KDD 2026 August 9–13, 2026 Jeju, Korea
32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining · ICC Jeju

Poster #346

SCIENTIFIC FOUNDATION MODELS

CoRank: LLM-Based Compact Reranking with Document Features for Scientific Retrieval

Runchu Tian[1,*] Xueqiang Xu[1,*] Bowen Jin[1] SeongKu Kang[2] Jiawei Han[1,†]

[*]Equal contribution · [†]Corresponding author

[1]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign · [2]Dept. of Computer Science and Engineering, Korea University

PAPER doi 10.1145/3770855.3818844

CODE & DATA github.com/Rachum-thu/CoRank

ONE-SENTENCE SUMMARY

Don't feed **full papers** to an LLM reranker. Feed LLM-extracted **category + section + keywords** so **10× more candidates** fit in one context window, then restore full text only for the shortlist. No training, no tuning, works with any LLM backbone.

+4.9 avg. nDCG@10 gain — 50.6 → 55.5, all backbones

200 vs 20 candidates per prompt — 10–50 tokens/doc, not ~200

40% of sliding-window cost — and +1.4 nDCG@10 better

1 The Problem

LLM *listwise* reranking hits two compounding bottlenecks in the **scientific** domain:

- **Weak first-stage retrieval.** Sparse and dense retrievers generalise poorly to science; relevant papers rank *far down* the list.
- **Full text is token-heavy.** One paper costs ~200 tokens (*up to 5,774*), so rerankers see only **~20 candidates per prompt**.

→ Relevant documents are excluded *before* reranking even begins — a hard ceiling on quality, no matter how good the reranker is.

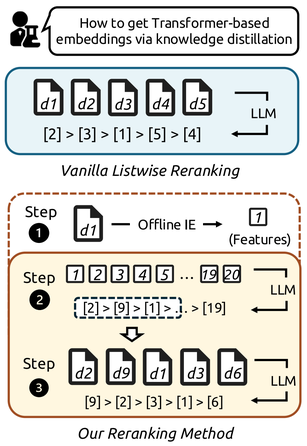


Fig. 1. Vanilla reranking scores a handful of full-text candidates. **CoRank** extracts features offline, reranks a far larger pool with them, then refines the shortlist with full text.

EXPERIMENTAL SETUP

Benchmarks	**LitSearch**, **CSFCube**, **NFCorpus**, **SciFact**, **Trec-Covid** — computer science, biomedical and clinical retrieval
First stage	Contriever (also BM25, Specter2, GTR-T5 for robustness)
Rerankers	Qwen3-32B · Gemini 2.0 Flash · GPT-4.1-mini
Extraction	Qwen3-8B, zero-shot, run offline and cached
Metrics	nDCG@10 and Recall@10

2 CoRank in Three Stages

1 Offline preprocessing
LLM(p_i) → Category$_i$, [Section$_{ij}$], [Keyword$_{ij}$].
Cached — no runtime cost.

2 Coarse reranking — compact features
200 compact reps in one prompt → keep the top-k.
Recovers papers the retriever buried.

3 Fine reranking — full text
Restore full text of the top **20** and rerank again.
Recovers detail lost in extraction.

A compact rep is **category + section + keywords — 10–50 tokens, not ~200. Training-free and orthogonal to sliding windows.**

3 Efficiency & Robustness

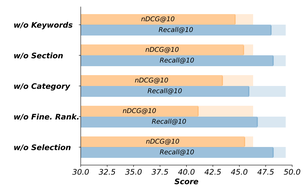
Method	N@10	R@10	Tokens	Cost
Vanilla	26.0	27.3	1.01M	$0.40
Sliding window	40.8	44.2	9.06M	$3.62
CoRank	**42.2**	**45.2**	**3.60M**	**$1.44**
Both	43.8	48.1	11.65M	$4.66

+1.4 nDCG@10 at 40% of the tokens. Robust too: small backbones gain +3.9 N@10, gains hold for BM25/Specter2/GTR-T5, and 5–20 keywords moves nDCG@10 by only 3.1%.

