Chain of Retrieval: Multi-Aspect Iterative Search Expansion and Post-Order Search Aggregation for Full Paper Retrieval

Anonymous ACL submission

Abstract

Scientific paper retrieval, particularly framed as document-to-document retrieval, aims to identify relevant papers in response to a long-form query paper, rather than a short query string. Previous approaches to this task have focused exclusively on abstracts, embedding them into dense vectors as surrogates for full documents and calculating similarity between them. Yet, abstracts offer only sparse and high-level summaries, and such methods primarily optimize one-to-one similarity, overlooking the dynamic relations that emerge across relevant papers during the retrieval process. To address this, we propose Chain of Retrieval (CoR), a novel iterative framework for full-paper retrieval. Specifically, CoR decomposes each query paper into multiple aspect-specific views, matches them against segmented candidate papers, and iteratively expands the search by promoting top-ranked results as new queries, thereby forming a tree-structured retrieval process. The resulting retrieval tree is then aggregated in a post-order manner: descendants are first combined at the query level, then recursively merged with their parent nodes, to capture hierarchical relations across iterations. To validate this, we present SciFullBench, a large-scale benchmark providing both complete and segmented contexts of full papers for queries and candidates, and results show that CoR significantly outperforms existing retrieval baselines.

1 Introduction

Information Retrieval (IR) is the task of searching for query-relevant documents from a large external corpus, evolving from sparse keyword matching (Sparck Jones, 1972; Robertson et al., 1995) to dense representation-based similarity (Karpukhin et al., 2020; Izacard et al., 2021). Notably, in the era of Large Language Models (LLMs) (Achiam et al., 2023; Team et al., 2023; Dubey et al., 2024; DeepSeek-AI et al., 2025), IR has become increasingly important, which allows LLMs to utilize up-to-date external information (Lewis et al., 2020).

In contrast to conventional retrieval tasks, whose queries are short (such as questions or keywords), scientific paper retrieval poses unique challenges. Specifically, queries are structured, long-form documents that encapsulate diverse aspects, ranging from research motivation and proposed methodology to experimental design and empirical findings. Also, relevance in this setting is inherently multi-faceted, as a paper may be considered relevant for various reasons, such as pursuing similar research objectives or employing comparable methods.

However, previous work has focused on abstract-level representations as a proxy for full papers (Cohan et al., 2020; Yasunaga et al., 2022; Ostendorff et al., 2022; Singh et al., 2023; Zhang et al., 2023a), fine-tuning models on citation-linked pairs of papers, each represented by its abstract and title. While they are effective to some extent, abstracts are inherently sparse and offer only high-level summaries, making them insufficient to capture the nuanced, multi-faceted relationships between scientific works. As a result, abstract-based approaches are limited to shallow similarity and struggle with tasks requiring deeper contextual understanding of full papers beyond surface-level abstracts, such as identifying complementary works, synthesizing literature reviews, or generating novel research directions (Asai et al., 2024; Baek et al., 2024; Chamoun et al., 2024; Jin et al., 2024; D'Arcy et al., 2024).

Nevertheless, it is non-trivial to represent and retrieve full scientific papers. On the one hand, full papers often exceed 100K tokens in length, far surpassing the context limits of most embedding models. Even if encoded into a single vector, representing its diverse aspects (e.g., motivation, methods, and experiments) within a single embedding may lead to oversimplification and blur fine-grained distinctions, particularly when relevance is tied to a specific aspect. Moreover, existing approaches typi-

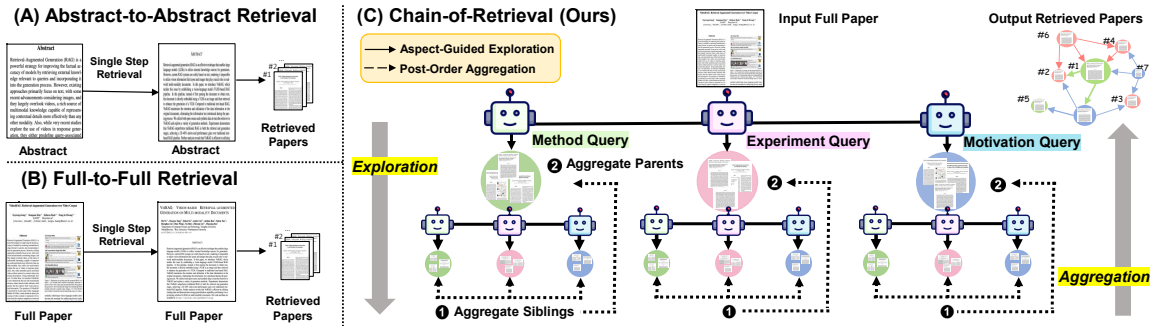

Figure 1: Conceptual illustration of our retrieval method (C) compared to prior retrieval approaches (A and B).

cally optimize semantic similarity between isolated pairs of papers, overlooking the broader, collective relations that emerge across relevant works during retrieval. This design stands in contrast to human information-seeking behavior in the information foraging theory (Pirolli and Card, 1995, 1999), which is dynamic and iterative: humans continuously refine their understanding by comparing newly encountered works with prior knowledge.

To address this, we propose Chain of Retrieval (CoR), a novel framework designed to capture dynamic inter-paper relevance while leveraging the full paper context, operated through a two-phase process: *Exploration* and *Aggregation*, illustrated in Figure 1. In the *Exploration* phase, the query paper is decomposed into multiple aspect-specific views, and each aspect is handled by a specialized query optimizer, which is further trained via Direct Preference Optimization (DPO) (Rafailov et al., 2023) on a self-generated preference to mitigate the shortage of realistic supervision for multi-retrieval systems. After retrieving with these aspect-centric queries, the retrieved (non-duplicate) papers most semantically aligned with the original query are promoted for the next iteration, recursively expanding the search with a tree structure.

In the *Aggregation* phase, results across different depths are merged to capture higher-order relations among retrieved works. Specifically, inspired by the notion of triadic closure in the network theory (Granovetter, 1973), we design an algorithm that first merges top-k results among sibling nodes (papers retrieved from the same parent), then combines these aggregated results with those of their parent node, and continues this process bottom-up until convergence at the root. This hierarchical aggregation reinforces strong multi-hop connections while naturally attenuating weaker, indirect signals, yielding a robust ranking of relevant papers.

We evaluate CoR by extending existing paper retrieval benchmarks, as they are originally designed for abstract-only retrieval, adapting them to include full papers and more recent publications (to mitigate the potential model contamination) in the ML and NLP domains, which we refer to as SciFullBench. Across experiments on SciFullBench, CoR consistently outperforms retrieval baselines (that either rely on abstracts or naively encode full papers), while remaining compatible with diverse, domain-agnostic embedding models, highlighting both the robustness and generality of our CoR.

2 Related Work

Scientific Paper Retrieval Scientific paper retrieval aims to identify relevant works given a query, which may take the form of short queries (Ajith et al., 2024), research proposals (Garikaparthi et al., 2025), or full papers, and the latter being the focus of this work. Pioneering work in scientific paper-to-paper retrieval used bibliometric statistics, such as influence scores (Zhou et al., 2012; Mohapatra et al., 2019), co-citations (Small, 1973; Haruna et al., 2018), and bibliographic coupling (Kessler, 1963). Recently, thanks to the capability of neural models, many studies have focused on calculating semantic similarities between abstracts of respective documents (Bhagavatula et al., 2018; Ostendorff, 2020), with the embedding models optimized to this domain. For example, Cohan et al. (2020) and Ostendorff et al. (2022) fine-tune BERT-based models (Devlin et al., 2019) using abstract pairs extracted from the citation graph, and Mysore et al. (2022) further consider the Wasserstein distance between sentence segments within abstract pairs. Moreover, recent studies target multiple tasks (such as paper classification and citation prediction in addition to paper retrieval) in a unified framework by considering their respective representations (Singh et al., 2023; Zhang et al., 2023a). Complementary to these, hybrid approaches integrate bibliomet-

ric signals with vector-based retrieval, reweighting ranks via post-hoc adjustments (Hamedani et al., 2016; Guo et al., 2022). However, such statistical indicators (e.g., citations) primarily reflect long-term scholarly accumulation and tend to capture retrospective importance rather than the contextual information expressed in the documents themselves. In contrast, we model paper relations dynamically during inference using aspect-driven queries generated from the full content of scientific papers.

Query Optimization with LLMs Effective formulation of queries is central to retrieval performance, and has long been studied from classical relevance feedback approaches (Rocchio Jr, 1971; Salton and Buckley, 1990) to query expansion techniques (Kuzi et al., 2016; Nogueira et al., 2019). More recent methods either leverage LLMs themselves to reformulate queries (Yu et al., 2023a; Wang et al., 2023; Gao et al., 2023), or further augment them with external query-relevant information retrieved in an auxiliary step (Yu et al., 2023b; Shen et al., 2024; Park and Lee, 2024; Lei et al., 2024). Another line of work seeks to reduce query ambiguity by decomposing a complex query into smaller queries (Zheng et al., 2024; Korikov et al., 2024). We note that our approach aligns with this decomposition paradigm but extends it to a more challenging setting: instead of handling short, user-issued queries, we decompose long-form scientific documents into multiple aspect-specific subqueries.

