Context at What Cost? A Cross-Model Study of Chunk Enrichment for Retrieval in RAG

Anonymous ACL submission

Abstract

Contextual Retrieval (CR) is widely adopted on the assumption that prepending LLM-generated context to each chunk improves retrieval. We present a cross-model study of this assumption across five BEIR datasets, seven embedding models, and a Late Chunking baseline. The main result is strongly model-dependent: causal last-token models are significantly harmed by CR on most datasets, the strongest native-bidirectional evidence is near-neutral, and retrofitted bidirectional models occupy an observed intermediate regime. Same-backbone pooling controls, LLM2Vec, and an exploratory attention probe show that directional aggregation co-varies with CR sensitivity while prefix-attention mass alone is insufficient. We synthesize the findings into a three-component mechanism view, a 4-step deployment decision rule, and a public diagnostic tool (`cr-diagnostic`) that estimates an embedder's CR sensitivity in roughly one GPU-hour. Within the tested Qwen-based pipeline paired with Qwen3-Embedding-8B, context-generation scale beyond 8B provides no measurable retrieval benefit.

1 Introduction

Retrieval-Augmented Generation systems (RAG) (Lewis et al., 2020) split documents into retrievable chunks. A key challenge is that individual chunks often lack context—a passage mentioning "the company" without naming it, or discussing "these results" without specifying the experiment, becomes difficult to match against user queries. This *decontextualization problem* motivated several chunk enrichment strategies.

Anthropic's Contextual Retrieval (CR) (Anthropic, 2024) addresses this by using an LLM to generate a brief contextual summary and prepending it to each chunk before embedding, reporting up to a 49% reduction in top-20 retrieval failure rate (when Contextual Embeddings are combined with Contextual BM25) across heterogeneous domains (codebases, fiction, ArXiv and science papers) without per-embedder or per-domain breakdowns. Late Chunking (Günther et al., 2024) takes a different approach: it encodes full documents through a long-context embedding model and extracts per-chunk embeddings from the contextualized representations, preserving document-level context without additional LLM calls. Both strategies have seen rapid adoption in production RAG systems, but neither claim has been replicated across embedding-model classes under a shared protocol.

Despite this adoption, no study has compared how these enrichment strategies interact with embedding-model choice under a shared experimental protocol. This interaction is critical: modern embedding models range from decoder-only LMs repurposed for embedding (Wang et al., 2024; Muennighoff et al., 2025) to natively bidirectional encoders (Chen et al., 2024), and recent work has shown that decoder-derived embedders are sensitive to input perturbations such as token prepending (Fu et al., 2025). Yet no study has tested whether CR's prefix prepending triggers these sensitivities consistently across model classes, or whether pooling and bidirectional retrofitting materially change the outcome. Merola and Singh (2025) compared CR and Late Chunking but used a single embedding model without statistical testing, leaving the cross-model question unaddressed. We additionally show that a coarse "causal vs. bidirectional" split is informative but incomplete: targeted controls indicate that pooling choice and retrofitted bidirectionality materially shape CR sensitivity even within the same backbone family, and an exploratory attention probe weakens the simple "more prefix attention implies more harm" account. These are patterns observed across seven models from limited backbone families, not a complete taxonomy.

Contributions.

1. A **cross-model evaluation of CR on five BEIR datasets** with statistical testing: CR significantly degrades causal-LM-based embedders on most datasets ($p < 0.05$ on 3–5 of 5 datasets per model), while the strongest native-bidirectional evidence is near-neutral.

2. Targeted mechanism controls: a same-backbone Qwen3 pooling comparison whose mean-pooling variant flips CR's sign, LLM2Vec as a second retrofitted bidirectional model, and an exploratory attention probe showing prefix-attention mass alone does not explain robustness.

3. A **random prefix ablation** that approximates the gap between CR's semantic enrichment and its prefix-induced structural distortion, applied across five models and five datasets as a practical diagnostic for prepend-based enrichment.

4. A **scale equivalence** result: within Qwen-generated contexts paired with Qwen3-Embedding-8B, context-generation scale beyond 8B yields no measurable retrieval benefit (TOST equivalence at δ=0.03 on 44/44 per-query-nDCG pairs on Qwen3-Embedding-8B; BGE-M3 shows the same directional pattern qualitatively but is not subjected to TOST); an append-vs-prepend and Gemma cross-family check bound the claim's scope.

5. A **three-component mechanism view, a 4-step decision rule, and the `cr-diagnostic` tool** that lets practitioners estimate CR's value on their own embedder in roughly one GPU-hour (§5.2, §6).

6. A Late Chunking evaluation across six architectures (including jina-v3) that exposes a parallel but partly confounded model-dependent pattern.

A supplementary downstream evaluation in Appendix X is broadly consistent with the retrieval findings but is not a primary contribution.

2 Related Work

Chunk Enrichment. Anthropic (2024) proposed CR using Claude as the context-generation model, demonstrating improvements on proprietary benchmarks without BEIR evaluation or cross-model analysis; because the method was released as a technical blog post rather than a refereed paper, we cite it as the original system description rather than as the primary evidential basis for our claims. Anantha et al. (2024) and Jain et al. (2025) provide peer-reviewed evidence that adding or restructuring context can materially affect retrieval quality in RAG systems. Merola and Singh (2025) compared CR and Late Chunking with a single embedding model (text-embedding-3-small), finding CR superior. A recent concurrent preprint, Shaukat et al. (2026), shows that embedding sensitivity to chunking strategy is substantial; we use it as recent context rather than as the primary basis for our claims.

Late Chunking. Günther et al. (2024) introduced Late Chunking for bidirectional models with mean pooling, demonstrating that encoding full documents preserves cross-chunk context. Their evaluation was limited to a single Jina embedding model (jina-embeddings-v2). We extend evaluation to six models including jina-embeddings-v3 (Sturua et al., 2024), which incorporates Late Chunking support during training, alongside causal architectures and models not trained for Late Chunking, revealing dramatic model-specific variation that the original work could not observe.

Decoder-Derived Embedders. A growing body of work converts causal LMs into embedding models (Wang et al., 2024; Muennighoff et al., 2025; BehnamGhader et al., 2024). Tang and Yang (2024) showed that last-token pooling dominates on standard benchmarks; Lee et al. (2025a) identified recency bias in last-token pooling and proposed latent attention. Fu et al. (2025) further show that decoder embedders are highly sensitive to prepended tokens. Together, these works motivate our same-backbone pooling control, which asks whether CR sensitivity is tied only to model family or more specifically to directional aggregation.

Hybrid Search. Bruch et al. (2024) showed convex combination outperforms RRF with proper tuning, while recent peer-reviewed RAG systems such as HybGRAG (Lee et al., 2025b) continue to rely on hybrid retrieval in practice. We include hybrid search mainly to test whether the same model dependence persists when sparse and dense evidence are fused.

3 Methodology

3.1 Experimental Setup

We evaluate on five BEIR (Thakur et al., 2021) datasets: SciFact (5,248 chunks), NFCorpus (3,666), FiQA (58,363), ArguAna (8,703), and SciDocs (26,070). Documents are chunked at 800 tokens with 200-token overlap.

Chunk enrichment strategies. (1) **Raw**: no enrichment (baseline). (2) **Title**: source document title prepended to each chunk (zero LLM cost). (3) **CR-8B/27B/397B**: LLM-generated context prepended to each chunk, using Qwen3-8B (Qwen Team, 2025), Qwen3.5-27B (Qwen Team, 2026a), and Qwen3.5-397B-A17B (MoE, 17B active) (Qwen Team, 2026b); see Appendix A for the prompt template. (4) **Late Chunking**: full-document encoding with per-chunk embedding extraction via tokenizer offset mapping.

Embedding models. We select seven models spanning two pretraining paradigms—causal LM foundations vs. natively bidirectional encoders—and a range of baseline quality levels (Table 1). Critically, GritLM-7B bridges the two categories: pretrained as a causal LM (Mistral-7B) but using bidirectional attention at inference via a 4D attention mask, enabling us to test whether inference-time attention or pretraining architecture governs CR sensitivity.

Targeted controls. To refine the main cross-model comparison, we add targeted follow-up experiments. First, we compare Qwen3-Embedding-8B with its standard last-token pooling against a mean-pooling variant using the same backbone. Second, we evaluate LLM2Vec (BehnamGhader et al., 2024) as an additional retrofitted bidirectional model. Third, we compare appending generated context against prepending it for Qwen3. Fourth, we report an NQ subset spot check and a Nemotron-exclusion robustness analysis. Finally, we include two additional checks: a prepend-based CR spot check on jina-v3 and a partial multi-seed robustness analysis for the random-prefix ablation. We treat these additions as targeted diagnostic controls rather than primary benchmark axes.

Search methods. BM25, Vector (FAISS), Hybrid-RRF (Cormack et al., 2009) (k=60), and Hybrid-CC ($\alpha \cdot s_{\text{vec}} + (1-\alpha) \cdot s_{\text{BM25}}$; $\alpha = 0.5$ by default, with optimal α^* determined by 5-fold CV in §4.3). For each fold, α^* is selected on the fold's held-out queries and then averaged across folds, so the reported test score for a given α is never tuned on that same query set.

Table 1: Embedding models evaluated. Models above the rule are pretrained as causal LMs; those below are natively bidirectional encoders. GritLM-7B (‡) uses bidirectional attention at inference despite causal LM pretraining. †jina-v3 is trained for Late Chunking. "Used Len" is the maximum sequence length we feed during chunk embedding; some models (e.g., Qwen3-Emb-8B, jina-v3) support longer native contexts but are run at the listed length for cross-model parity. Late Chunking uses extended lengths (Appendix B).

Model	Params	Dim	Used Len	Pooling
Qwen3-Emb-8B(Zhang et al., 2025)	8B	4096	512	Last-token
E5-Mistral-7B(Wang et al., 2024)	7B	4096	4096	Last-token
Qwen3-Emb-0.6B(Zhang et al., 2025)	0.6B	1024	512	Last-token
GritLM-7B‡(Muennighoff et al., 2025)	7B	4096	4096	Mean (bidir.)
Nemotron-8B(Babakhin et al., 2025)	7.5B	4096	4096	Mean (bidir.)
BGE-M3(Chen et al., 2024)	568M	1024	512	CLS (bidir.)
jina-v3†(Sturua et al., 2024)	570M	1024	8192	Mean (bidir.)

Reranking. We use Qwen3-Reranker-8B (Zhang et al., 2025) (primary) and BGE-Reranker-v2-M3 (568M, XLM-RoBERTa; validation) to avoid single-family bias.

Evaluation. Retrieval operates at chunk level, but evaluation is at document level: when multiple chunks from the same document appear in the top-k, we retain only the maximum score per document, then compute nDCG@10 against BEIR's document-level relevance judgments. This max-pooling strategy mirrors standard IR practice—a document is considered retrieved if any of its chunks scores sufficiently high. We do not report chunk-level relevance because BEIR provides document-level judgments; a chunk-level evaluation would therefore require additional annotation and would not be directly comparable to the benchmark. Retrieval-side statistical significance uses paired permutation tests (n=10,000) and bootstrap 95% CIs (n=10,000). For the supplementary downstream evaluation (Appendix X), FiQA Raw vs. CR-8B comparisons are tested with paired randomization tests over query-matched judge scores, SciFact Raw vs. CR-8B comparisons with McNemar's exact test on paired query-level correctness, and 95% bootstrap CIs are reported for both tasks; we apply Benjamini–Hochberg correction across the 16 downstream model$\times$metric comparisons and make inferential claims only for effects that survive correction.