4 Main Results

Reranker	Strategy	LitSearch N@10	LitSearch R@10	CSFCube N@10	CSFCube R@10	NFCorpus N@10	NFCorpus R@10	SciFact N@10	SciFact R@10	Trec-Covid N@10	Trec-Covid R@10
None	First-stage only	14.2	21.2	23.6	16.4	32.8	15.5	67.7	81.3	59.6	1.6
RankZephyr	Vanilla$_{Sliding}$	34.1	39.0	34.1	22.7	28.6	15.1	44.4	80.0	59.5	1.6
Qwen3-32B	Vanilla$_{Full}$	25.7	27.3	28.4	20.2	36.8	17.2	77.6	85.7	67.0	1.7
	CoRank$_{Full}$	**39.6**	**41.9**	**31.8**	**23.4**	**39.8**	**18.5**	**80.8**	**89.1**	**76.4**	**2.0**
Qwen3-32B	Vanilla$_{Sliding}$	39.8	43.5	31.2	22.9	38.8	18.0	80.0	89.8	76.7	2.1
	CoRank$_{Sliding}$	**42.2**	**46.2**	**35.5**	**26.6**	**39.8**	**18.7**	**81.7**	**90.3**	**78.1**	**2.1**
Gemini 2.0 Flash	Vanilla$_{Full}$	26.1	27.3	28.5	18.8	35.9	16.8	75.6	85.4	68.9	1.8
	CoRank$_{Full}$	**40.7**	**44.2**	**35.1**	**27.8**	**38.0**	**17.5**	**77.6**	**88.3**	**76.1**	**2.0**
Gemini 2.0 Flash	Vanilla$_{Sliding}$	40.8	44.6	32.2	23.2	38.3	17.9	77.2	88.8	79.7	2.2
	CoRank$_{Sliding}$	**43.3**	**48.9**	**38.1**	**31.0**	**39.0**	**18.6**	**78.0**	**91.1**	**79.2**	**2.2**
GPT-4.1-mini	Vanilla$_{Full}$	26.1	27.2	28.7	19.7	37.1	16.8	76.7	84.9	68.4	1.8
	CoRank$_{Full}$	**46.3**	**49.4**	**39.0**	**30.5**	**41.4**	**19.1**	**79.6**	**89.8**	**80.7**	**2.2**
GPT-4.1-mini	Vanilla$_{Sliding}$	41.9	44.4	34.9	25.3	40.0	17.9	79.3	90.5	81.0	2.2
	CoRank$_{Sliding}$	**45.8**	**49.1**	**39.4**	**30.4**	**41.4**	**19.2**	**80.4**	**91.6**	**80.5**	**2.2**

Blue rows are CoRank — it wins on every backbone, with and without sliding windows. RankZephyr is the strongest supervised reranker; RankVicuna and RankMistral score lower. Trec-Covid R@10 is low by construction.



Every component matters

Fig. 4. Solid bars = score after removing that component; pale bars = full CoRank. Dropping *category* or *fine-grained reranking* hurts most — the feature types carry complementary signal.

5 Building the Compact Representations

UNSTRUCTURED PAPER ~200 tokens, up to 5,774

Full text
TinyBERT: Distilling BERT for Natural Language Understanding. Language model pre-training, such as BERT, has significantly improved the performances of many natural language processing tasks. However, pre-trained language models are usually computationally expensive, so it is difficult to efficiently execute them on resource-restricted devices ...

→ zero-shot LLM IE (offline)

STRUCTURED FEATURES extracted once, then cached

CATEGORY — 3-LEVEL PATH
NLP → Language Models → Knowledge Distillation

SECTION
Two-Stage Learning Framework for Enhanced Knowledge Transfer

PSEUDO QUERY
"How to compress BERT with transformer distillation?"

KEYWORDS
Distillation · Transformers · Embedding

→ combine

REPRESENTATION OPTIONS coarse → fine granularity

0 Full text
1 Query
2 Category
3 Category + Section
4 Category + Section + Keywords ← CoRank

Form 0 ≈ 200 tok → Forms 1–4 = 10–50 tok

Adaptive selection at query time: keep only what matches the query by embedding similarity — 5 keywords of 30 · 1 section of 3 · 1 pseudo query of 20.

6 Why Compact Features Work

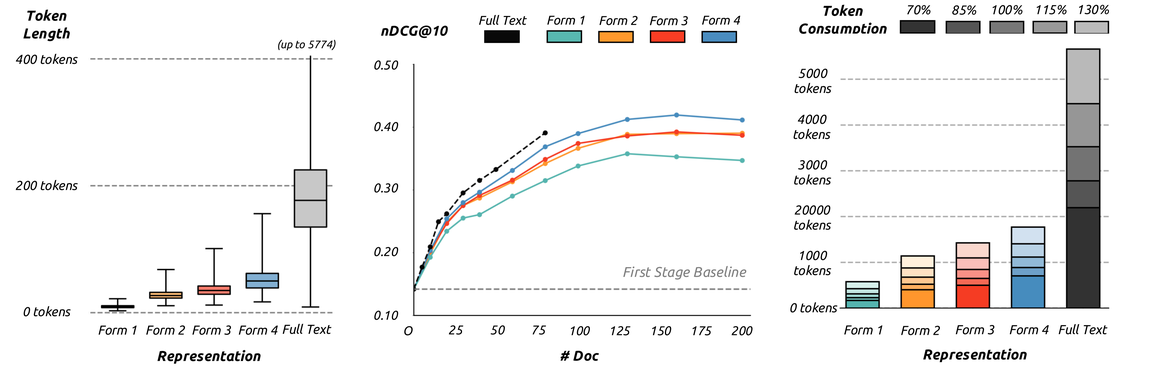


Fig. 3. (a) Tokens per document. **(b)** nDCG@10 vs. documents in one 32k context. **(c)** Tokens needed for equal reranking quality.

- Past **~80 candidates** full text overflows the context; Form 4 scales to 200 at **~1,000** tokens instead of **~3,500**.
- Below ~80 candidates full text still wins → **breadth first, then depth.**
- **Cost is a dial, not a cliff.** Sweeping 5–20 keywords or a 5–100 document pool moves nDCG@10 by only a few points.

MOLECULE MAKER LAB INSTITUTE · I-GUIDE · NSF · ACCESS

NSF awards **2505932** (MMLI), **IIS 25-37827** and **2118329** (I-GUIDE); computation on Delta/DeltaAI via ACCESS **#250851**.

Scan for paper & code → github.com/Rachum-thu/CoRank