3 Methodology

We define the paper retrieval problem and present Chain of Retrieval (CoR), illustrated in Figure 2.

3.1 Preliminary

Paper-to-Paper Retrieval Given a query paper D (with its abstract D_{abstract}), the goal of paper retrieval is to return a ranked list of relevant papers from the corpus $\mathcal{C}$, where relevance could be defined by signals, such as citations, co-citations, and human labels. In contrast to existing studies that use D_{abstract}, we utilize its complete version D for query formulation and corpus construction.

3.2 Aspect-Aware Multi-Vector Retrieval

Aspect-Aware Query Optimization Scientific papers encapsulate multiple facets, such as research motivation, method design, and experimental validation following Baek et al. (2024) and Moussa et al. (2025), which might not be represented by a single query. To capture this diversity, we formulate the query optimization process that transforms the full paper D into a set of aspect-specific queries. Formally, we define a family of query optimization functions: $\mathcal{F} = \{f_{\text{R}}, f_{\text{M}}, f_{\text{E}}\}$, where each function f maps D to a query $q = f(D)$ that targets a specific aspect. Notably, each function is instantiated with an LLM guided by an aspect-specific template (see Appendix J). In addition to these fine-grained queries, we consider the abstract D_{abstract}, since it reflects a broad view of the overall content (that can complement specialized views), yielding the final query set: $\mathcal{Q} = \{f_i(D) \mid f_i \in \mathcal{F})\} \cup \{D_{\text{abstract}}\}$.

Retrieval with Multi-View Corpora Once the optimized query set $\mathcal{Q}$ is formulated, we perform retrieval individually for each query $q \in \mathcal{Q}$. Specifically, for the abstract-based query D_{abstract}, we use a corpus $\mathcal{C}_{\text{abstract}}$ that contains candidate abstracts (as they are comparable in length and structure). For the other aspect-specific queries derived from the full paper, we use a segmented version of the full corpus: $\mathcal{C}_{\text{chunked}}$, where each paper is split into fixed-length chunks (likely to capture its specific aspect) and indexed using multi-vector representations (Khattab and Zaharia, 2020; Santhanam et al., 2021). Each query then retrieves its top-k relevant segments, which are mapped back to their source papers via the mapping function $h(x)$, as follows: $\mathcal{R}_q = \{h(x) \mid x \in \texttt{Retrieve}(q, \mathcal{C}, k)\}$. Finally, the resulting ranked lists of candidate papers from all queries are aggregated to form the multi-view retrieval pool, denoted as follows: $\mathcal{R} = \bigcup_{q \in \mathcal{Q}} \mathcal{R}_q$.

3.3 Iterative Chain-of-Retrieval

Iterative Aspect-Aware Search Expansion The exploration phase aims to iteratively expand the search space by promoting promising retrieved results as new queries. Specifically, starting from the root query paper, we branch into aspect-specific queries (namely, motivation, methods, and experiments), using the process described in Section 3.2, while excluding the abstract view. Then, each query retrieves a set of top-k candidate papers from the corpus, which serve as the first layer of expansion. At each subsequent depth r, every aspect-specific branch selects a single most-relevant representative work from its top-k results and promotes it as the next query paper, ensuring continuity of semantic relevance while avoiding divergence from the original query, and this query paper is further decomposed into aspect-specific queries for the

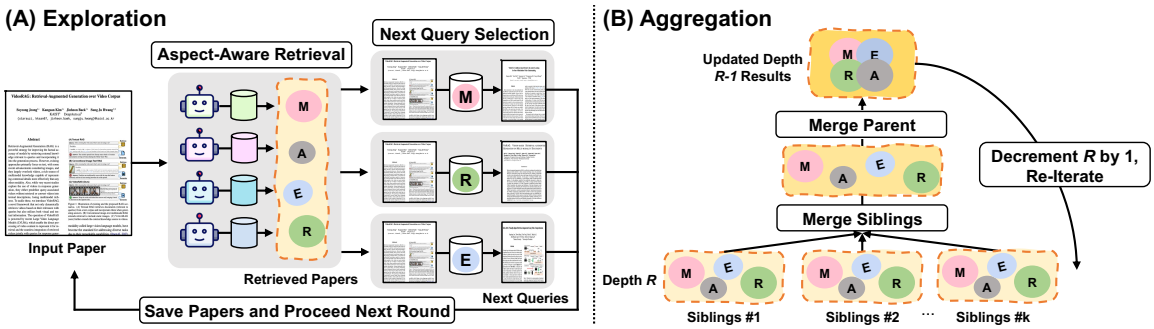

Figure 2: Overview of Chain-of-Retrieval (CoR), consisting of (A) Exploration and (B) Aggregation (formalized in Algorithm 1).

next iteration. At the end, this procedure grows a tree-structured retrieval space: after r iterations, the search expands into $4(3)^{r-1}$ paths, each corresponding to an aspect-driven exploration chain.

To avoid redundancy and ensure efficiency, we store previously selected papers in aspect-specific caches: $\textsc{Cache}[a]$ for $a \in \{\text{R}, \text{M}, \text{E}\}$, where each cache records documents with respect to the initial branching aspects of the root query paper, so that any paper stored in the corresponding cache is excluded from future expansions within that branch (to avoid repeated visits to the same works). Further, every retrieved top-k set across depths is preserved in a memory M, which latter supports aggregation across multiple levels of the search tree.

Recursive Post-Order Aggregation After R rounds of expansion, we aggregate results in a bottom-up, post-order fashion. Starting from the leaf nodes (i.e., the final retrieved sets at depth R), we first merge sibling results that share the same parent using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), defined as follows: $\text{RRF}(P) = \sum_{q \in \mathcal{Q}} \frac{1}{k + \text{rank}_q(P)}$, where $\text{rank}_q(P)$ denotes the position of paper P in the ranked list retrieved by query q, and k is a smoothing constant. This step ($\textsc{MergeSiblings}$ in Algorithm 1) emphasizes documents consistently ranked highly across different aspect queries. The merged sibling sets are then recursively combined upward with their parent results ($\textsc{MergeEdges}$), gradually consolidating evidence across the retrieval tree until the root.

It is worth noting that, through this recursive process, rank signals from deeper layers are naturally decayed: RRF suppresses the influence of lower-ranked items, and thus papers retrieved only at later depths contribute less to the final ranking. Formally, the effective weight of ranks from depth R decreases at a rate proportional to $\sigma(2\sigma)^{R-1}$, where σ is the number of query sets per retrieval. As a result, papers strongly and repeatedly supported across multiple aspects and iterations are reinforced, while weaker or spurious signals are attenuated, yielding a robust ranking for the query.

Algorithm 1 Chain-of-Retrieval (CoR).
Please see Appendix G for the detailed algorithms.

Require: Root paper D; Max depth R; Top-k per query K; Corpora $\mathcal{C} = \{\mathcal{C}_{\text{abstract}}, \mathcal{C}_{\text{chunked}}\}$
Ensure: Final top-k candidates for D

$Q \leftarrow [(D, \textsc{Root})]$
$\mathcal{M} \leftarrow [\,]$
$\textsc{Cache}[a] \leftarrow \emptyset \;\; \forall a \in \{\text{R}, \text{M}, \text{E}\}$
for $r = 0$ **to** $R - 1$ **do** ▷ **Exploration**
  $\mathcal{M}[r] \leftarrow [\,], \;\; Q_{\text{next}} \leftarrow [\,]$
  for all $(d, \textsc{parent}) \in Q$ **do**
    $P \leftarrow \textsc{OneHopRetrieval}(d, \textsc{parent}, \mathcal{C}, K)$
    $\mathcal{M}[r].\textsc{Extend}(P)$
    $Q_{\text{next}} \leftarrow \textsc{SelectNextQuery}(P, \textsc{Cache}, \gamma)$
  $Q \leftarrow Q_{\text{next}}$
for $r = R - 1$ **downto** 1 **do** ▷ **Aggregation**
  $S \leftarrow \textsc{MergeSiblings}(\mathcal{M}[r])$
  $\mathcal{M}[r-1] \leftarrow \textsc{MergeEdges}(\mathcal{M}[r-1], S)$
return $\textsc{MergeSiblings}(\mathcal{M}[0])$

3.4 Multi-Aspect Preference Optimization

Offline Policy Exploration To improve the quality of aspect-specific queries, we further train each query-generation agent f with the Direct Preference Optimization (DPO) algorithm (Rafailov et al., 2023). A key challenge, however, lies in constructing reliable agent-specific preference datasets, as collecting human-labeled preferences for multi-agent systems is both costly and impractical. Moreover, recent studies emphasize that initializing DPO with the Supervised Fine-Tuned (SFT) model is crucial to mitigate distributional shift issues (Feng et al., 2024; Xu et al., 2024; Meng et al., 2024).