Infrastructure and reproducibility. 8× NVIDIA H200. Chunks use an 800-token ceiling but are typically much shorter (FiQA median 119); for 512-used-length models, 2.3%–9.9% of raw chunks truncate, and a non-truncated-subset analysis confirms truncation is not the driver (Appendix R). Context generation is deterministic (temperature 0). An anonymized supplementary repository implementing the random-prefix diagnostic accompanies this submission; full code and pipeline will be released publicly after review anonymity constraints are lifted. Full details in Appendix B–Q.

4 Results

4.1 CR Impact Across Embedding Models

Table 2 presents vector search nDCG@10 across the four primary embedding models and five BEIR datasets; Table 3 extends the evaluation with three additional models that probe edge cases (low-quality causal, retrofitted bidirectional, and LC-trained bidirectional). The central empirical finding is that all four causal-LM-based models are **significantly degraded by CR on most datasets** ($p < 0.05$ on 3–5 of 5 datasets each), while the strongest native-bidirectional evidence is near-neutral. In §5.1, we introduce a random prefix ablation that helps diagnose this pattern by approximating the contrast between CR's semantic enrichment and its structural distortion.

Causal-LM-based models are hurt on most datasets. All four causal-LM-foundation models are significantly degraded by CR on 3–5 of 5 datasets each ($p < 0.05$), spanning 0.6B to 8B parameters; the pattern is therefore not driven by overall model quality. The remaining cells (e.g., Qwen3 / NFCorpus: $+0.004$, $p = 0.40$; Qwen3-0.6B / NFCorpus: $+0.002$) are non-significant near-zero effects rather than positive evidence. GritLM-7B is also hurt (SciFact -0.018, FiQA -0.039, ArguAna -0.031; $p < 0.01$) despite bidirectional inference, though with smaller magnitudes—§5.1 quantifies this partial mitigation. Late Chunking degrades Qwen3-Embedding-8B (-0.19 mean) and E5-Mistral (-0.39) heavily but is partly confounded with context-window mismatch (§4.4). E5-Mistral uses a generic query instruction; Appendix V shows that switching to task-specific instructions does not change the direction of CR's effects.

Bidirectional evidence is narrower and more qualified. BGE-M3 and Nemotron-8B show near-neutral CR effects (mean $|\Delta| < 0.01$) but are not equal-strength evidence. BGE-M3 has stable baselines across all five datasets and small CR deltas, with one significant but modest degradation on ArguAna (-0.020), making it the clearest native-bidirectional case but not evidence of universal improvement. Nemotron is supportive but caveated: instruction-prefix dominance in mean pooling produces anomalously low baselines on short-query datasets (FiQA 0.029), and its small significant changes include both gains and losses; its non-pathological cells (SciFact -0.002, ArguAna -0.009, SciDocs $+0.014$) remain directionally consistent. A prepend-CR check on jina-v3 shows small degradations on all five datasets (-0.003 to -0.021), indicating that native bidirectionality alone does not guarantee prepend robustness. The GritLM contrast is informative: bidirectional inference does not erase the inherited causal-LM vulnerability.

Figure 1 summarizes this descriptively: model-level Spearman $\rho = -0.94$ ($p = 0.005$, n=6); point-level $\rho = -0.60$ ($p < 0.001$, n=30). Because the model-level view uses only six aggregates and points cluster by model, we treat these correlations as descriptive summaries rather than standalone inferential evidence. Causal-LM-based models cluster below zero regardless of inference-time pooling; the strongest native-bidirectional evidence remains near zero; GritLM is intermediate. Minor exceptions (e.g., Qwen3/NFCorpus: $+0.004$, $p = 0.40$) likely reflect NFCorpus's short, dense abstracts already carrying sufficient intra-chunk context. The trend is not driven by Nemotron alone: excluding it still yields $\rho = -0.90$ ($p = 0.037$, n=5).

BM25 nearly always benefits. In contrast to dense retrieval, CR-8B and CR-27B improve BM25 in all 10/10 dataset–strategy cells we evaluate, with the largest relative gain ($+14.8\%$) on FiQA from CR-397B; the only exception is CR-397B on SciFact ($\Delta = -0.002$, within noise), so we characterize BM25 as nearly always benefiting from CR rather than uniformly so. This asymmetry reflects a fundamental difference: CR adds domain-specific keywords valuable for lexical matching but introduces distributional noise for embedders that already capture these semantics implicitly, consistent with classic document-expansion and lexical-

Table 2: Vector search nDCG@10 (no reranking). Causal-LM models are hurt by CR; bidirectional encoders are robust. See Table 3 for GritLM, Qwen3-0.6B, jina-v3.

		Causal LM foundations				**Bidirectional encoders**			
		Qwen3-8B		E5-Mistral		Nemotron		BGE-M3	
Dataset	Strategy	nDCG	Δ	nDCG	Δ	nDCG	Δ	nDCG	Δ
SciFact	Raw	**.796**	—	**.773**	—	.633	—	**.645**	—
	CR-8B	.775	−.021	.752	−.021	.631	−.002	.644	−.001
	CR-27B	.782	−.014	.736	−.037	**.669**	+.036	.641	−.004
	Late	.595	−.201	.159	−.614	.734	+.101	.580	−.065
NFCorpus	Raw	.407	—	**.400**	—	.053	—	.317	—
	CR-8B	**.411**	+.004	.389	−.011	.066	+.013	**.318**	+.001
	CR-27B	.398	−.009	.385	−.015	.064	+.011	.318	+.001
	Late	.280	−.127	.076	−.324	**.157**	+.104	.284	−.033
FiQA	Raw	**.612**	—	**.533**	—	.029	—	**.411**	—
	CR-8B	.509	−.103	.476	−.057	.033	+.004	.399	−.012
	CR-27B	.516	−.096	.495	−.038	**.035**	+.006	.415	+.004
	Late	.290	−.322	.005	−.528	.029	+.000	.348	−.063
ArguAna	Raw	**.510**	—	**.454**	—	.450	—	**.399**	—
	CR-8B	.474	−.036	.444	−.010	.441	−.009	.379	−.020
	CR-27B	.484	−.026	.445	−.009	.450	+.000	.392	−.007
	Late	.321	−.189	.164	−.290	**.491**	+.041	.377	−.022
SciDocs	Raw	**.266**	—	**.208**	—	.030	—	.164	—
	CR-8B	.260	−.006	.194	−.014	.045	+.015	**.168**	+.004
	CR-27B	.259	−.007	.201	−.007	.038	+.008	.166	+.002
	Late	.162	−.104	.001	−.207	**.060**	+.030	.152	−.012

Table 3: Additional models (vector, no reranking). Qwen3-0.6B confirms quality-independent CR degradation. GritLM-7B shows partial degradation despite bidirectional inference.

		Qwen3-0.6B		GritLM-7B		jina-v3	
Dataset	Strategy	nDCG	Δ	nDCG	Δ	nDCG	Δ
SciFact	Raw	**.701**	—	**.789**	—	**.730**	—
	CR-8B	.696	−.005	.771	−.018	.727	−.003
	Late	—	—	.789	+.000	.689	−.041
NFCorpus	Raw	.358	—	.403	—	**.371**	—
	CR-8B	**.360**	+.002	**.405**	+.003	.360	−.011
	Late	—	—	.398	−.005	.343	−.028
FiQA	Raw	**.464**	—	**.604**	—	**.477**	—
	CR-8B	.426	−.038	.565	−.039	.456	−.021
	Late	—	—	.599	−.006	.450	−.026
ArguAna	Raw	**.478**	—	.431	—	**.357**	—
	CR-8B	.449	−.030	.401	−.031	.353	−.003
	Late	—	—	**.433**	+.002	.348	−.009
SciDocs	Raw	**.225**	—	.222	—	**.205**	—
	CR-8B	.217	−.008	.219	−.003	.195	−.010
	Late	—	—	**.225**	+.003	.203	−.002

recall effects.

Model scale within the tested setting. Paired permutation tests between CR-8B, CR-27B, and CR-397B show no significant differences ($p > 0.05$) on any of the 44 dataset–method combinations actually tested (Appendix E; the nominal grid is 5 datasets $\times$ 4 methods $\times$ 3 pairs $=$ 60 cells, but CR-397B is available only for SciFact, NFCorpus, and FiQA, reducing the realised count to 44). This comparison should be interpreted narrowly: CR-397B is available only on three datasets, and it is an MoE model with 17B active parameters. To go beyond "absence of difference" to positive evidence of equivalence, we apply TOST (Two One-Sided Tests) on per-query nDCG@10 for Qwen3-Embedding. We choose $\delta = 0.03$ nDCG@10 because it corresponds to roughly 4–5% relative change at the 0.6–0.8 baseline range we observe and is below the smallest gains typically claimed as meaningful on BEIR-scale leaderboards; we also report the tighter $\delta = 0.02$ as a stricter check. At $\delta = 0.03$, 44/44 tested pairs achieve equivalence ($p < 0.05$); at $\delta = 0.02$, 34/44 (77%) do. Within this evaluated Qwen-based pipeline and with Qwen3-Embedding-8B as the embedder, an 8B context-generation model is sufficient; we do not claim that 8B is universally sufficient across embedders, context LLM families, or corpora. Full per-method breakdowns are in Appendix C (Qwen3-Embedding) and Appendix D (BGE-M3, where the same direction holds qualitatively but with smaller absolute differences).

LLM family independence. Regenerating contexts on three datasets (SciFact, FiQA, ArguAna) with Gemma-3-12B-it and re-evaluating yields identical directional effects: Gemma contexts degrade Qwen3-Embedding by -0.019 to -0.068 and GritLM by -0.008 to -0.043, matching the Qwen-context pattern. CR's model-dependent effect is therefore primarily a property of the embedder rather than of the context-generation family.

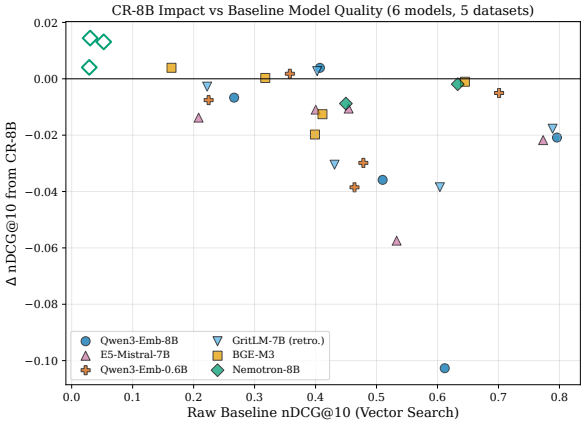

Figure 1: CR-8B impact (Δ nDCG@10 from raw, y-axis) vs. raw vector-search baseline (x-axis), one point per (model, dataset) pair, 6 models × 5 BEIR datasets = 30 points. We omit jina-v3 from this figure because its primary intended use is Late Chunking rather than prepend-CR; jina-v3 prepend-CR deltas are reported separately in Appendix K. Causal-LM-foundation last-token models (Qwen3-Emb-8B, E5-Mistral-7B, Qwen3-Emb-0.6B; circles/triangles/plus) and the retrofitted bidirectional GritLM-7B (down-triangle) cluster below the zero line; native bidirectional encoders BGE-M3 (square) and Nemotron-8B (diamond) sit near zero. Open green diamonds mark Nemotron's anomalously weak FiQA and NFCorpus baselines (raw < 0.1), included as caveated supporting evidence only. Spearman correlations: model-level $\rho = -0.94$ (p=0.005, n=6); point-level $\rho = -0.60$ (p<0.001, n=30); the trend is meant descriptively rather than as standalone inferential evidence.

