.
- Nogueira, R.; and Cho, K. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.
- Nogueira, R.; Yang, W.; Lin, J.; and Cho, K. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.
- Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 3982–3992.

- Reuter, M.; Lingenberg, T.; Liepina, R.; Lagioia, F.; Lippi, M.; Sartor, G.; Passerini, A.; and Sayin, B. 2025. Towards Reliable Retrieval in RAG Systems for Large Legal Datasets. In Aletras, N.; Chalkidis, I.; Barrett, L.; Goanță, C.; Preotiuc-Pietro, D.; and Spanakis, G., eds., *Proceedings of the Natural Legal Language Processing Workshop 2025*, 17–30. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-338-8.
- Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, 60: 503–520.
- Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4): 333–389.
- Salton, G.; and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5): 513–523.
- Sarathi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; and Manning, C. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, 32628–32649.
- Singh, S.; Reddy, A.; and Chopra, M. 2026. Beyond the Reranker: Do RAG Retrieval Enhancements Help Once a Strong Reranker Is Present? *arXiv preprint arXiv:2606.28367*.
- SU, H.; Yen, H.; Xia, M.; Shi, W.; Muennighoff, N.; yu Wang, H.; Haisu, L.; Shi, Q.; Siegel, Z. S.; Tang, M.; Sun, R.; Yoon, J.; Arik, S. O.; Chen, D.; and Yu, T. 2025. BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval. In *The Thirteenth International Conference on Learning Representations*.
- Sun, W.; Yan, L.; Ma, X.; Wang, S.; Ren, P.; Chen, Z.; Yin, D.; and Ren, Z. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 14918–14937.
- Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; and Gurevych, I. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Tombros, A.; and Sanderson, M. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, 2–10. New York, NY, USA: Association for Computing Machinery. ISBN 1581130155.
- Wang, S.; Zhang, H.; Li, J.; and Liu, Y. 2026. PaperRegister: Hierarchical register indexing for academic document retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*.
- Wilkinson, R. 1994. Effective retrieval of structured documents. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, 311–317. Springer.
- Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; and Nie, J.-Y. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 641–649.
- Xiong, L.; Xiong, C.; Li, Y.; Tang, K.-F.; Liu, J.; Bennett, P.; Ahmed, J.; and Overwijk, A. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.
- Xu, A.; Lin, B.; Xue, B.; Wang, B.; Xu, B.; Wu, B.; Zhang, B.; Lin, C.; Dong, C.; Ling, C.; et al. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.
- Zhang, Y.; Wu, R.; Cai, P.; Wang, X.; Yan, G.; Mao, S.; Wang, D.; and Shi, B. 2026. Leanrag: Knowledge-graph-based generation with semantic aggregation and hierarchical retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 34862–34869.