To address these challenges, we follow the synthetic preference construction procedure of Meng et al. (2024), treating an off-the-shelf instruction-tuned LLM ($\mathcal{P}_\theta$) as the SFT policy ($\mathcal{P}_{\text{SFT}}$). Specifically, for each input paper D, we generate a set of k rolled-out candidate queries from every functional

agent: $Q_d^{\text{Exp}} = \bigcup_{f \in \{f_{\text{R}}, f_{\text{M}}, f_{\text{E}}\}} \{q_{(D,f)}^{(1)}, \ldots, q_{(D,f)}^{(k)}\}$. After that, among these candidates, we identify the best and worst queries according to a reward function: $q_{(D,f)}^{\text{Best}} = \arg\max_j \texttt{Reward}(q_{(D,f)}^{(j)})$ and $q_{(D,f)}^{\text{Worst}} = \arg\min_j \texttt{Reward}(q_{(D,f)}^{(j)})$. In addition, a preference pair is retained only if the reward gap between those two exceeds a margin τ, ensuring that $\mathcal{D}_{\text{pref}}$ contains discriminative signals for training.

Reward Formulation Designing an effective reward function is crucial for preference optimization, as it determines how query-generation agents are guided to improve. Our principle is to provide reward signals that are reliable and reproducible, cost-efficient to compute, and directly aligned with retrieval quality; thus, instead of relying on qualitative judgments from LLMs (LLMs-as-judge), which are costly and potentially inconsistent, we leverage citation links as environmental feedback, providing direct, objective reward signals (where we use Recall@k in practice). Also, for each agent $f \in \{f_{\text{R}}, f_{\text{M}}, f_{\text{E}}\}$, we define an independent reward function $\texttt{Reward}_f$, such that the reward for one agent does not depend on the outputs of others, under the assumption that reinforcing the quality of aspect-specific queries individually is sufficient, since the final retrieval results are aggregated across agents via a linear combination of their outputs.

4 Experiments

4.1 Experimental Setup

We describe here the setup used to validate our method. Further implementation details and training procedures are in Appendix B and Appendix C.

Datasets To evaluate full paper-to-paper retrieval, we build upon and extend existing abstract-only benchmarks, which we refer to as SciFullBench. Specifically, it preserves the original retrieval formulation and relevance definition, while augmenting each query-candidate pair with full-text content and refreshing the corpus with recent publications. Notably, we also evaluate CoR under abstract-only configurations on SciFullBench (Table 17) and on an existing benchmarks that use other relevance signals (Tables 19 and 21), as well as in the full-paper-to-full-paper variant of the same benchmark (Table 20), to ensure that our gains are not specific to the full-paper scenario and generalize across retrieval regimes under diverse relevance signals, on which CoR consistently outperforms baselines.

To build SciFullBench, we follow prior work on scientific paper retrieval (Singh et al., 2023; Medić and Snajder, 2022; Cohan et al., 2020) and define relevance using neighboring relationships in the academic citation graph, including both incoming citations and outgoing references. Query papers are collected from major ML (NeurIPS, ICLR) and NLP (ACL, EMNLP) venues using OpenReview API[1], ACL-Anthology[2], and SEA (Yu et al., 2024). We then retain papers with at least ten relevant neighbors and randomly sample 400 query papers per venue. For the retrieval corpus, we collect CS-related papers from arXiv (2020-2025)[3], where full text is publicly available, resulting in approximately 40K papers. Lastly, we post-process papers by removing reference sections, filtering citation markers, and eliminating their in-text mentions, to avoid hindsight leakage. Further dataset construction details are provided in Appendix A.

Retrieval Models We compare our method with existing retrievers developed for scientific paper retrieval, as follows: **SPECTER2 Base** (Singh et al., 2023); **SciNCL** (Ostendorff et al., 2022); **SciMult-MHAExpert** (Zhang et al., 2023a); **SPECTER2 Adapters + MTL CTRL** (Singh et al., 2023). We also consider off-the-shelf general-purpose embedding models, such as **Jina-Embeddings-V2-Base-EN** (Günther et al., 2023), **BGE-M3** (Chen et al., 2024a), and **Inf-Retriever-v1-1.5B** (Yang et al., 2025) as our primary experiments, while also utilizing **Granite-Embeddings-English-R2** (Awasthy et al., 2025), **QWEN3-Embedding-0.6B** (Zhang et al., 2025b), and **Dewey-EN-Beta** (Zhang et al., 2025a) to broaden evaluation in Appendix E and F.

Retrieval Units We consider various retrieval units, where we denote **A** for *Abstract*, **F** for *Full paper*[4], and **C** for *chunked context* (i.e., segmented units of the full paper, each capped at 3K tokens). Using them, we consider four retrieval setups: **A2A** (Abstract-to-Abstract), **F2F** (Full-to-Full), **A2C** (Abstract-to-Chunk), and **F2C** (Full-to-Chunk).

Query Optimizers We instantiate query optimizers using LLMs: the open-weight instruct models Llama-3.2-3B-Instruct (Meta, 2024) and QWEN-2.5-3B-Instruct (Yang et al., 2024; Team, 2024).

[1] https://openreview.net/
[2] https://aclanthology.org/
[3] https://arxiv.org/
[4] Since full papers often exceed the context length of embedding models, we truncate them up to the maximum length.

Table 1: Main results of various domain-specific and domain-agnostic retrievers on SciFullBench.

	Methods	ICLR-NeurIPS				ACL-EMNLP			
		References		Citations		References		Citations	
		nDCG@300	Recall@300	nDCG@300	Recall@300	nDCG@300	Recall@300	nDCG@300	Recall@300
	Lexical-Based Retrievers								
	BM-25 (A2A)	28.29	38.21	27.52	37.41	21.02	30.79	20.86	29.37
	BM-25 (F2F)	26.34	37.02	36.48	47.47	23.38	35.10	28.08	38.58
	Domain-Specific Retrievers (A2A)								
	SciNCL	36.24	51.80	33.07	47.90	26.12	40.71	25.58	38.91
	SPECTER2-Base	35.24	50.41	34.48	49.09	25.07	39.02	26.85	40.21
	SPECTER2-Adapter-MTL CTRL	36.22	51.07	33.12	47.38	25.32	38.86	25.48	38.54
	SciMult-MHAExpert	30.95	45.04	28.32	41.07	23.11	36.12	22.57	33.79
Domain-Agnostic Retrievers	**Jina-Embeddings-v2-BASE-EN**								
	Abstract-to-Abstract (A2A)	36.78	49.92	33.85	47.89	24.96	37.90	26.45	38.92
	Full-to-Full (F2F)	37.05	50.60	35.80	49.63	27.55	41.69	28.34	40.65
	CoR w/ Llama-3.2-3B-Instruct	**38.88**	**55.50**	**38.58**	**56.21**	**28.34**	**44.44**	**31.80**	**47.03**
	CoR w/ QWEN-2.5-3B-Instruct	**38.91**	**55.78**	**39.36**	**56.97**	**28.21**	**44.83**	**32.54**	**48.30**
	BGE-M3								
	Abstract-to-Abstract (A2A)	32.08	44.09	30.31	42.85	22.95	34.55	24.41	36.28
	Full-to-Full (F2F)	33.71	45.84	32.22	44.01	24.13	35.88	24.85	35.87
	CoR w/ Llama-3.2-3B-Instruct	**33.81**	**49.39**	**34.54**	**49.27**	**24.62**	**39.22**	**28.87**	**42.67**
	CoR w/ QWEN-2.5-3B-Instruct	**34.84**	**50.43**	**35.72**	**50.77**	**25.22**	**39.84**	**29.33**	**43.72**
	Inf-Retriever-v1-1.5B								
	Abstract-to-Abstract (A2A)	45.84	61.72	38.61	54.63	33.43	50.10	30.94	45.53
	Full-to-Full (F2F)	31.09	42.27	35.58	48.79	32.86	48.00	30.31	43.64
	CoR w/ Llama-3.2-3B-Instruct	**47.31**	**65.31**	**43.66**	**61.54**	**34.48**	**53.63**	**36.18**	**52.95**
	CoR w/ QWEN-2.5-3B-Instruct	**47.21**	**64.78**	**44.21**	**62.06**	**34.41**	**53.30**	**36.87**	**53.59**

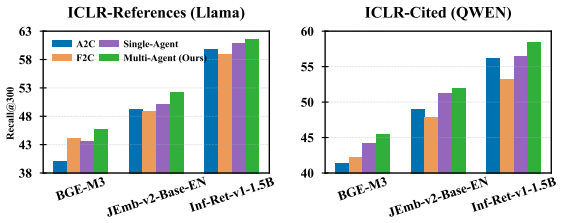

Figure 3: Retrieval performance comparing different query formulations against our multi-agent CoR framework (after a single depth of retrieval and with untrained query optimizers).