4.2 Pooling and Retrofitted Bidirectionality

Table 4 tightens the interpretation. For the *same* Qwen3 backbone, switching from last-token to mean pooling flips CR's sign on all three targeted datasets (FiQA Δ: $-0.103 \rightarrow +0.015$; SciFact $-0.022 \rightarrow +0.011$; ArguAna $-0.034 \rightarrow +0.006$). Mean pooling also lowers the absolute baseline (e.g., FiQA Raw $0.612 \rightarrow 0.094$), so the sign reversal cannot be attributed to pooling alone; we read it as evidence that CR sensitivity is governed by pooling-and-aggregation choices in conjunction with model family rather than by family membership alone.

The intermediate regime is also no longer unique to GritLM. LLM2Vec, a second retrofitted bidirectional model built from a causal base, shows near-neutral mixed effects: $+0.017$ on FiQA, -0.006 on SciFact, and -0.013 on ArguAna. This profile is noticeably less harmful than the strongly negative last-token models, but less robust than the strongest native bidirectional encoder. Together with GritLM, it supports treating the retrofitted bidirectional case as an *observed intermediate regime* rather than a one-model anecdote. An exploratory attention probe in Appendix L points in the same direction mechanistically: Nemotron allocates the most attention to the CR prefix yet remains near-neutral, whereas GritLM allocates far less prefix attention but still degrades, so vulnerability cannot be explained by prefix-attention mass alone.

Table 4: Targeted controls for CR sensitivity on three datasets. Cells report Raw nDCG@10 and Δ (CR-8B minus Raw). The same Qwen3 backbone changes sign under mean pooling, while retrofitted bidirectional models (GritLM, LLM2Vec) occupy an intermediate regime. Numbers come from a controlled re-evaluation pass executed on the same fixed query/corpus splits but in a separate batch, and may therefore differ by ≤ 0.002 from the corresponding cells in Tables 2–3 due to floating-point non-determinism in batched inference (e.g., Qwen3 last-token ArguAna Raw: 0.512 here vs. 0.510 in Table 2). The sign and order-of-magnitude of every Δ reported here is unchanged when the main-run Raw values are used as the reference. Because the mean-pooling variant also changes the raw baseline substantially (e.g., FiQA $0.612 \rightarrow 0.094$), this is strong but not definitive evidence about mechanism.

Model	**FiQA**		**SciFact**		**ArguAna**	
	Raw	Δ	Raw	Δ	Raw	Δ
Qwen3 (last-token)	.612	−.103	.796	−.022	.512	−.034
Qwen3 (mean pool)	.094	+.015	.631	+.011	.406	+.006
GritLM	.604	−.039	.789	−.018	.431	−.031
LLM2Vec	.183	+.017	.527	−.006	.318	−.013

4.3 Reranking and Hybrid Search

Reranking exhibits a **baseline-quality-correlated pattern** (Table 8): it hurts high-quality models on most datasets (Qwen3 -0.041 averaged across datasets, GritLM -0.001 but with large per-dataset variation: -0.093 on FiQA, $+0.124$ on ArguAna), is near-neutral for E5-Mistral ($+0.004$), and helps lower-quality models (Qwen3-0.6B $+0.027$, jina-v3 $+0.052$, BGE-M3 $+0.073$, Nemotron $+0.022$). The qualitative direction is consistent with reranking complementing weaker first-stage retrievers and competing with stronger ones, although part of the strong-baseline penalty may also reflect query/passage instruction-format match between the embedder and the reranker (Appendix W).

Validation with BGE-Reranker-v2-M3 (568M, XLM-RoBERTa) reproduces the same overall pattern (helping weaker models and hurting stronger

ones) on the three models for which we have full coverage with both rerankers (Qwen3-Embedding-8B, BGE-M3, E5-Mistral-7B); per-(model,dataset) sign-direction matches on 10/15 vector conditions, with the disagreements concentrated on cases where BGE-Reranker's smaller capacity produces large negative deltas absent from Qwen3-Reranker. On BM25, reranking helps uniformly across the seven models we evaluate ($+0.022$ to $+0.093$). Full results in Appendix F.

α sensitivity. Hybrid-CC combines vector and BM25 scores via $\alpha \cdot s_{\text{vec}} + (1-\alpha) \cdot s_{\text{BM25}}$. The 5-fold-CV optimum α^* varies widely across six models, from 0.94 (Qwen3) to 0.38 (Nemotron); the common default $\alpha = 0.5$ is near-optimal only for BGE-M3 and can cost up to -0.05 nDCG@10 for models at either extreme. A descriptive linear fit $\alpha^* \approx 0.89 \cdot \bar{b} + 0.33$ ($R^2 = 0.87$, $n = 6$) captures the rough trend but is too thinly sampled to act as a tuning rule. Combined with the stable BM25 gains from CR, hybrid retrieval may remain worthwhile even when dense-only CR is harmful, provided α is validated rather than fixed (full sweep in Appendix P, Figure 7).

4.4 Late Chunking under Unequal Context-Window Stress

Late Chunking exposes the same model dependence under partly confounded context-window stress: 512-used-length models (Qwen3-Emb-8B, BGE-M3) are pushed past their used regime, while jina-v3 (8192), Nemotron (4096), and E5-Mistral (4096) operate near or within their native ranges. Magnitudes co-vary with this stratification and with pooling: E5-Mistral -0.39 on average, Qwen3 -0.19, BGE-M3 -0.04, Nemotron $+0.06$, GritLM near-neutral (-0.001). Even jina-v3 (Sturua et al., 2024), trained with Late Chunking support, shows small degradations on all five datasets versus its chunk-level baseline (-0.002 to -0.041). The qualitative pattern—causal last-token pooling fares worst, mean-pooled bidirectional encoders fare best—mirrors our CR findings; we do not claim a clean separation of LC from context-window stress (Appendix O–P).

4.5 Counter-intuitive findings

Three observations cut against common expectations. **(i) Shorter chunks do not benefit more from CR**: Qwen3 on FiQA degrades by -0.071 at chunk length ≤ 256 versus -0.102 at ≤ 800 tokens (Appendix S), refuting the "CR rescues short chunks" framing and supporting a structural-distortion reading. **(ii) LC-specific training does not confer prepend robustness**: jina-v3, trained with Late Chunking support, still degrades under prepend-CR on all five datasets (-0.003 to -0.021; Appendix K), so the CR and LC interventions cannot be assumed to share robustness profiles. **(iii) Reranking harms high-quality first-stage retrievers**: on vector + raw, Qwen3-Reranker-8B *lowers* nDCG@10 by -0.041 on Qwen3-Embedding-8B while improving weaker first-stage models by $+0.022$ to $+0.073$ (Table 8); BGE-Reranker-v2-M3 reproduces the direction, ruling out reranker-family artifacts.

Significance summary. Detailed permutation tests and bootstrap confidence intervals are reported in Appendix H. In brief, the main dense-retrieval findings are statistically stable: CR-8B significantly degrades the primary causal last-token models on most datasets, BGE-M3 and Nemotron show small mixed effects overall, and Late Chunking strongly degrades the primary last-token models.

Supplementary downstream analysis. A downstream evaluation on FiQA and SciFact is reported in Appendix X. It is broadly consistent with the retrieval-side findings, with the clearest supported effect being a drop in FiQA judged Correctness for Qwen3 under CR-8B. For SciFact, we interpret the weaker propagation as a property of claim verification: small retrieval perturbations can change which evidence chunks are retrieved without changing the final class label, especially when the top-ranked evidence is redundant enough to preserve the final verdict.

5 Analysis

5.1 Separating Semantic Enrichment from Structural Distortion

The results in §4 show that CR helps some models and hurts others, but do not fully explain *why*. We hypothesize that CR exerts two opposing effects: **semantic enrichment** (the context adds useful information) and **structural distortion** (the prefix perturbs the model's learned input distribution). To approximate this contrast, we introduce a *random prefix* (RP) ablation: for each chunk, we sample a passage from a random chunk in the same dataset, truncate it to match CR's word count (seed 42), and prepend it. RP preserves positional pertur-

bation while removing CR-specific semantics, so the CR$-$RP gap isolates the value CR carries *beyond* length-matched in-domain noise. RP itself can carry domain vocabulary, so the gap is a lower bound on semantic value rather than a clean estimate; the shuffled-prefix control (Appendix T) separates lexical from syntactic contributions.

Multi-seed RP variance is small (std. 0.002–0.007; Appendix K); a shuffled-prefix control (Appendix T) shows BGE-M3 extracts value from shuffled domain vocabulary alone, while Qwen3/FiQA shows CR $\approx$ RP $\approx$ Shuffled, indicating no recoverable enrichment for that cell.

Three observed behaviors. CR outperforms RP on 23/25 conditions (Appendix G); the CR$-$RP gap reveals *how much* semantic value each model extracts. **(1) Indistinguishable (causal last-token):** Qwen3 on FiQA gives CR $\approx$ RP ($\Delta{=}-0.000$), structural distortion dominating; on other datasets Qwen3 and E5-Mistral show modest CR$-$RP gaps ($+0.014$ to $+0.095$) but negative net effect. **(2) Intermediate (retrofitted bidirectional):** GritLM/FiQA shows CR $\gg$ RP ($\Delta{=}+0.100$, recovering $\sim$70% of RP's -0.138), net still negative; LLM2Vec mirrors this. **(3) Near-full compensation (native bidirectional):** BGE-M3/FiQA shows CR $\gg$ RP ($\Delta{=}+0.099$) with near-zero net effect (-0.012). Append-vs-prepend (Appendix J) and prefix-length ablations (Appendix P) further support this stratification: prepend is more harmful than append for Qwen3, and length-dependent degradation is monotonic for causal last-token, intermediate for GritLM, and broadly absent for BGE-M3/Nemotron.

5.2 A predictive claim: prefix-sensitivity is forecastable

Across our nine configurations (seven primary models plus the two Qwen3 pooling variants), two scalar features track CR's effect: raw baseline $\bar{b}$ (model-level Spearman $\rho{=}-0.94$ with CR-8B Δ) and pooling type; the same features also predict the hybrid-CC optimum ($\alpha^* \approx 0.89\,\bar{b} + 0.33$, $R^2{=}0.87$, $n{=}6$). We therefore propose, as a testable hypothesis: **the CR-sensitivity of an unseen embedder can be estimated without running CR by a small *prefix-sensitivity probe***—a $\sim$200-query raw vs. RP comparison combined with pooling type, which jointly matched CR's sign in 25/25 of our (model, dataset) cells. The sample is too small to operationalize here but is the most concrete predictive claim our data supports.