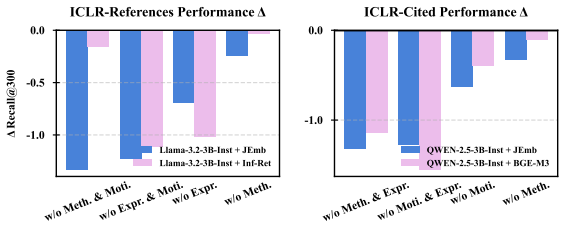

Figure 4: Change (Δ) in retrieval performance (relative to using all aspects) when excluding individual scientific aspects (Moti = Motivation, Meth = Method, Expr = Experiment).

For reproducibility, we set the temperature to 0 and the maximum generation tokens to 2,000.

Training and Inference Configuration For each query, we retrieve the top-300 candidate papers. To prevent repeatedly selecting the same candidate as the next query, we always skip the highest-ranked results (selected previously) and instead promote the next-best. During preference pair construction, we generate 16 candidate queries per aspect and evaluate them based on Recall@30, while excluding pairs whose reward margin is smaller than 3%.

4.2 Experimental Results and Analysis

We now provide an in-depth analysis of the experimental results. Please refer to Appendix E and Appendix F for more experiments and analyses.

Main Results Table 1 presents the main results, where CoR outperforms all baselines across various settings, validating the effectiveness of our proposed framework for full paper-to-paper retrieval[5]. Notably, when using the same domain-agnostic retriever, CoR surpasses abstract-to-abstract (A2A) baselines by an average of 6.37% in Recall, demonstrating that simply relying on abstracts is suboptimal compared to our aspect-driven approach. Also, interestingly, general-purpose embedding models such as Jina-Embeddings-V2-Base-EN and BGE-M3 (whose A2A performance lags behind representative domain-specific retrievers) achieve competitive or superior performance once integrated into CoR, emphasizing that CoR delivers strong performance regardless of the retrievers by serving as a plug-and-play framework. Lastly, when compared against F2F retrieval with long-context embedding models, CoR achieves average gains of 7.82% in Recall, confirming that simply embedding entire papers is insufficient, and that structured, iterative search expansion with post-order aggregation is crucial for effective full paper-to-paper retrieval.

Effectiveness of Specialized Query Optimizers Our results in Figure 3 confirm the effectiveness of using multiple specialized LLM agents, each dedicated to a single aspect. More specifically, using multiple specialized agents outperforms a single base agent tasked with generating multiple queries

[5]We report results on lower @K values and other relevance signals in Appendix F, where CoR yields consistent gains.

Table 2: Performance comparison with and without the Aspect-Aware Cache in the selection algorithm, reported after retrieval depths of three using DPO-trained query optimizers. We report the mean and standard deviation over three independent runs, and underline statistically significant improvements.

{Optimizer} + {Retriever}	ICLR-References Recall@300	ICLR-Citations Recall@200
Llama-3.2-3B-Inst + JEmb-v2		
w/o Aspect-Aware Cache	53.63 ± 0.11	46.80 ± 0.03
w/ Aspect-Aware Cache	**54.15** ± 0.39	**47.08** ± 0.04
Llama-3.2-3B-Inst + BGE-M3		
w/o Aspect-Aware Cache	46.88 ± 0.16	40.82 ± 0.18
w/ Aspect-Aware Cache	**47.35** ± 0.13	**41.16** ± 0.27
Llama-3.2-3B-Inst + Inf-Ret-v1		
w/o Aspect-Aware Cache	62.65 ± 0.10	51.83 ± 0.05
w/ Aspect-Aware Cache	**62.98** ± 0.19	**52.31** ± 0.11

(without aspect separation). Also, when contrasted with setups that use either full papers or abstracts as queries (matched with the same chunked corpus) and that retrieve the same number of total candidates as CoR of a single retrieval round, our multi-aspect queries generated with specialized agents achieve gains of 3.36% and 3.33%, respectively, highlighting the advantage of decomposing noisy, monolithic queries into aspect-aware subqueries.

Importance of Multi-Aspect Coverage To examine the role of aspect diversity in query formulation, we analyze how the retrieval performance changes as we vary the set of participating query optimizer agents. As shown in Figure 4, excluding any single aspect leads to measurable performance degradation, which supports the necessity of capturing the multifaceted nature of scientific papers. Not all aspects, however, contribute equally: excluding experimental or research-motivation queries causes larger drops in performance than omitting methodological queries. We attribute this to the inherent characteristics of each aspect. Specifically, experimental and research-motivation views often yield stronger semantic overlap across related papers, grounded in shared entities such as tasks, baselines, and datasets. In contrast, methodological details are more heterogeneous and less entity-driven, demanding deeper semantic abstraction and reasoning to identify commonalities.

Role of Aspect-Aware Cache Table 2 shows the benefits of incorporating the aspect-aware cache into our *Next Query Selection* strategy (described in Section 3.3), which enables preventing redundant exploration within each aspect branch and thus encouraging more diverse search trajectories. From this, CoR with the aspect-aware cache consistently outperforms those without it, demonstrating its effectiveness in steering the retrieval process toward broader coverage and stronger overall performance.

Table 3: Performance comparison of specialized agents trained with DPO versus untrained query optimizers, evaluated after a single round of retrieval (i.e., one depth). We report the mean and standard deviation across three independent trials, and underline statistically significant improvements.

{Optimizer} + {Retriever}	ICLR-References Recall@100	ICLR-References Recall@300	NeurIPS-Citations Recall@100	NeurIPS-Citations Recall@300
Llama-3.2-3B-Inst + JEmb-v2				
w/o DPO	36.20 ± 0.00	52.14 ± 0.00	41.70 ± 0.00	57.36 ± 0.00
w/ DPO	**37.10** ± 0.12	**52.60** ± 0.05	41.33 ± 0.06	**57.95** ± 0.10
Llama-3.2-3B-Inst + BGE-M3				
w/o DPO	31.32 ± 0.00	45.61 ± 0.00	36.24 ± 0.00	50.85 ± 0.00
w/ DPO	**31.59** ± 0.02	**45.91** ± 0.02	**36.55** ± 0.07	50.49 ± 0.09
Llama-3.2-3B-Inst + Inf-Ret-v1				
w/o DPO	45.15 ± 0.05	61.75 ± 0.05	47.91 ± 0.00	63.04 ± 0.00
w/ DPO	**45.92** ± 0.06	**62.15** ± 0.06	**48.57** ± 0.05	**63.61** ± 0.07
QWEN-2.5-3B-Inst + JEmb-v2				
w/o DPO	36.72 ± 0.00	52.43 ± 0.00	41.82 ± 0.00	58.20 ± 0.00
w/ DPO	**37.41** ± 0.17	**53.00** ± 0.10	**42.21** ± 0.15	**58.64** ± 0.12
QWEN-2.5-3B-Inst + BGE-M3				
w/o DPO	30.94 ± 0.00	45.68 ± 0.00	37.09 ± 0.00	51.23 ± 0.00
w/ DPO	**31.52** ± 0.14	**46.60** ± 0.04	**37.75** ± 0.15	**52.08** ± 0.21

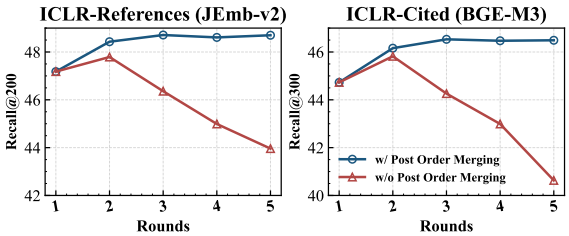

Figure 5: Retrieval results as a function of retrieval depths, with DPO-trained Llama-3.2-3B-Instruct as optimizers.

Efficacy of Query Preference Optimization To assess the contribution of reinforcing aspect-aware query optimizers with DPO, we conduct an ablation study isolating the DPO training. As shown in Table 3, integrating DPO consistently improves performance across diverse LLM-retriever configurations in both the reference and citation setups. Notably, although the preference pairs are constructed in the Reference (outgoing-links) setting using Recall as the reward signal, the gains transfer to other evaluation setups of Citation (incoming-links).

Analysis on Post-Order Aggregation To verify whether our aggregation mechanism is responsible for performance improvements beyond those obtained from multi-aspect expansion, we conduct an analysis. As shown in Figure 5, CoR yields performance gains across successive retrieval rounds, with the gap over baselines widening as depth increases. This confirms the effectiveness of our post-order aggregation, which progressively attenuates weak signals from distant connections and thereby enhances robustness against noise in deeper rounds. By contrast, the baseline strategy (merging results from all branches at once) fails to adaptively penalize such signals, leading to steadily diminishing performance. However, improvements plateau beyond depth 3, which suggests diminishing returns, likely due to added noise with deeper expansions.