6 Discussion and Conclusion

Mechanism: a three-component view. We synthesize the cross-model, pooling, and ablation evidence into a three-component view of any prepend-based enrichment. Prepending a context simultaneously (i) **shifts position-sensitive aggregation** toward the prefix, (ii) **injects lexical signal** that may be query-bridging, and (iii) **re-distributes self-attention**. Causal last-token embedders are dominated by (i): the pooled state migrates onto inserted tokens, explaining the structural-distortion signature in the random-prefix and prefix-length ablations. BM25 captures (ii) almost in isolation (10/10 cells positive). Native bidirectional CLS/mean encoders absorb (i) via their fixed pooling location and benefit modestly from (ii) when CR carries domain vocabulary (BGE-M3 shuffled-prefix). The attention probe shows (iii) is not decisive: Nemotron allocates the most prefix attention yet is robust; GritLM allocates less yet still degrades. This aligns our findings with Fu et al. (2025)'s decoder-embedder sensitivity to prepended tokens and Lee et al. (2025a)'s last-token recency bias: CR's model-dependent effect is a RAG-scale consequence of known decoder-embedder properties, not a chunking artifact.

Practical decision rule. **(1) Embedder check**: if the production embedder is a causal LM with last-token pooling, expect dense nDCG@10 to drop 0.02–0.10; default to raw chunks. **(2) BM25 layer**: if BM25 or hybrid is in the stack, CR-8B is almost always worth indexing for that layer alone (BM25 gains hold in 10/10 cells). **(3) Hybrid weight**: validate α on a held-out probe; the common $\alpha{=}0.5$ default costs up to -0.05 nDCG@10 for high-quality dense models (α^* range 0.38–0.94). **(4) Scale check**: an 8B context generator is sufficient within a Qwen-style pipeline with Qwen3-Embedding-8B (44/44 TOST equivalence at $\delta{=}0.03$); larger generators add cost without measurable gain.

Conclusion and tool release. CR's effect is strongly model-dependent: causal last-token embedders are harmed, native bidirectional encoders are near-neutral, and within Qwen-style pipelines an 8B context generator suffices. We release `cr-diagnostic`, a tool that estimates an embedder's prefix-sensitivity in $\sim$1 GPU-hour.

Limitations

Our evaluation covers 5 of 18 BEIR datasets, all in English, and at most 58K chunks, so larger or multilingual production settings remain unvalidated. The strongest native-bidirectional evidence comes from BGE-M3; Nemotron and jina-v3 provide only narrower support and boundary cases, so we do not claim broad native-bidirectional coverage. The retrofitted-bidirectional regime is still represented by only two models, so we treat it as an observed pattern rather than a complete taxonomy. The random-prefix ablation is approximate because same-dataset random text can itself carry lexical value. The scaling comparison is also narrow: CR-397B is evaluated on only three datasets and uses an MoE architecture with 17B active parameters. Late Chunking should be interpreted cautiously because context-length mismatch is uneven across models, especially for 512-token embedders. Our downstream analysis is supplementary rather than decisive: it uses one generator, one judge, and limited tasks. We therefore present the paper primarily as retrieval-side evidence with targeted diagnostic controls, not as a complete account of all chunk-enrichment settings.

Ethical Considerations

This work studies retrieval pipelines rather than directly generating user-facing content, but it still has ethical implications. First, our supplementary FiQA evaluation uses a single LLM judge, so judge bias or hidden preference shifts could affect downstream conclusions; we therefore keep those claims narrow. Second, CR-style enrichment adds substantial inference cost and energy use at indexing time, which matters for large deployments; one practical contribution of this paper is showing when that cost is unlikely to be justified. Third, retrieval interventions can amplify or suppress evidence unevenly across domains, so practitioners should validate any enrichment strategy on their own corpus before deployment, especially in high-stakes settings. We use public benchmark and model artifacts, and the supplementary package records the code license and bundled SciFact sample license; we do not collect new human-subject data. In line with the ACL Policy on Publication Ethics, we used large language model assistants for (i) English-language polishing and proofreading of author-written English drafts (comparable in scope to grammar checkers, spell checkers, and dictionaries; usage was limited to sentence- and phrase-level support and did not include whole-passage or whole-draft machine translation), and (ii) coding assistance during implementation in the form of code completion, refactoring, and debugging suggestions on author-designed algorithms and pipelines. The assistants were not used to suggest, draft, or restructure new technical content, nor to design any experimental algorithm or analysis logic. All scientific claims, experimental design, numerical results, the final wording, and the algorithmic logic of the released code are the authors' own work; all code was reviewed and tested by the authors. We accompany this submission with an anonymized supplementary repository to support independent verification, and will release the full pipeline publicly after anonymity constraints are lifted.

References

Raviteja Anantha, Tharun Bethi, Danil Vodianik, and Srinivas Chappidi. 2024. Context tuning for retrieval augmented generation. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 15–22. Association for Computational Linguistics.

Anthropic. 2024. Introducing contextual retrieval. https://www.anthropic.com/news/contextual-retrieval. Blog post, September 2024.

Anthropic. 2026. Claude sonnet 4.6 system card. https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf. System card, February 17, 2026.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling (COLM)*.

Sebastian Bruch, Siyu Gai, and Amir Ingber. 2024. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, 42(1):20:1–20:35.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In

Findings of the Association for Computational Linguistics: ACL 2024, pages 2318–2335. Association for Computational Linguistics.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 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 (SIGIR)*, pages 758–759. ACM.

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. 2025. Token prepending: A training-free approach for eliciting better sentence embeddings from LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3181. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Michael Günther, Isabelle Mohr, Daniel J. Williams, Bo Wang, and Han Xiao. 2024. Late chunking: Contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*.

Arihant Jain, Purav Aggarwal, and Anoop Saladi. 2025. Autochunker: Structured text chunking and its evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 983–995. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626. ACM.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2025b. HybGRAG: Hybrid retrieval-augmented generation on textual and relational knowledge bases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 879–893. Association for Computational Linguistics.

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 (NeurIPS)*, volume 33, pages 9459–9474.

Carlo Merola and Jaspinder Singh. 2025. Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation. *arXiv preprint arXiv:2504.19754*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qwen Team. 2026a. Qwen3.5-27B model card. `https://huggingface.co/Qwen/Qwen3.5-27B`. Hugging Face model card.

Qwen Team. 2026b. Qwen3.5-397B-A17B model card. `https://huggingface.co/Qwen/Qwen3.5-397B-A17B`. Hugging Face model card.

Muhammad Arslan Shaukat, Muntasir Adnan, and Carlos C. N. Kuhn. 2026. A systematic investigation of document chunking strategies and embedding sensitivity. *arXiv preprint arXiv:2603.06976*.

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. *arXiv preprint arXiv:2409.10173*.

Yixuan Tang and Yi Yang. 2024. Pooling and attention: What are effective designs for LLM-based embedding models? *arXiv preprint arXiv:2409.02727*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916. Association for Computational Linguistics.

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. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Context Generation Prompt

The following template is used for all context-generation models (Qwen3-8B, Qwen3.5-27B, Qwen3.5-397B-A17B-FP8), following Anthropic's original design (Anthropic, 2024):

```
<document>
{document}
</document>
Here is the chunk we want to situate
within the overall document:
<chunk>
{chunk}
</chunk>
Please give a short succinct context to
situate this chunk within the overall
document for the purposes of improving
search retrieval of the chunk. Answer
only with the succinct context and
nothing else.
```

B Infrastructure Details

All experiments run on 8× NVIDIA H200 (141GB each). Context generation uses vLLM (Kwon et al., 2023) with TP=8 for 397B and DP=8 for 8B/27B/Gemma-12B, with deterministic decoding (`temperature=0`, `top_p=1.0`, `max_tokens=200`). Our environment uses vLLM 0.16.0, transformers 4.57.6, and PyTorch 2.9.1; vector search uses FAISS `IndexFlatIP` (exact inner product) and BM25 uses the `bm25s` library (Lucene scoring, k_1=1.5, b=0.75, no stemming). Embedding uses batch size 32 with the per-model used length: 512 for Qwen3-Embedding-8B/Qwen3-Embedding-0.6B/BGE-M3, 4096 for E5-Mistral/Nemotron/GritLM, 8192 for jina-v3. Although Qwen3-Embedding and jina-v3 support longer native contexts, we cap their used length to keep cross-model comparisons aligned; Late Chunking is run at extended lengths as described below. For Late Chunking, max lengths are extended to encode full documents: 8192 for Qwen3, BGE-M3, Nemotron, and jina-v3; 4096 for E5-Mistral (its standard max). GritLM-7B uses bidirectional attention via a 4D attention mask during embedding. Bootstrap CIs and permutation tests use seed 42. Reranking uses top-30 results with batch size 32 (Qwen3-Reranker) or 64 (BGE-Reranker), parallelized across 8 GPUs. Summing logged `time_sec` fields across the retained experiment artifacts used in the paper, while excluding downstream generation/judging and aggregate summary files, yields approximately 124 hours of recorded GPU-process runtime; because exploratory reruns and some preprocessing are not logged uniformly, we report this as a conservative lower bound rather than an exact end-to-end budget.

Query instruction prefixes. Qwen3-Embedding uses `Instruct: Given a query, retrieve relevant passages that answer the query\nQuery: {text}`. E5-Mistral uses the generic main-experiment instruction `Instruct: Given a web search query, retrieve relevant passages that answer the query\nQuery: {text}`; see Appendix V for the task-specific comparison. Nemotron uses `Instruct: Given a question, retrieve passages that answer the question\nQuery: {text}`. BGE-M3 uses no query instruction prefix. Passage inputs receive no instruction prefix for any model. All embeddings are L2-normalized before FAISS IndexFlatIP search, making inner-product search equivalent to cosine similarity.

Late Chunking implementation. Full documents are encoded in a single forward pass. Chunk token spans are identified via `offset_mapping`. For bidirectional models, chunk embeddings use mean pooling over the span. For causal models, we extract the last token's hidden state within each span.

C Full Main Results (Qwen3)

D Full BGE-M3 Results

E Model Scale Analysis

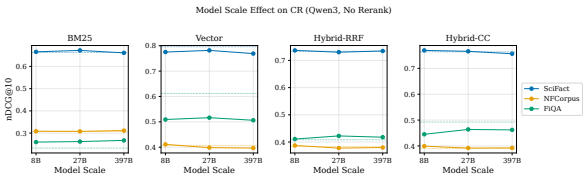

Figure 2: nDCG@10 by context-generation model scale. Performance differences are within bootstrap CIs.

TOST equivalence tests. To positively demonstrate equivalence (rather than merely failing to reject equality), we apply Two One-Sided Tests (TOST) on per-query nDCG@10 scores for Qwen3-Embedding across all dataset–method combinations. At equivalence margin $\delta = 0.03$

Table 5: Main results (nDCG@10) with Qwen3-Embedding-8B, no reranking. Bold indicates the best per dataset-method. CR-397B is only available for SciFact, NFCorpus, and FiQA.