Table 4: Analysis on the semantic coverage of the Top-300 retrieved documents over ground truth documents with Convex Hull Volume. The results of CoR are reported after 3 rounds (i.e., depth) with Llama-3.2-3B-Inst as query optimizers.

	ACL (Citations)	NeurIPS (Citations)
{Type} w/ {Retriever}	CHV Ratio (Δ)	CHV Ratio (Δ)
Relative Coverage over A2A		
F2F w/ JEmb-v2	1.126 (+12.6%)	1.135 (+13.5%)
F2F w/ BGE-M3	1.088 (+8.8%)	1.069 (+6.9%)
F2F w/ Inf-Ret-v1	1.128 (+12.8%)	1.097 (+9.7%)
Relative Coverage over A2A		
CoR w/ JEmb-v2	**1.190 (+19.0%)**	**1.198 (+19.8%)**
CoR w/ BGE-M3	**1.107 (+10.7%)**	**1.097 (+9.7%)**
CoR w/ Inf-Ret-v1	**1.167 (+16.7%)**	**1.140 (+14.0%)**
Relative Coverage over F2F		
CoR w/ JEmb-v2	**1.095 (+9.5%)**	**1.078 (+7.8%)**
CoR w/ BGE-M3	**1.044 (+4.4%)**	**1.042 (+4.2%)**
CoR w/ Inf-Ret-v1	**1.060 (+6.0%)**	**1.072 (+7.2%)**

Diversity Analysis Recall that CoR is designed to expand the search space iteratively across multiple aspects of the query paper, and we further analyze this by measuring the semantic coverage of retrieved candidates over ground-truth in the embedding space using the Convex Hull Volume Ratio (CHV Ratio), which compares coverage between two configurations, with the same set (see Appendix C.3 for details). As shown in Table 4, CoR consistently improves the diversity of semantic coverage among the retrieved documents: relative to the baseline A2A setup, it achieves a 14.98% gain in CHV Ratio, surpassing the 10.72% gain of the F2F setup. Moreover, when directly compared against F2F, CoR still provides an additional 6.52% improvement. These results highlight that beyond improving retrieval accuracy, the proposed CoR enables broader semantic exploration, which allows discovering relevant papers that would likely be missed under the vanilla retrieval strategies.

Generalization to Other Domains To examine the generality of CoR beyond scientific paper retrieval, we further evaluate its performance on a different but challenging document-to-document retrieval task: patent-to-patent retrieval, which has substantially long queries (please see Appendix D for details on task design and implementation). As shown in Table 5, CoR delivers robust gains over both the domain-agnostic and domain-specific baselines, surpassing F2F retrieval by 2.01% and A2A retrieval by 6.96%. Importantly, these improvements are achieved without any training (i.e., using the off-the-shelf GPT-4o for query optimization), highlighting the versatility and broad applicability of CoR for long document-to-document retrieval.

Human Evaluation To further assess the quality of retrieved results beyond the citation-based ground truth in SciFullBench, we conduct a human evaluation. A total of 15 participants evaluate the relevance of the Top-5 retrieved candidates for three methods (A2A, F2F, CoR) across 15 query papers, assigning one of four labels to each candidate (motivation-relevant, method-relevant, experiment-relevant, or irrelevant). We then compute per-query precision by treating any of the three aspect labels as relevant and the remaining label as irrelevant. As shown in Table 6, CoR achieves the highest precision, surpassing A2A and F2F by 3.34% and 10.84%, respectively, indicating that its improvements extend beyond citation-based metrics to human-perceived relevance. To ensure the reliability of human judgments, we measure inter-annotator agreement on it with Cohen's Kappa coefficient (Cohen, 1960), obtaining a score of 0.43, which corresponds to moderate agreement and supports the reliability of our human evaluation. For further details, please refer to Appendix C.4.

Table 5: Experimental results on patent-to-patent retrieval using PatentFullBench (self-constructed; See Appendix D). Results for CoR are reported after retrieval depth of two.

	References	Citations
Methods	Recall@100	Recall@100
Baseline		
PAT-SPECTER (A2A) (Ghosh et al., 2024)	47.36	52.16
PaECTER (A2A) (Ghosh et al., 2024)	52.77	56.25
BGE-M3 (A2A)	46.42	49.57
BGE-M3 (F2F)	52.30	55.81
Inf-Ret-v1 (A2A)	52.78	57.94
Inf-Ret-v1 (F2F)	56.23	58.59
JEmb-v2-Base-EN (A2A)	49.39	53.27
JEmb-v2-Base-EN (F2F)	56.57	59.61
CoR (Ours): {Optimizer} + {Retriever}		
GPT-4o-Mini-2024-0718 + BGE-M3	**54.71**	**56.78**
GPT-4o-Mini-2024-0718 + Inf-Ret-v1	**58.62**	**63.06**
GPT-4o-Mini-2024-0718 + JEmb-v2	**57.23**	**60.74**

Table 6: Human study on the relevance of ground truth candidates. The results are reported after 3 rounds (i.e., depth) using DPO-trained QWEN-2.5-3B-Instruct models for query optimizer, and jina-embeddings-v2-base-en model for retriever.

Retrieval Methods	**Precision@5**
Abstract-to-Abstract (A2A)	60.83
Full-to-Full (F2F)	53.33
Chain-of-Retrieval (CoR; Ours)	**64.17**

5 Conclusion

In this work, we introduced Chain of Retrieval (in short, CoR), a novel framework for paper-to-paper retrieval that iteratively expands the search space via aspect-aware query optimization (reinforced with DPO training) and recursively merges results via post-order aggregation. For evaluation, we presented SciFullBench, a large-scale dataset for full-context scientific retrieval, and results show that CoR consistently outperforms the abstract-level and full-context baselines, even with off-the-shelf embedding models and for a patent retrieval task, highlighting its robustness and generality.

Limitations

While our work introduces a novel retrieval approach that iteratively expands the search space via multi-aspect query optimization over the full context of scientific papers, it still has room for future work. First, an adaptive branch pruning mechanism could be further considered during the exploration phase to more dynamically guide the search results with the semantic alignment of individual papers with the root query, which would be an interesting direction for future work. Second, while our framework retrieves diverse and relevant papers, deeper exploration and repeated LLM inference introduce additional computational overhead. While we partially mitigate this with a lightweight selection mechanism (in place of expensive LLM verifiers), further optimizations on it would be a valuable direction, which we leave as future work.

Ethics Statement

Although our Chain of Retrieval (CoR) framework improves retrieval performance over prior approaches, it can still retrieve irrelevant or misleading papers, which may propagate incorrect or harmful information to both human users and downstream AI models. To ensure the development of trustworthy automated systems, future work should incorporate robust verification or re-ranking mechanisms that can effectively filter irrelevant or potentially harmful documents from the retrieved set.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15068–15083, Miami, Florida, USA. Association for Computational Linguistics.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, and 3 others. 2025. Granite embedding models. *Preprint*, arXiv:2502.20204.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Brown and Yaoqi Zhou. 2019. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2019:baz085.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jianlyu Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, and Zheng Liu. 2024b. Air-bench: Automated heterogeneous information retrieval benchmark. *Preprint*, arXiv:2412.13102.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. 2024. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*.

Aniketh Garikaparthi, Manasi Patwardhan, Aditya Sanjiv Kanade, Aman Hassan, Lovekesh Vig, and Arman Cohan. 2025. MIR: Methodology inspiration retrieval for scientific research problems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28614–28659, Vienna, Austria. Association for Computational Linguistics.

Mainak Ghosh, Sebastian Erhardt, Michael E. Rose, Erik Buunk, and Dietmar Harhoff. 2024. Paecter: Patent-level representation learning using citation-informed transformers. *Preprint*, arXiv:2402.19411.

Ronald L. Graham. 1972. An efficient algorithm for determining the convex hull of a finite planar set. *Info. Proc. Lett.*, 1:132–133.

Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, and 1 others. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Hongmei Guo, Zhesi Shen, Jianxun Zeng, and Na Hong. 2022. Hybrid methods of bibliographic coupling and text similarity measurement for biomedical paper recommendation. In *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation*, pages 287–291. IOS Press.

Masoud Reyhani Hamedani, Sang-Wook Kim, and Dong-Jin Kim. 2016. Simcc: A novel method to consider both content and citations for computing similarity of scientific papers. *Information Sciences*, 334:273–292.

Khalid Haruna, Maizatul Akmar Ismail, Abdullahi Baffa Bichi, Victor Chang, Sutrisna Wibawa, and Tutut Herawan. 2018. A citation-based recommender system for scholarly paper recommendation. In *Computational Science and Its Applications–ICCSA 2018: 18th International Conference, Melbourne, VIC, Australia, July 2-5, 2018, Proceedings, Part I 18*, pages 514–525. Springer.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.

Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. 2019. A scalable hybrid research paper recommender system for microsoft academic. In *The world wide web conference*, pages 2893–2899.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Maxwell Mirton Kessler. 1963. Bibliographic coupling between scientific papers. *American documentation*, 14(1):10–25.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Anton Korikov, George Saad, Ethan Baron, Mustafa Khan, Manav Shah, and Scott Sanner. 2024. Multi-aspect reviewed-item retrieval via LLM query decomposition and aspect fusion. In *Proceedings of the Workshop Information Retrieval's Role in RAG Systems (IR-RAG 2024) co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024), Washington DC, USA, 07 18, 2024*, volume 3784 of *CEUR Workshop Proceedings*, pages 23–33. CEUR-WS.org.

Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1929–1932.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph 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*, pages 611–626.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *Preprint*, arXiv:2107.00414.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. 2024. Corpus-steered query expansion with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–401, St. Julian's, Malta. 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, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zoran Medić and Jan Snajder. 2022. Large-scale evaluation of transformer-based article encoders on the task of citation recommendation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 19–31, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Dattatreya Mohapatra, Abhishek Maiti, Sumit Bhatia, and Tanmoy Chakraborty. 2019. Go wide, go deep: Quantifying the impact of scientific papers through influence dispersion trees. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 305–314. IEEE.

Hanane Nour Moussa, Patrick Queiroz Da Silva, Daniel Adu-Ampratwum, Alyson East, Zitong Lu, Nikki Puccetti, Mingyi Xue, Huan Sun, Bodhisattwa Prasad Majumder, and Sachin Kumar. 2025. Scholareval: Research idea evaluation grounded in literature. *arXiv preprint arXiv:2510.16234*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - a test collection of computer science research articles for faceted query by example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Malte Ostendorff. 2020. Contextual document similarity for content-based literature recommender systems. *arXiv preprint arXiv:2008.00202*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. *arXiv preprint arXiv:2202.06671*.

Jeonghyun Park and Hwanhee Lee. 2024. Conversational query reformulation with the guidance of retrieved documents. *arXiv preprint arXiv:2407.12363*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Peter Pirolli and Stuart Card. 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–58.

Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review*, 106(4):643.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and 1 others. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*.

Gerard Salton and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4):288–297.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15933–15946, Bangkok, Thailand. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Henry Small. 1973. Co-citation in the scientific literature: a new measure of the relationship between documents. *J. Am. Soc. Inf. Sci.*, 42:676–684.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Junhan Yang, Jiahe Wan, Yichen Yao, Wei Chu, Yinghui Xu, and Yuan Qi. 2025. inf-retriever-v1 (revision 5f469d7).

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023b. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Dun Zhang, Panxiang Zou, and Yudong Zhou. 2025a. Dewey long context embedding model: A technical report. *Preprint*, arXiv:2503.20376.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang, and Jianfeng Gao. 2023a. Pre-training multi-task contrastive learning models for scientific literature understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12259–12275, Singapore. Association for Computational Linguistics.

Yu Zhang, Bowen Jin, Xiusi Chen, Yanzhen Shen, Yunyi Zhang, Yu Meng, and Jiawei Han. 2023b. Weakly supervised multi-label classification of full-text scientific papers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3458–3469.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*.

Yan-Bo Zhou, Linyuan Lü, and Menghui Li. 2012. Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New Journal of Physics*, 14(3):033033.

A SciFullBench

In this section, we provide more details on the construction process of SciFullBench, with comparison against existing benchmarks. The detailed statistics can be found in Tables 7, 8.

A.1 Construction Procedure

Step 1: Collect Data We collected research papers from leading machine learning venues including ICLR 2024, 2025, NeurIPS 2024, 2023, ACL 2024, 2023, and EMNLP 2024, 2023 for our query documents. We scraped existing publications from the main conference for ACL and EMNLP, including both long- and short- papers from the ACL-Anthology website[6], and used the OpenReview[7] API to obtain submitted research papers for ICLR and NeurIPS, using a subset of documents uploaded by authors from SEA (Yu et al., 2024). Afterwards, we obtain metadata from the arXiv database uploaded on the Kaggle[8] website, with papers from 2020 to January of 2025 that belong in the machine learning domain.

Step 2: Build a Raw Dataset We subsequently collect the information of papers that cite our potential query documents using the Semantic Scholar API[9] and check the existence of a paper in the arxiv database, and rule out any potential query documents that had fewer than ten remaining candidates after such process. We then use Allen AI Science Parse[10] released under Apache 2.0 license to obtain reference information for respective documents, and filter documents based on the implemented criteria from above. Using such data, we formulate our raw (query document, gold candidate) pairs where its abstract information is intact and where more than 10 deduplicated candidates exist, organized by splits in respective years belonging to each venue. For each split, we aggregate all filtered (query, candidate) set from each venue without distinction of its published year. From this filtered pool, we randomly sampled 500 potential queries from each split and formulated a corpus containing all ground truth candidates of sampled queries, keeping our corpus at manageable size. Next, we collected the pdf files of the papers, using the Google Cloud API[11] to access the full paper dump directly and collect the latest updated versions for each paper.

Step 3: Filter and Sanitize Dataset Using the Allen AI Science Parse tool, we parse the query and candidate pdf files collected in Step 2 into a json format file containing title, abstract, main content list, and reference section along with its inline-citation mentions for both queries and candidates. Since we use citation information as the main signal for measuring document similarities, excluding the in-line citations and the reference section was crucial to validate the fairness of our benchmark. Although our parsing tool generally separates reference from main content, reference information is frequently intermixed within the main content.

Thus, we applied an auxiliary filtering algorithm to remove such issues. Since our tool parses a pdf file into a structured json file, we were able to obtain text information segmented into multiple passages. Hence, prior to reformatting it back to complete text, we removed sections or headers that contain at least one reference title information. To mitigate concerns regarding the severe loss of information from reformatted query papers, we excluded documents in which such problematic passages made up the majority. Furthermore, we ensured to remove any residing reference title information within the formatted text.

Step 4: Finalize Benchmark Based on the preprocessed full document, we formulate a pseudo-definitive benchmark consisting of (query, ground-truth candidate), and target corpus set by random sampling 400 query documents per split in each venue that still 10 gold candidates residing subsequent to the filtering procedure. Ultimately, we formulate a total of 3200 query documents, along with large-scale target corpora consisting of all the gold candidates of such queries. Finally, we remove citations and interlinked mentions from the entire set of formatted papers and finalize our benchmark using the information provided from our parsing tool, capturing the patterns using the Python regular expression. Consequently, we construct final benchmark in which all documents for both query and candidate corpus consist of title, abstracts, full-paper text, and list of paper segments.

[6] https://aclanthology.org/
[7] https://openreview.net/
[8] https://www.kaggle.com/
[9] https://www.semanticscholar.org
[10] https://github.com/allenai/science-parse
[11] https://cloud.google.com/apis?hl=ko

Table 7: Statistics of query papers in the SCIFULLBENCH benchmark.

	ICLR		NeurIPS		EMNLP		ACL	
	References	Citations	References	Citations	References	Citations	References	Citations
domain	ML		ML		ML/CL		ML/CL	
years included	2024,2025		2023,2024		2023,2024		2023,2024	
Maximum # of Ground Truth Candidates	62	750	133	763	160	583	60	320
# of papers per year	(141,259)	(324,76)	(170, 230)	(353,47)	(226, 174)	(334, 66)	(184,216)	(256,144)
# of tokens per paper	9402.35	8183.37	11184.39	9035.43	5908.36	6226.45	6190.06	6614.6
average # of candidates per sample	21.74	42.25	24.58	40.37	19.64	39.24	18.05	33.89
minimum # of candidates per sample	10	10	10	10	10	10	10	10

Table 8: Statistics of the corpus in the SCIFULLBENCH.

	SCIFULLBENCH
Total # of papers	40,782
Avg. # of tokens per abstract	200.39
Avg. # of tokens per paper	6987.67
Total # of segmented corpus	115,044
Avg. # of tokens per segment	2479.90

A.2 Comparison with Prior Benchmarks

Previous benchmarks for scientific literature search, where the full context is available for both query and candidates, rarely exist. Kanakia et al. (2019) introduce expert-annotated paper recommendation benchmark using abstract and citations within Microsoft Academic Graph (MAG). Cohan et al. (2020) and Singh et al. (2023) reveal an evaluation set to search relevant scientific literature within a pool of 30 candidates per query document with 5 gold labels in the computer science domain, while Medić and Snajder (2022) disclose a benchmark with 60 candidates per query from 19 scientific fields sourced from MAG. Brown and Zhou (2019) presents RELISH, which also provides expert-annotated ground-truth candidates in the biomedical domain.