Dataset	Strategy	BM25	Vector	Hybrid-RRF	Hybrid-CC
SciFact	Raw	0.6627	**0.7960**	0.7334	0.7640
	CR-8B	0.6656	0.7751	**0.7368**	**0.7694**
	CR-27B	**0.6721**	0.7816	0.7302	0.7658
	CR-397B	0.6610	0.7691	0.7347	0.7566
	Late Chunk.	0.6627	0.5954	0.6619	0.7170
NFCorpus	Raw	0.3074	0.4072	0.3836	0.3903
	CR-8B	0.3084	**0.4111**	**0.3873**	**0.3992**
	CR-27B	0.3078	0.3983	0.3777	0.3918
	CR-397B	**0.3114**	0.3966	0.3797	0.3924
	Late Chunk.	0.3074	0.2802	0.3186	0.3310
FiQA	Raw	0.2327	**0.6119**	0.4084	**0.4924**
	CR-8B	0.2597	0.5092	0.4108	0.4451
	CR-27B	0.2619	0.5164	**0.4224**	0.4639
	CR-397B	**0.2672**	0.5062	0.4177	0.4620
	Late Chunk.	0.2327	0.2895	0.3202	0.3389
ArguAna	Raw	0.3606	**0.5101**	**0.4499**	**0.4922**
	CR-8B	0.3691	0.4742	0.4382	0.4680
	CR-27B	**0.3720**	0.4835	0.4464	0.4756
	Late Chunk.	0.3606	0.3211	0.3890	0.3193
SciDocs	Raw	0.1536	**0.2665**	**0.2110**	0.2224
	CR-8B	0.1548	0.2598	0.2108	0.2233
	CR-27B	**0.1570**	0.2589	0.2104	**0.2241**
	Late Chunk.	0.1536	0.1620	0.1735	0.1896

nDCG@10, all 44 tested pairs (CR-8B vs CR-27B: 20; CR-8B vs CR-397B: 12; CR-27B vs CR-397B: 12) achieve equivalence at $\alpha = 0.05$. At the tighter margin $\delta = 0.02$, 34/44 (77%) achieve equivalence; the 10 non-equivalent pairs involve hybrid search on datasets where the absolute difference, while statistically detectable, remains < 0.02 nDCG@10 and is practically negligible.

F Reranking Analysis

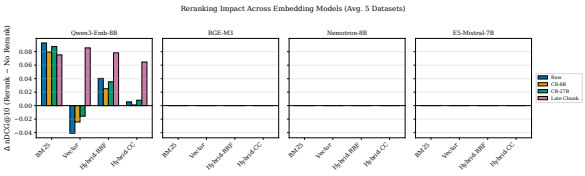

Figure 3: Reranking impact by embedding model (vector, raw). Reranking hurts high-quality models (Qwen3, GritLM on SciFact/FiQA) and helps lower-quality models, exhibiting a baseline-quality dependency.

G Random Prefix Ablation (Full)

Full 5-dataset enrichment comparison results (Table 9). CR outperforms random prefixes on 18/20 conditions (90%) for the four primary models, and on 5/5 for GritLM-7B, giving 23/25 (92%) overall across the five embedding models. The single

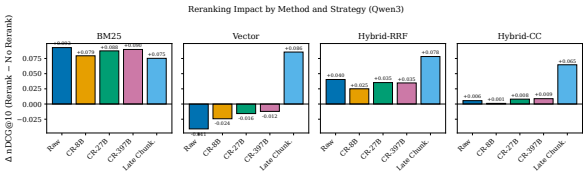

Figure 4: ΔnDCG@10 from reranking by method and strategy.

negative case is Nemotron on SciDocs (CR−RP $= -0.006$). The Qwen3/FiQA cell where CR $\approx$ RP ($\Delta = -0.000$) is the limit case discussed in §5.1.

H Statistical Significance

Multiple comparison correction. We apply Benjamini–Hochberg (BH) FDR correction across all 328 permutation tests. These are decomposed as follows: for the four primary models (Qwen3-Embedding-8B, E5-Mistral-7B, BGE-M3, Nemotron-8B), CR-8B / CR-27B / Late each contribute $4\times5\times4=80$ comparisons (4 methods $\times$ 5 datasets $\times$ 4 models) and CR-397B contributes $4\times3\times4=48$ comparisons (CR-397B is available on three datasets only), giving 288. The remaining 40 comparisons come from CR-8B on GritLM-7B and Qwen3-Embedding-0.6B (2 models $\times$ 5 datasets

Table 6: Results (nDCG@10) with BGE-M3 embedding, no reranking. CR-397B is only available for SciFact, NFCorpus, and FiQA.

Dataset	Strategy	BM25	Vector	Hybrid-RRF	Hybrid-CC
SciFact	Raw	0.6627	**0.6449**	0.6809	0.7085
	CR-8B	0.6656	0.6439	0.6876	0.7091
	CR-27B	**0.6721**	0.6411	**0.6892**	**0.7135**
	CR-397B	0.6610	0.6317	0.6724	0.7022
	Late Chunk.	0.6627	0.5804	0.6524	0.6905
NFCorpus	Raw	0.3074	0.3175	0.3434	0.3460
	CR-8B	0.3084	0.3178	0.3394	**0.3471**
	CR-27B	0.3078	**0.3182**	0.3437	0.3432
	CR-397B	**0.3114**	0.3178	**0.3452**	0.3468
	Late Chunk.	0.3074	0.2840	0.3258	0.3310
FiQA	Raw	0.2327	0.4113	0.3562	0.3854
	CR-8B	0.2597	0.3987	0.3671	0.3869
	CR-27B	0.2619	**0.4149**	0.3822	0.4034
	CR-397B	**0.2672**	0.4119	**0.3871**	**0.4079**
	Late Chunk.	0.2327	0.3477	0.3364	0.3557
ArguAna	Raw	0.3606	**0.3990**	0.4019	**0.4116**
	CR-8B	0.3691	0.3793	0.4007	0.4019
	CR-27B	**0.3720**	0.3919	**0.4060**	0.4114
	Late Chunk.	0.3606	0.3770	0.3921	0.3951
SciDocs	Raw	0.1536	0.1636	0.1792	0.1794
	CR-8B	0.1548	**0.1675**	**0.1793**	**0.1842**
	CR-27B	**0.1570**	0.1663	0.1784	0.1839
	Late Chunk.	0.1536	0.1525	0.1748	0.1757

$\times$ 4 methods). At $\alpha = 0.05$, 151 of 160 originally significant tests survive (94.4%); at $\alpha = 0.01$, 135 of 138 survive (97.8%). All main findings reported in the paper survive FDR correction: Qwen3 CR-8B vector degradation on FiQA ($p_{\text{FDR}} < 0.001$), ArguAna ($p_{\text{FDR}} < 0.001$), SciFact ($p_{\text{FDR}} < 0.01$), and SciDocs ($p_{\text{FDR}} < 0.05$); E5-Mistral CR-8B vector degradation on all 5 datasets ($p_{\text{FDR}} < 0.05$); GritLM-7B vector degradation on SciFact, FiQA, and ArguAna ($p_{\text{FDR}} < 0.01$); Late Chunking degradation for both primary last-token models on all 5 datasets ($p_{\text{FDR}} < 0.001$).

I Pooling Comparison Discussion

The same-backbone Qwen3 control reported in Table 4 (main body) shows that CR sensitivity changes substantially with pooling. Last-token pooling is consistently harmful on the three evaluated datasets, whereas mean pooling is slightly positive on all three. Because the raw baseline also changes substantially, we interpret this as strong but not definitive evidence that pooling and directional aggregation materially contribute to CR sensitivity. LLM2Vec, added as a second retrofitted bidirectional model, shows a near-neutral mixed profile that supports the intermediate regime beyond GritLM alone.

J Append vs. Prepend

K Large-Scale Spot Check and Robustness

NQ subset. To probe whether the CR pattern survives on a much larger corpus, we evaluate a Natural Questions subset ($n_{\text{queries}} = 3{,}452$ test queries with relevance judgments, ≈ 2.68M passages in the retrieval corpus) with Qwen3 and BGE-M3. The effect sizes are very small but in the saturation regime: Qwen3 changes from 0.9669 to 0.9648 ($\Delta = -0.002$), while BGE-M3 changes from 0.9341 to 0.9348 ($\Delta = +0.001$). Both deltas are within the noise floor for nDCG@10 at this scale and neither is individually statistically meaningful; we report this only as a directional consistency check rather than a full large-scale benchmark and do not run separate significance tests for it.

Nemotron exclusion. Because Nemotron has anomalously weak baselines on some short-query datasets, we recompute the model-level correlation between baseline quality and CR impact after excluding it. The trend remains qualitatively unchanged: with Nemotron, Spearman $\rho = -0.943$ ($p = 0.0048$, $n = 6$); without it, $\rho = -0.900$ ($p = 0.037$, $n = 5$). The cross-model relationship is therefore not driven by Nemotron alone.

Table 7: Reranking effect (Δ nDCG@10) with Qwen3-Reranker-8B. Positive values indicate improvement from reranking.

Dataset	Strategy	BM25	Vector	Hybrid-RRF	Hybrid-CC
SciFact	Raw	**+0.0370**	-0.0723	-0.0064	-0.0369
	CR-8B	**+0.0693**	-0.0105	**+0.0286**	-0.0040
	CR-27B	**+0.0663**	-0.0088	**+0.0442**	+0.0082
	CR-397B	**+0.0748**	-0.0035	**+0.0300**	+0.0084
	Late Chunk.	**+0.0327**	**+0.0890**	**+0.0483**	+0.0077
NFCorpus	Raw	**+0.0466**	-0.0011	**+0.0210**	**+0.0155**
	CR-8B	**+0.0508**	+0.0007	**+0.0247**	**+0.0127**
	CR-27B	**+0.0536**	**+0.0159**	**+0.0384**	**+0.0239**
	CR-397B	**+0.0513**	**+0.0173**	**+0.0363**	**+0.0239**
	Late Chunk.	**+0.0344**	**+0.0639**	**+0.0573**	**+0.0465**
FiQA	Raw	**+0.1510**	-0.1415	**+0.0623**	-0.0214
	CR-8B	**+0.1265**	-0.0800	**+0.0187**	-0.0153
	CR-27B	**+0.1580**	-0.0585	**+0.0356**	-0.0055
	CR-397B	**+0.1435**	-0.0502	**+0.0379**	-0.0059
	Late Chunk.	**+0.1038**	**+0.0986**	**+0.1070**	**+0.0738**
ArguAna	Raw	**+0.1739**	**+0.0419**	**+0.1018**	**+0.0593**
	CR-8B	**+0.0957**	-0.0018	**+0.0345**	+0.0044
	CR-27B	**+0.1008**	-0.0023	**+0.0357**	+0.0049
	Late Chunk.	**+0.1653**	**+0.1294**	**+0.1344**	**+0.1652**
SciDocs	Raw	**+0.0561**	-0.0319	**+0.0233**	**+0.0114**
	CR-8B	**+0.0549**	-0.0297	**+0.0191**	+0.0075
	CR-27B	**+0.0589**	-0.0249	**+0.0227**	+0.0088
	Late Chunk.	**+0.0401**	**+0.0472**	**+0.0442**	**+0.0301**

Table 8: Reranking impact (Δ nDCG@10) on vector search (raw strategy) across all seven models. Primary reranker is Qwen3-Reranker-8B; the Nemotron column uses BGE-Reranker-v2-M3 because Qwen3-Reranker results for Nemotron on the raw-vector condition were not retained, so its column is not directly comparable in absolute magnitude to the other six. The qualitative pattern, however, is consistent with Appendix F: reranking hurts high-quality causal-LM models (Qwen3, GritLM on SciFact/FiQA) and helps lower-quality or bidirectional models.

	Qwen3	GritLM	E5-Mis.	Q-0.6B	jina-v3	BGE-M3	Nemot.
SciFact	$-.072$	$-.059$	$-.048$	$+.013$	$+.001$	$+.066$	$+.051$
NFCorp.	$-.001$	$+.016$	$+.009$	$+.034$	$+.023$	$+.062$	$+.066$
FiQA	$-.142$	$-.093$	$-.058$	$+.004$	$+.008$	$+.040$	$+.022$
ArguAna	$+.042$	$+.124$	$+.097$	$+.079$	$+.196$	$+.145$	$-.045$
SciDocs	$-.032$	$+.010$	$+.019$	$+.007$	$+.030$	$+.053$	$+.014$
Mean	$-.041$	$-.001$	$+.004$	$+.027$	$+.052$	$+.073$	$+.022$

Random-prefix multi-seed stability. We repeated the random-prefix ablation across multiple seeds on five representative model–dataset pairs. For Qwen3, we used 10 seeds on SciFact and ArguAna; for BGE-M3, we used 5 seeds on FiQA, SciFact, and ArguAna. RP variance was small in all cases (Qwen3 std. 0.0046 on SciFact and 0.0024 on ArguAna; BGE-M3 std. 0.0051 on FiQA, 0.0066 on SciFact, and 0.0059 on ArguAna). The qualitative interpretation was unchanged across seeds: Qwen3 retained negative net CR effects with modest positive CR$-$RP gaps ($+0.028$ on SciFact, $+0.038$ on ArguAna), while BGE-M3 consistently showed much larger CR$-$RP gaps ($+0.093$ on FiQA, $+0.074$ on SciFact, $+0.087$ on ArguAna). We did not complete the exact FiQA/Qwen3 case where CR and RP are closest, so we present this as a partial robustness check rather than an exhaustive seed-sensitivity analysis.

jina-v3 prepend-based CR check. Because jina-v3 is trained specifically for Late Chunking, we additionally test whether standard prepend-based CR helps it under ordinary chunk embedding. It does not: across all five datasets, CR-8B produces small degradations relative to raw chunks (FiQA -0.021, SciFact -0.003, ArguAna -0.003, NFCorpus -0.011, SciDocs -0.010). This indicates that LC-specific training does not imply robustness to prepend-based enrichment, further separating the CR and LC interventions.

L Exploratory Attention Probes

Attention analysis. We extract final-layer attention weights from the pooled token on n=100 CR-enriched chunks per dataset on SciFact and FiQA (n_{total}=200 CR chunks per model; Figure 5). Reported $\pm$ values are the sample standard deviation of the per-chunk prefix-attention fraction. The re-

Table 9: Random Prefix (RP) ablation: vector nDCG@10 across five datasets and five embedding models. "Title" = source document title prepended (zero LLM cost); GritLM "Title" is left blank because the Title condition was not evaluated for this model. Δ = CR-8B $-$ RP.

Dataset	Model	Raw	Title	CR-8B	RP	Δ
SciFact	Qwen3	.796	.787	.775	.752	+.024
	BGE-M3	.645	.649	.644	.574	+.070
	E5-Mis.	.773	.762	.752	.657	+.095
	Nemot.	.633	.644	.631	.617	+.014
	GritLM	.789	—	.771	.755	+.017
NFCorpus	Qwen3	.407	.408	.411	.375	+.036
	BGE-M3	.317	.316	.318	.258	+.060
	E5-Mis.	.400	.395	.389	.329	+.060
	Nemot.	.053	.053	.066	.055	+.010
	GritLM	.403	—	.405	.374	+.031
FiQA	Qwen3	.612	.612	.509	.510	−.000
	BGE-M3	.411	.411	.399	.299	+.099
	E5-Mis.	.533	.533	.476	.438	+.038
	Nemot.	.029	.029	.033	.025	+.008
	GritLM	.604	—	.565	.465	+.100
ArguAna	Qwen3	.510	.502	.474	.398	+.076
	BGE-M3	.399	.365	.379	.269	+.110
	E5-Mis.	.454	.455	.444	.393	+.051
	Nemot.	.450	.440	.441	.276	+.165
	GritLM	.431	—	.401	.362	+.039
SciDocs	Qwen3	.266	.263	.260	.227	+.033
	BGE-M3	.164	.173	.168	.116	+.052
	E5-Mis.	.208	.206	.194	.177	+.017
	Nemot.	.030	.053	.045	.051	−.006
	GritLM	.222	—	.219	.195	+.024

sults argue against a simple "attention diversion" account: Nemotron allocates the *highest* prefix attention ($82\% \pm 3\%$) yet is CR-robust, while GritLM allocates only $36\% \pm 6\%$ (comparable to BGE-M3 at $34\% \pm 8\%$) yet is CR-vulnerable. This suggests that vulnerability depends not only on *how much* attention is assigned to the prefix but also on how the model integrates that signal. We treat this as exploratory because raw attention weights are known to overstate or understate the actual contribution of input tokens once value-vector norms are taken into account (Kobayashi et al., 2020).

GritLM attention-mode ablation. Disabling GritLM's bidirectional mask collapses quality (FiQA: $0.604 \rightarrow 0.099$) and reverses CR's sign ($+0.029$). Because this change also destroys baseline retrieval quality, we do not treat it as a clean identification strategy. It nevertheless illustrates the model's intermediate position: bidirectional inference elevates quality but does not fully erase inherited structural vulnerability.

Table 10: Statistical significance summary (permutation test, $p < 0.05$). Fraction of retained model–dataset comparisons where each strategy significantly differs from Raw; denominators vary because CR-397B is available on three datasets and the added GritLM/Qwen3-0.6B checks cover CR-8B only.

Strategy	BM25	Vector	Hybrid-RRF	Hybrid-CC
CR-8B	12/30	19/30	5/30	11/30
CR-27B	12/20	13/20	7/20	6/20
CR-397B	4/12	7/12	2/12	4/12
Late Chunk.	0/20	19/20	20/20	19/20
Avg. $\lvert\Delta\rvert$ when significant (nDCG@10):				
CR-8B	0.0178	0.0263	0.0180	0.0201
CR-27B	0.0167	0.0234	0.0167	0.0197
CR-397B	0.0345	0.0380	0.0356	0.0313
Late Chunk.	–	0.1779	0.0587	0.0859

Table 11: Appending versus prepending the same CR-generated content for Qwen3. Δ is relative to Raw. Adding generated content is harmful in both positions; prepend is somewhat more harmful on FiQA, but the direction is not uniform across datasets.

Dataset	**Raw**	**Prepend Δ**	**Append Δ**
FiQA	.612	−.103	−.082
SciFact	.796	−.022	−.025
ArguAna	.512	−.034	−.040

M Cost Comparison

N BM25 Keyword Analysis

O Case Studies

P Additional Figures

Q Input Length and Truncation Analysis

Table 14 reports per-dataset token length statistics for raw chunks, CR-8B enriched chunks, and full documents (for Late Chunking), all tokenized with the Qwen3-Embedding-8B tokenizer.

R Non-Truncated Subset Analysis

To rule out truncation as a confound (CR-8B increases truncation from 2.3–9.9% to 3.6–17.0% for 512-max models), we re-embed and re-evaluate using only chunks whose CR-enriched text remains under 512 tokens.

All five datasets show consistent direction between the non-truncated subset and the full dataset for both Qwen3-Embedding-8B (causal) and BGE-M3 (bidirectional). The largest truncation rate (NFCorpus, 17%) produces no discernible difference ($+0.003$ vs. $+0.004$ for Qwen3). This is consistent with CR's degradation being driven by structural

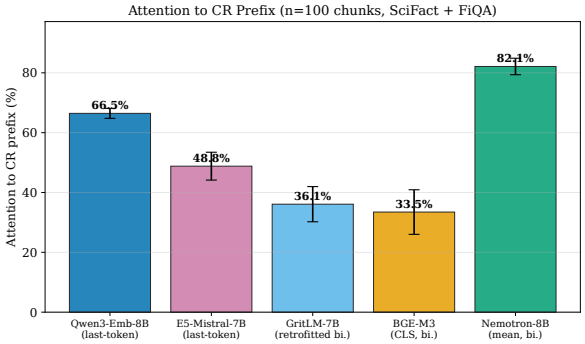

Figure 5: Attention to CR prefix (n=100 CR-enriched chunks per dataset on SciFact and FiQA, n_{total}=200 per model). Nemotron allocates the most attention (82%) yet is CR-robust, suggesting that prefix-attention mass alone does not explain CR vulnerability.

Table 12: Context generation cost (pricing: typical cloud API rates as of 2026-03). Δ nDCG@10 averaged across dataset–method pairs for Qwen3-Emb-8B.

Model	Active Params	Input Tokens	Output Tokens	Est. API Cost	Avg Δ nDCG@10
CR-8B	8B	50.9M	4.0M	\$1.8	-0.009
CR-27B	27B	51.6M	4.1M	\$5.7	-0.007
CR-397B	17B (MoE)	32.5M	2.7M	\$6.5	-0.011

No pairwise comparison is significant ($p > 0.05$). Δ for Qwen3-Emb-8B.

distortion of the embedding representation rather than by information loss from truncation.

S Chunk Size Sensitivity

A natural concern is whether CR becomes more beneficial for shorter, more decontextualized chunks. We test this by filtering existing chunks to three size buckets ($\leq$256, $\leq$512, $\leq$800 tokens) and re-evaluating CR's effect for Qwen3-8B, E5-Mistral, and BGE-M3 on FiQA, SciFact, and ArguAna (Figure 10).

Counter to the expectation that shorter chunks—which have less inherent context—would benefit most from CR enrichment, we find that CR's degradation for causal models is actually *smaller* for shorter chunks (e.g., Qwen3 FiQA: -0.071 at $\leq$256 vs. -0.102 at $\leq$800). BGE-M3 remains consistently robust across all sizes ($|\Delta| < 0.02$). This suggests that CR's harm is not driven by chunk decontextualization but by the structural distortion of prepending text to causal embedders, which is proportionally larger for longer chunks that carry more positional information.

T Shuffled Prefix Control

To further dissect CR's semantic contribution, we create a *shuffled prefix* condition: for each chunk, we take its CR context and randomly shuffle the words (seed 42), preserving vocabulary and length but destroying syntactic structure and meaning.

Two findings emerge (Table 16): **(1)** For Qwen3, shuffled prefix performs *between* CR and RP on most datasets (e.g., FiQA: CR $\approx$ RP = 0.510, Shuffle = 0.537), and on FiQA specifically, CR $\approx$ RP, confirming that on this dataset Qwen3 extracts no measurable value from CR content—even domain vocabulary. **(2)** For BGE-M3, shuffled prefix consistently outperforms RP (e.g., FiQA: Shuffle 0.382 vs. RP 0.299), suggesting that domain vocabulary alone—even without syntactic coherence—carries value for bidirectional models. The gap between CR and Shuffle is small ($|\Delta| \leq 0.02$), indicating that most of CR's semantic contribution for BGE-M3 comes from keyword-level rather than sentence-level semantics.