In addition, Mysore et al. (2021) formulates a CFSCube benchmark to evaluate fine-grained sentence-wise alignment between abstract passages. Wu et al. (2024) also presents SciMMIR, a multi-modal document retrieval evaluation set to evaluate the performance of figure-wise document retrieval frameworks. However, such benchmarks do not include candidates or queries in which their entire lexical content is fully disclosed. Although Zhang et al. (2023b) intend to evaluate their classifier pipeline using the complete scientific context, it contains five potential candidates per query, which is infeasible to evaluate large-scale literature retrieval frameworks.

Meanwhile, Lauscher et al. (2021) provides full context with fine-grained citation-intent annotation on the paper's citation context (sentences), to evaluate the model's ability to accurately identify citation intents. However, its scope is inherently limited to a trivial classification task, and does not reveal metadata information (especially full context) on the cited papers nor properly filtered and preprocessed. Conversely, our proposed benchmark consists of fully-disclosed document context, along with massive number of target candidates, adhering to the realistic setting for evaluating scientific literature retrieval.

B Training Procedures

In this section, we provide details for the training data construction process, followed by the preference set construction and detailed settings, such as hyper-parameters and training environment.

B.1 Query and Corpus Construction

Composition Our input query document for offline rollout includes publications and submissions from leading venues in ML, with different published or submitted years from SCIFULLBENCH. We incorporate publications and submissions from **ICLR** 2017 to 2023, **NeurIPS** 2016 to 2022 from the dataset revealed in REVIEWER2 (Gao et al., 2024), **ACL** 2017 to 2022, **EMNLP** 2017 to 2022, and **NAACL** 2017 to 2022. As for the target candidate corpus construction, we follow a similar approach to SciFullBench in which we crawled ML papers in the arXiv database. Ultimately, we formulate 15K input query documents with at least 10 ground-truth candidates, along with 97k target corpus consisting of all the ground-truth candidates of respective input query documents.

Formulation Workflow The formulation procedure for building the training set is similar to the process explained in Appendix A, with several key differences.

First, to prepare query documents that do not overlap with our SCIFULLBENCH benchmark (to

prevent data leakage), we collected publications and submissions from different venues (or years) from ones in our evaluation set. Yet, duplicate papers still may exist in both the benchmark and train set since research papers can be resubmitted to other venues if rejected. To preclude such corner cases, we excluded query documents from our train set where its abstracts match the ones in query document of SCIFULLBENCH.

Second, compared to the benchmark construction process that filters query documents with at least 10 ground truth documents for each citation and reference split, we view ground truth candidates as an aggregation of both relevance signals. Moreover, since the number of our input query documents drastically exceeds those in SCIFULLBENCH, our target candidate corpus size inevitably becomes larger. Thus, we randomly sampled 140K samples from the initial filtered target corpus, and reformulated our queries accordingly with at least 10 candidates.

B.2 Preference Set Construction

Using the query documents and the formulated corpus from Appendix B.1, we exclusively used references (out-going links) as ground truth annotations. To ensure fairness in evaluating the effectiveness of our algorithm, we excluded potential query documents from our training set that exist in the corpus used for the inference setting, as the documents in the corpus are used as query for later rounds of retrieval. Moreover, to ensure that data leakage does not occur, we additionally checked the SciFullBench queries for those whose title information is available, and additionally filtered queries in our train set with similar or same titles.

During the rollout, we set the maximum input tokens to 60k and the maximum completion tokens to 2000. Moreover, we set the temperature to 0.7 and the nucleus sampling ratio to 0.8 to promote exploration, with the branching factor of 16 per aspect-aware LLM agent, using a single 48GB NVIDIA A6000 GPU with VLLM gpu memory utilization rate of 0.8 for each respective LLM agent + Retriever setup. Note that we arbitrarily stopped exploration under our notion that enough preference data was accumulated through self-exploration to enhance training efficiency. Subsequently, we measured the recall@30 reward of generated queries and paired the query response with the highest reward and the lowest reward, sampling only those with a recall@30 difference of at least 3%.

B.3 Training Details and Hyper-Parameters

We trained two instruct models, Llama-3.2-3B-Instruct and QWEN-2.5-3B-Instruct, using the Unsloth (Daniel Han and team, 2023) library to reduce memory consumption. We trained the respective query optimizer agents with LoRA (Hu et al., 2022) adapters and 4-bit quantization on the default models. In addition, we set the maximum total number of tokens to 4000, allocating 2000 tokens each for input and output. This ensures that the responses from our preference set match the output token limits in the inference setup, allowing us to exclusively assess the effect of training rather than other factors. Furthermore, we trained each LLM agent with a linear scheduler, a learning rate of 1e-5, a weight decay rate of 0.01, and a maximum gradient norm of 0.6 for 3 epochs, with a global batch size of 32 on a single NVIDIA A5000 GPU with 24GB size VRAM, while occasionally using NVIDIA A6000 GPU with 48GB size VRAM to speedup training process. Under such setup, it required approximately 14 hours to train a single agent. As for the LoRA configuration, we set both the rank and α scale to 64, and applied the adapters to both the attention layers and the feedforward network, without leveraging dropout or bias factors.

C Experiment Details

In this section, we provide detailed setup for experiments, including prompts, neural retrievers, query optimizers, experiment environment, metrics, and implementation details. We also provide a detailed description for human evaluation results in Table 6.

C.1 Inference Settings

Prompt Construction We constructed prompts for scientific paper-to-paper retrieval based on the three major aspects that comprise a scientific paper: **Methodology**, **Experiments**, and **Research Questions**, inheriting the approaches in conventional research for the AI-for-Scientific Discovery domain (Baek et al., 2024; Wang et al., 2024; Moussa et al., 2025), widely regarded as the most common and foundational aspects that constitute scientific papers. The initial raw prompts were first created using ChatGPT, and human-in-the loop revision was conducted, to devise the optimal prompt suitable for our framework. The prompts used for COR can be found in Appendix J.

Retrieval Setting We primarily use the Euclidean distance to measure similarities between

our adaptively generated queries and candidates, where we acquired candidates specifically with the minimum L2 distance. Since minimizing the L2 distance was objective for most of our baseline domain-specific embedding models such as SciNCL, SPECTER2-Base, SPECTER2-Adapter-MTL CTRL, we matched such settings when we used jina-embeddings-v2-base-en, BGE-M3, and Inf-Retriever-v1-1.5b. However, for SciMult, we measured MIPs (Maximum Inner Product), since it was trained to maximize MIPs in the original paper. In addition, we utilize the FAISS library (Douze et al., 2024), which enables efficient retrieval from a large-scale corpora. We implemented an efficient L2 search using FlatL2 and FlatIP for MIPs.

In addition, we concatenated the title and abstract information for our baseline A2A retrieval settings, following the conventional retrieval setup from (Cohan et al., 2020; Singh et al., 2023) for both query and target candidates. As for our CoR framework, when using abstracts as query and candidates, we use the concatenated version of title and abstracts for both query and candidates, with the exception of root depth to balance the effect of title information of root query paper, where the query optimizers receive full paper which occasionally receives title information.

Query Optimizers We mainly used the VLLM (Kwon et al., 2023) framework for faster inference of open source models, setting the repetition penalty to 1.2. For Llama-3.2-3B-Instruct, we used the default context window size of 131072, while using YARN (Peng et al., 2023) to extrapolate the default context window size of QWEN-2.5-3B-Instruct from 32768 to 131072. Moreover, for inference, we used combinations of three NVIDIA A5000/RTX3090/4090 GPUs with 24GB VRAM size, along with vllm gpu memory utilization rate of 1.0, using fp16 precision for faster inference.

As for the inference setup for untrained query optimizer models, we used a combination of single NVIDIA A6000/RTX 4090 GPU with 48GB VRAM using the same context window and fp16 precision, along with a fixed gpu memory utilization rate of 0.7. When it comes to experiments using GPT-4o-Mini-2024-0718 and GPT-4o-2024-11-20 as query optimizers, we used temperature 0 and default hyper parameters of the OpenAI client.chat.completions API[12], while completely using the default hyperparameter settings for gpt-4.1-2025-04-14 (temperature is also set to default). Moreover, 60 is used as the hyperparameter k for Reciprocal Rank Fusion. For base agent experiments in Table 3, we equally set the hyper parameters to those used in its respective counterpart comparison groups, and are prompted to generate three different queries in a structured manner, using the Python Pydantic BaseModel[13] module.

C.2 Models

Domain Specific Retrievers We compare our method with retrieval using State-of-The-Art domain-specific embedding models, to demonstrate robust improvement over approaches that seek to devise optimal representations of paper abstracts through extensive training. We mainly use SPECTER2 models, where its variant with adapters and its multitask control codes have been reported to achieve SoTA performance on the MDCR benchmark (Medić and Snajder, 2022). We used proximity control code adapter, and base models on huggingface[14]. Meanwhile, we also use SciNCL and SciMult models as baselines, where SciNCL[15] achieves the strongest performance on the SciDocs benchmark to date, while SciMult-MHAExpert model implemented with Mixture-of-Experts (Fedus et al., 2022) architecture is reported to outperform previous models on the recommendation benchmark (Kanakia et al., 2019). For SciMult-MHAExpert, we used the model released in the official SciMult GitHub repository [16], specifically an expert model trained for link prediction.