Table 13: Types of keywords added by CR context on FiQA (BM25 ΔnDCG@10: $+14.8\%$). CR enriches chunks with domain-specific terms, entity names, and topical keywords absent from raw text, directly improving BM25 lexical matching.

Category	Example	Freq.
Domain terms	*financial, investment, strategies, risks, economic, dividend, portfolio*	14.9K
Entity resolution	"the company" $\rightarrow$ *Ayn Rand, New York, Sergey Brin, San Francisco*	1.2K
Topic framing	*advice, considerations, benefits, personal, challenges*	12.5K
Query-bridge terms	terms appearing in both CR context and $\geq$1 query: *interest rates, savings account, insurance products*	8.4K

Illustrative example. *Raw chunk:* "Savings accounts with 8% APY? Unheard of these days. You're lucky if you find one at 1%..." *CR context:* "Current savings account interest rates are very low, typically around 1% APY, making high-yield savings and money market funds recommended for short-term holding." The CR prefix adds 18 tokens that also appear in related queries (*rates, yield, funds, term*), directly improving BM25 recall.

U Context-Generation LLM Independence

To test whether our findings are specific to the Qwen context-generation family, we regenerate contexts for three datasets using Gemma-3-12B-it (Gemma Team, 2025) and evaluate with both Qwen3-Embedding-8B and GritLM-7B. Table 17 shows that both LLM families produce the same directional effects: CR degrades Qwen3-Embedding on all three datasets regardless of the context source, and likewise for GritLM. The magnitudes differ slightly (Gemma contexts are less harmful on FiQA for Qwen3, -0.068 vs. -0.103), but the direction is consistent. This argues against the pos-

Case 1: CR rescues weak embedding *E5-Mistral / SciFact* nDCG@10: 0.000 → **1.000**
Query: "Hypocretin neurones induce panic-prone state in rats."
Raw top-1 (irrelevant): "Structural plasticity of the adult brain: how animal models help us understand brain changes in depression…"
CR context added: "*The chunk discusses the role of orexin in panic anxiety, highlighting its involvement in the pathophysiology of panic disorder through animal models…*"
CR top-1 (relevant): "A key role for orexin in panic anxiety. Panic disorder is a severe anxiety disorder…"
→ CR introduces domain terms (*orexin*, *panic anxiety*) absent from the raw chunk, enabling the weak embedding to match the query.

Case 2: CR degrades strong embedding *Qwen3 / FiQA* nDCG@10: 1.000 → **0.000**
Query: "How is the opening-day price of a stock decided?"
Raw top-1 (relevant): "When a stock is going to become public there's a level of analysis required to figure out the range of IPO price…"
CR context added: "*The chunk discusses how a stock opens on its first day of trading, reflecting pre-market news and order imbalances…*"
CR top-1 (different, less relevant): "I'd add, this is actually the way any stock opens every day…the closing price of the prior day is what it is…"
→ The CR prefix shifts the last-token embedding, causing a relevant chunk to be displaced by a tangentially related one. Qwen3 already captured the semantics; the prefix adds noise.

Case 3: Late Chunking recovers weak embedding *Nemotron / SciFact* nDCG@10: 0.000 → **1.000**
Query: "High cardiopulmonary fitness causes increased mortality rate."
Raw top-1 (irrelevant): "Clinical epidemiology of heart failure. The aim of this paper is to review…" (score: 0.349)
Late Chunking: Same document encoded with full-document context → chunk embedding incorporates surrounding sections on fitness and mortality.
→ Nemotron's mean pooling distributes document context across all positions; Late Chunking provides the global context that the short chunk lacked.

Case 4: Late Chunking destroys strong embedding *Qwen3 / SciFact* nDCG@10: 1.000 → **0.000**
Query: "Recurrent mutations occur frequently within CTCF anchor sites adjacent to oncogenes."
Raw top-1 (relevant): "3D Chromosome Regulatory Landscape of Human Pluripotent Cells…transcriptional enhancers and insulators…CTCF-CTCF loops…" (score: 0.546)
Late Chunking: Last-token pooling now captures only the final token's representation from the long-document encoding, losing the specific CTCF/mutation information present in the chunk.
→ Qwen3's last-token pooling creates an information bottleneck: in standard mode, the last token summarizes the chunk effectively. In Late Chunking, the last token of the chunk span in a long document captures only local context, losing the chunk-level semantics.

Figure 6: Illustrative case studies showing how chunk enrichment strategies interact with embedding model quality. Cases 1–2 demonstrate the baseline-quality dependency of CR: it helps E5-Mistral by adding domain keywords the weak model cannot infer, but hurts Qwen3 by perturbing embeddings that already capture the semantics. Cases 3–4 show the model-specific nature of Late Chunking: mean pooling (Nemotron) distributes document context beneficially, while last-token pooling (Qwen3) creates an information bottleneck that loses chunk-specific semantics.

sibility that our findings are an artifact of within-family bias between Qwen context generation and Qwen embedding.

V E5-Mistral Instruction Sensitivity

E5-Mistral-7B-Instruct is designed to use task-specific query instructions (e.g., *"Given a scientific claim, retrieve documents that support or refute the claim"* for SciFact). Our main experiments use a generic instruction (*"Given a web search query, retrieve relevant passages"*), which is a common deployment scenario when task-specific instructions are unavailable. Table 18 compares both conditions.

Three findings emerge: (1) Task-specific instructions produce small, inconsistent differences from the generic instruction ($|\Delta| \leq .037$ on raw chunks), with 2 of 5 datasets showing *lower* nDCG@10 under correct instructions (SciFact, SciDocs). (2) CR's relative effect is non-positive under both conditions: all datasets show $\Delta \leq 0$ for CR-8B and CR-27B relative to raw, indicating that the instruction is unlikely to be a major confound for our central finding that CR hurts causal-LM-based models. (3) Late Chunking remains heavily degraded under both conditions (.001–.243), making instruction mismatch an unlikely explanation.

This analysis supports that our main findings—CR hurts causal-LM-based models, Late Chunking is heavily degraded for causal last-token models—hold regardless of instruction configuration. The generic-instruction setting in our main experiments represents a realistic deployment scenario.

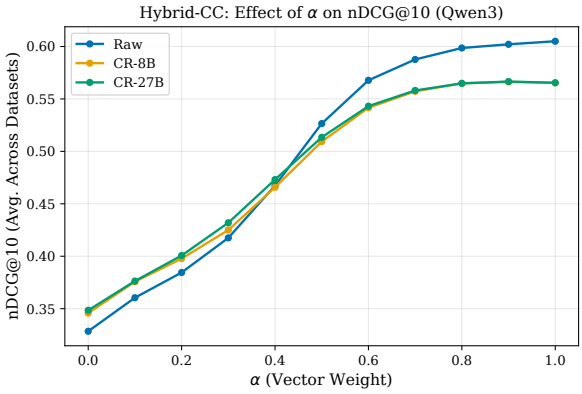

Figure 7: Optimal Hybrid-CC α from 5-fold CV across six models. α^* correlates with baseline quality: Qwen3 (0.94) > E5-Mistral (0.84) > GritLM (0.81) > Qwen3-0.6B (0.79) > BGE-M3 (0.57) > Nemotron (0.38). The fitted line is based on n=6 models and should be interpreted only as a preliminary heuristic, not a tuning rule.

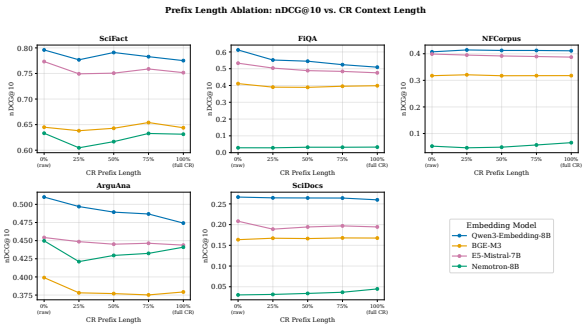

Figure 8: Prefix length ablation across five datasets and five models. Causal last-token models degrade monotonically; GritLM-7B shows intermediate sensitivity; natively bidirectional encoders are insensitive.

W Extended Limitations Discussion

This appendix expands on the main *Limitations* section with additional implementation-level caveats. It is supplementary; the main *Limitations* section above remains the canonical statement.

- We evaluate on 5 of 18 BEIR datasets; our five datasets range from 3.7K to 58K chunks, but results on larger corpora (NQ: 2.7M passages, MS MARCO: 8.8M, HotpotQA: 5.2M) and domain-specific benchmarks (TREC-COVID) remain unvalidated. Production RAG systems often operate at scales orders of magnitude beyond our evaluation.

- CR-397B is only evaluated on 3 datasets. Furthermore, CR-397B is a Mixture-of-Experts model (17B active parameters), making the scaling comparison indirect.

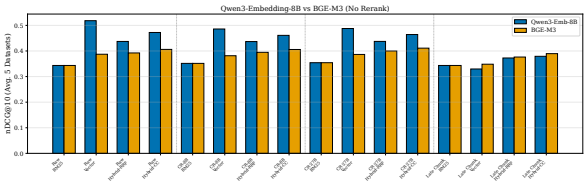

Figure 9: Vector nDCG@10 across embedding models and CR strategies.

Table 14: Input length statistics and truncation rates. “>512”: percentage exceeding Qwen3/BGE-M3 max. 4,096-max models have 0% truncation.

Dataset	Condition	Mean	Median	Max	>512
SciFact	Raw	328	304	801	9.2%
	CR-8B	374	351	873	13.9%
	Doc (Late)	329	303	1,928	—
NFCorpus	Raw	355	353	801	9.9%
	CR-8B	397	394	869	17.0%
	Doc (Late)	356	352	2,335	—
FiQA	Raw	174	119	801	4.5%
	CR-8B	210	155	912	5.6%
	Doc (Late)	173	117	3,483	—
ArguAna	Raw	214	192	801	2.3%
	CR-8B	255	232	885	3.6%
	Doc (Late)	214	191	1,398	—
SciDocs	Raw	229	206	801	3.3%
	CR-8B	269	245	887	4.3%
	Doc (Late)	228	203	6,969	—

- **Dual-effect framework scope.** We now observe the intermediate regime in two retrofitted bidirectional models (GritLM and LLM2Vec), but this category remains undersampled relative to the causal last-token and native bidirectional groups. We therefore do not treat the retrofitted bidirectional case as a complete taxonomy, only as an observed intermediate regime supported by two models. The GritLM attention-mode ablation provides within-model evidence but is confounded: disabling bidirectional attention collapses nDCG@10 from 0.60 to 0.10, so the CR sign reversal may partly reflect a floor effect rather than an architectural mechanism. Qwen3-0.6B (low-quality, causal) shows the same CR degradation pattern as Qwen3-8B, indicating that quality is not the primary driver among causal models.

- **E5-Mistral instruction mismatch.** Main experiments use a generic query instruction rather than E5-Mistral’s recommended per-dataset task-specific instructions. Appendix V

Table 15: CR effect on non-truncated subsets only (vector, no reranking). Direction and magnitude match the full-dataset results, confirming that truncation is not a driver.

Dataset	Subset	Qwen3-8B		BGE-M3	
	%	Δ	Full Δ	Δ	Full Δ
SciFact	86%	−0.022	−0.021	−0.005	−0.001
NFCorpus	83%	+0.003	+0.004	−0.001	+0.001
FiQA	94%	−0.094	−0.103	−0.014	−0.012
ArguAna	96%	−0.037	−0.036	−0.020	−0.020
SciDocs	96%	−0.007	−0.006	+0.004	+0.004

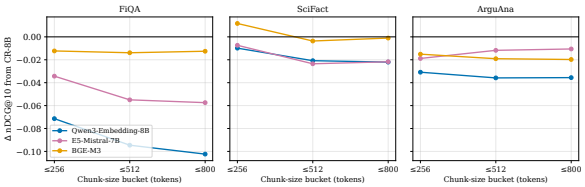

Figure 10: CR effect (Δ nDCG@10) across chunk size buckets. Shorter chunks do *not* benefit more from CR; degradation for causal models is actually smaller at shorter lengths.

shows that task-specific instructions produce small, inconsistent differences ($|\Delta| \leq .037$) and that the direction of CR effects is preserved under both conditions.

- **E5-Mistral padding.** The Mistral tokenizer defaults to left-padding, which causes batch-dependent RoPE position shifts and non-deterministic embeddings. All E5-Mistral experiments in this paper use right-padding (`padding_side="right"`) to ensure deterministic results.

- **Nemotron-8B query collapse.** Nemotron shows extremely low baselines on short-query datasets (NFCorpus: 0.053, FiQA: 0.029) due to instruction prefix dominance in mean pooling (Lee et al., 2025a).

- **Late Chunking context length mismatch.** Late Chunking requires encoding full documents, which exceeds the trained max sequence length for Qwen3 and BGE-M3 (512 tokens; documents reach 6,969 tokens). Their degradation may partly reflect this context-length mismatch rather than Late Chunking's intrinsic limitations. We evaluate jina-embeddings-v3 (Sturua et al., 2024), the only model in our set trained with Late Chunking support, and find small but uniformly negative deltas relative to its standard chunk-level baseline (ranging from −0.002 to −0.041), suggesting that even LC-specific training does not guarantee consistent improvement under our chunking and document-length distribution.

- Both rerankers are cross-encoders; ColBERT-style may differ.

- **Reranker instruction-format match.** Qwen3-Reranker-8B is trained to consume queries together with task-specific instructions, and the strong-baseline penalty we observe on Qwen3-Embedding (e.g., FiQA −0.142) could partly reflect query/passage instruction-format mismatch between the embedder and the reranker rather than a pure baseline-quality dependency. We therefore frame reranking as exhibiting a *baseline-quality-correlated pattern* on our pipeline rather than a clean causal claim. The BGE-Reranker-v2-M3 cross-validation only confirms direction, not the exact magnitude.

Table 16: Shuffled prefix control (vector, no reranking). Shuf. = word-shuffled CR prefix. For Qwen3, CR ≈ RP ≈ Shuffle on FiQA. For BGE-M3, Shuffle > RP, suggesting domain vocabulary itself has value.

Model	Dataset	Raw	CR	Shuf.	RP
Qwen3	FiQA	.612	.510	.537	.510
	SciFact	.796	.774	.787	.752
	ArguAna	.510	.474	.498	.398
BGE-M3	FiQA	.411	.399	.382	.299
	SciFact	.645	.644	.648	.574
	ArguAna	.399	.379	.393	.269

Table 17: LLM family independence. Gemma-3-12B contexts produce the same directional effects as Qwen3-8B contexts.

		Qwen3-Emb-8B		GritLM-7B	
Dataset	Strategy	nDCG	Δ	nDCG	Δ
SciFact	Raw	.796	—	.789	—
	CR (Qwen-8B)	.775	−.021	.771	−.018
	CR (Gemma-12B)	.777	−.019	.781	−.008
FiQA	Raw	.612	—	.604	—
	CR (Qwen-8B)	.509	−.103	.565	−.039
	CR (Gemma-12B)	.544	−.068	.561	−.043
ArguAna	Raw	.510	—	.431	—
	CR (Qwen-8B)	.474	−.036	.401	−.031
	CR (Gemma-12B)	.488	−.022	.408	−.023

Table 18: E5-Mistral nDCG@10: generic vs. task-specific instructions. Differences are small; CR direction is preserved.

		Strategy				
Dataset	Instruction	Raw	CR-8B	CR-27B	Late	Title
SciFact	Generic	**.773**	.752	.736	.159	.762
	Task-Spec.	.769	.748	.755	.146	.768
NFCorpus	Generic	.400	.389	.385	.076	.395
	Task-Spec.	.401	.399	**.410**	.044	.406
FiQA	Generic	.533	.476	.495	.005	.533
	Task-Spec.	**.570**	.505	.516	.005	.570
ArguAna	Generic	.454	.444	.445	.164	.455
	Task-Spec.	**.471**	.468	.474	.243	.467
SciDocs	Generic	**.208**	.194	.201	.001	.206
	Task-Spec.	.174	.179	.179	.001	.172

- **Score normalization.** Hybrid-CC uses min-max normalization (Bruch et al., 2024); z-score may yield different α^*.
- α^* **prediction** is based on $n = 6$ models (4 degrees of freedom; $R^2 = 0.87$). BGE-M3 is the largest outlier ($\alpha^* = 0.57$ vs. predicted 0.69).
- **Multiple comparisons.** 328 permutation tests. BH-FDR correction (Appendix H) retains 94.4% of significant results at $\alpha = 0.05$; all main findings survive.
- **Chunk length and truncation.** For 512-max models, 3.6%–17.0% of CR-8B chunks are truncated. A non-truncated subset analysis (Appendix R) shows that CR's degradation direction and magnitude are preserved when truncated chunks are excluded, indicating that truncation is unlikely to be a major confound.
- **Pooling control caveat.** The same-backbone Qwen3 pooling experiment changes both the pooling rule and the absolute retrieval quality. It therefore provides strong evidence that pooling matters, but it does not isolate pooling as the sole mechanism behind CR sensitivity.
- **Large-scale validation scope.** Our NQ experiment is only a subset-level spot check. It supports directional consistency, but it is not a full-scale replication on a web-scale benchmark.
- **Random-prefix seed sensitivity.** We added multi-seed RP checks on five representative model–dataset pairs and found low variance with unchanged qualitative interpretations. However, we did not complete the exact FiQA/Qwen3 case where CR and RP are closest, so the seed-stability evidence is partial rather than exhaustive.
- **Limited downstream evaluation.** Our end-to-end evaluation (Appendix X) covers two datasets (FiQA, SciFact) with four embedding models, a single generator (Qwen3-8B, temperature 0.7), and judge (Claude Sonnet 4.6). We clarify the SciFact "paradox" as a task property: claim-verification labels can remain stable under small retrieval perturbations. However, we do not present new multi-seed generation results in this submission.

X Supplementary Downstream Evaluation

To test whether retrieval-level changes can propagate to end-to-end RAG quality, we evaluate on FiQA (648 queries, LLM-as-a-Judge with Claude Sonnet 4.6 (Anthropic, 2026)) and SciFact (188 queries, claim verification accuracy) using four embedding models under Raw vs. CR-8B. We treat this as a supplementary analysis rather than a primary claim: FiQA uses a single LLM judge, and generation uses one stochastic run per condition (Qwen3-8B, temperature 0.7). The generator always receives *raw* chunk text—CR affects only which chunks are retrieved. See Appendix Y for the judge prompt.

Table 19 provides supplementary downstream evidence. On FiQA, Qwen3 shows a reliable drop in judged Correctness under CR-8B ($\Delta = -0.13$, paired randomization $p < 0.001$, BH-FDR $q < 0.01$, 95% bootstrap CI $[-0.19, -0.07]$); this is the only downstream comparison that survives multiple-comparison correction across the 16 downstream model$\times$metric tests. The remaining FiQA judge dimensions are directionally consistent with the retrieval results but do not survive correction (Qwen3 Relevance q=0.14; all other model$\times$metric pairs $q \geq 0.74$). On SciFact, none of the four model-level accuracy comparisons reaches significance under McNemar's exact test (Qwen3 p=0.24, GritLM p=0.29, E5-Mistral p=0.65, BGE-M3 p=0.80): retrieval degradation does not translate cleanly into lower accuracy. We interpret this as a task effect rather than a contradiction: claim verification is label-stable when

Table 19: Downstream evaluation. **(a)** FiQA: LLM-as-a-Judge (1–5). **(b)** SciFact: claim verification. Bold marks the only downstream comparison that survives paired randomization plus Benjamini–Hochberg FDR correction across the 16 downstream model×metric comparisons (Qwen3 FiQA Correctness: $\Delta = -0.13$, paired randomization $p < 0.001$, $q < 0.01$, 95% bootstrap CI $[-0.19, -0.07]$). All other comparisons are directionally consistent with the retrieval-side findings but not statistically supported after correction.

(a) FiQA — LLM-as-a-Judge ($n = 645$–648)

Model	Strat.	Rel.	Fai.	Cor.	nDCG
Qwen3	Raw	4.09	4.25	3.32	.612
	CR-8B	4.02	4.27	3.19	.509
	Δ	−.07	+.02	**−.13**	−.103
GritLM	Raw	4.07	4.24	3.27	.604
	CR-8B	4.05	4.23	3.24	.565
	Δ	−.02	−.00	−.04	−.039
E5-Mis.	Raw	3.97	4.19	3.14	.533
	CR-8B	3.95	4.17	3.12	.476
	Δ	−.02	−.02	−.03	−.057
BGE-M3	Raw	3.83	4.12	2.98	.411
	CR-8B	3.81	4.12	2.99	.399
	Δ	−.02	−.00	+.01	−.012

(b) SciFact — Claim Verification ($n = 188$)

Model	Strat.	Acc.	F1†	NEI FP	nDCG
Qwen3	Raw	.771	.775	18	.796
	CR-8B	.803	.814	16	.775
	Δ	+.032	+.039	−2	−.021
GritLM	Raw	.787	.789	16	.789
	CR-8B	.819	.835	17	.771
	Δ	+.032	+.046	+1	−.018
E5-Mis.	Raw	.798	.785	14	.773
	CR-8B	.782	.797	20	.752
	Δ	−.016	+.012	+6	−.021
BGE-M3	Raw	.771	.797	23	.645
	CR-8B	.761	.766	22	.644
	Δ	−.011	−.031	−1	−.001

†Macro-F1 over SUPPORT/CONTRADICT only.

multiple retrieved chunks already provide sufficient evidence, and the top retrieved sets remain highly redundant, so modest retrieval perturbations can change which evidence chunk is surfaced without changing the final class prediction. We therefore use the downstream analysis only to check gross inconsistencies, not to override the retrieval-side conclusions.

Y LLM-as-a-Judge Prompt (FiQA)

The judge (Claude Sonnet 4.6, temperature 0) receives the query, top-5 passages, generated answer, and gold documents:

```
You are evaluating a RAG system's answer.
## Query: {query}
## Retrieved Passages: [1]...[5]
## Generated Answer: {answer}
## Gold Reference: {gold_documents}
Rate (1-5):
Relevance, Faithfulness, Correctness
Respond JSON: {"relevance": <int>,
 "faithfulness": <int>, "correctness": <int>}
```