Domain Agnostic Retrievers Since general full-documents exceed the context window of most embedding models, we specifically chose long-context window embedding models to construct a fair evaluation suite. In particular, we mainly use three different neural retrievers, Jina-Embeddings-v2-Base-EN, and BGE-M3. both highly cited general-purpose embedding models with a context window of 8192 tokens, and Inf-Retriever-v1-1.5B (context window of 32768), a lightweight version of Inf-Retriever-v1, demonstrating SoTA performance among open source embedding models for long-context retrieval (LongEmbed) in MTEB (Muennighoff et al., 2022)[17] and AIR-Bench (Chen et al.,

[12]https://platform.openai.com/docs/guides/

[13]https://docs.pydantic.dev/latest/api

[14]https://huggingface.co/allenai/specter2

[15]https://huggingface.co/malteos/scincl

[16]https://github.com/yuzhimanhua/SciMult

[17]https://huggingface.co/spaces/mteb/leaderboard

2024b) (general QA tasks) as of September 12, 2025.

In addition, we provide additional experiments using other recent general-purpose embedding models varying in context window size, such as QWEN3-Embedding-0.6B (Zhang et al., 2025b), Dewey-EN-Beta (Zhang et al., 2025a), and Granite-Embeddings-English-R2 (Awasthy et al., 2025). QWEN3-Embedding-0.6B is a light weight version of QWEN3-Embedding model series, recently released model that displays SoTA performance on MTEB-multilingual evaluation split, while Granite-Embeddings-English-R2 is the most recently released model as of August 2025. In addition, Dewey-EN-Beta is another recent model that has a context window size of 131072, which is the longest among up-to-date existing models, attaining SoTA performance on the MTEB LongEmbed benchmark.

C.3 Metrics

In our experiments, we adopt four primary metrics. **Recall@K**, **Normalized Discount Cumulative Gain (nDCG)**, **Mean Average Precision (mAP)**, and **Mean Reciprocal Rank (MRR)**. Unlike the standard nDCG metric, which assigns higher weights to gold candidates with higher significance, we treat all gold candidates for a given query equally, assigning each a uniform weight of 1 for SciFullBench experiments.

We use large cutoff value K following the convention from previous works such as SPECTER (Cohan et al., 2020), which sets K based on the maximum number of ground-truth candidates, as well as SciMult (Zhang et al., 2023a), which reports metrics with K up to 100. In our benchmark, the number of ground-truth relevant papers per query is often very large, frequently reaching several hundreds and exceeding 700 in some cases (Table 7), which reduces the bias of evaluation and robustly measures the capacity of retrieval systems. As a result, choosing a small K (e.g., K = 20) would inherently fail to cover a substantial portion of the ground-truth set, leading to artificially low recall and NDCG scores that reflect truncation effects rather than true retrieval quality. Thus, we report results with relatively high K values (e.g., 300), which covers the 99.5th percentile of ground-truth candidates per query.

In addition, we provide **Convex Hull Volume** (Graham, 1972) as a metric to quantitatively measure the dispersion of the embedding representations of retrieved candidates, computing the volume of the smallest possible bounding convex closure of the data points after projecting the embedding vectors to the three-dimensional vector space.

Notably, we employ convex hull volume in two different ways. As for the analysis provided in Table 4,we measured the average relative convex hull volume ratio of retrieved ground-truth candidates between binary samples from two distinct retrieval variants, representing the relative coverage expansion on a sample level. Meanwhile, in Figure 9, we measured the average absolute convex hull volume of total Top@K retrieved samples per retrieval variant, to analyze the trend of semantic coverage with respect to retrieval depth progression.

C.4 Human Evaluation Setup

For sampling query papers for human study, we use NeurIPS, ACL, EMNLP split in SciFullBench where its query paper exists for both reference and citations split, and randomly sampled 15 different query papers. For A2A, F2F, and CoR variant, we used result variants where jina-embeddings-v2-base-en is utilized as backbone retriever, and DPO-trained QWEN-2.5-3B-Instruct model for query optimizers in CoR, with retrieval depth of 3, and utilized retrieved results from Citations (inward citation) split. When selecting the top-5 results for each variant, we excluded the 0th index result, since the same paper is mostly retrieved due to the structure of our target corpus. Moreover, we exclude ground truth candidates for reference and citations split, in order to measure the efficacy of CoR on candidates that are regarded as incorrect by our benchmark, and formulated top-k results for qualitative evaluation. For each paper we made two duplicate tasks for evaluation and originally hired 15 human evaluators studying the field of machine learning, who willingly consented to perform evaluation on two different query papers, for approximately 20 U.S. Dollars for respective evaluator. During the evaluation process, each participant is provided three retrieval variants with its order randomly shuffled, each variant consisting of 15 pairs of query, retrieved paper with title and abstracts. The participant is then instructed to choose one of the following (**Method**, **Experiment**, **Motivation/Research Question**, and **Irrelevant**) for every retrieved result, checking the relevance with the query paper. The details on human evaluation guideline can be found in Appendix I.

D Patent-to-Patent Retrieval

In this section, we provide detailed procedures for constructing the benchmark and experimental settings for the patent-to-patent retrieval task. The statistics for PATENTFULLBENCH can be observed at Table 9 and Table 10.

D.1 PatentFullBench Construction Workflow

Step 1: Collecting Raw Data: We initially obtained a full XML dump from the USPTO Open Data portal website[18], and initially crawled US patents from 2020 to 2025 June 17th.

Step 2: Formatting Raw Data: We parsed XML dump from the USPTO Open Data portal website, obtaining metadata such as publication number, publication date, title, abstract, inventors, priority claims, claims, descriptions, references (citations from given patent). As for citations and references, we initially filtered U.S patents exclusively, especially ones where its date is over 2020, and obtained metadata such as classified country, date, document number, collecting a group of raw patents where each patent at least had 10 citations from the parsed XML dump.

Step 3: Initial Gold Label Annotation and Target Corpus Construction: After forming a pool of raw extracted patent contents from XML dump data, we obtained patent invention title, abstracts,id, and date, using PatentsView API[19]. Since PatentsView API does not support full-content access, we made sure to exclude patents absent from the extracted content from **step 2**. After obtaining actual reference to full patents per input query, we attempt to further label each query document with inward citations (incoming links) by labeling with gold documents that cite the query document. We further process citation relations through traversing all the patent files, and record the query patent B that cites A patent as ground-truth candidate for inward citation of A. Subsequent to obtaining both inward citations and outward citations (references) for respective query patent documents, we removed duplicate labels that have either the same title or abstract content. For respective splits, we excluded query documents from our benchmark set that had less than 10 ground truth candidates, and randomly sampled 800 query documents for respective splits. Successively, we formulated our target corpus via accumulating the ground truth candidates that exist within our randomly sampled query document splits without duplicates, excluding documents that either had the exact same title and abstracts. Then, we removed ground-truth candidates from the initially sampled query documents that were absent from the target corpus. This process ensures that our potential ground-truth candidates for queries always exist in the target corpus. Finally, we randomly sampled 400 queries that had at least 10 ground-truth candidates after the above filtering step, and reformulated our target corpus, making sure that duplicate candidates do not exist.

Step 4: Formatting Full Context of Query and Candidate Documents and Final Filtering We concatenated title, abstract, claims, and description content to formulate a full context for query and candidate patents. In order to eliminate potential cues that reveal hindsight information on cited patents, we used python regular expression to remove such signals, while using the citation information obtained from **step 2** to eliminate in-line citation mentions, and document numbers, along with the title and abstract information of ground truth candidates that exist in the full content of respective query patents.

As for candidate documents, we removed the query title and abstract from candidate documents whenever the candidate served as a ground-truth document for that query, preventing candidate documents from revealing direct hints about the relevant query patents. After sanitizing the full documents, we segmented the full context of candidate documents for every 3000 tokens (based on NLTK), and eliminated the patent candidate from our corpus and ground truth labels that had at least a single duplicate partial chunk, and finalized our patent retrieval benchmark.

D.2 Experiment Details

Prompt Construction We constructed prompts for patent-to-patent retrieval based on the three major aspects that comprise a patent: **Claim**, **Backgrounds**, and **Method**. **Method** aspect refers to technical details and implementation structure of the invention, while **Backgrounds** provide background information on the technical problems of previous works. Moreover, **Claim** aspects comprise of the scope of its legal protection. The prompt construction process took roughly one hour, following the same procedure demonstrated in the

[18] https://data.uspto.gov/bulkdata/datasets
[19] https://patentsview.org/apis/purpose

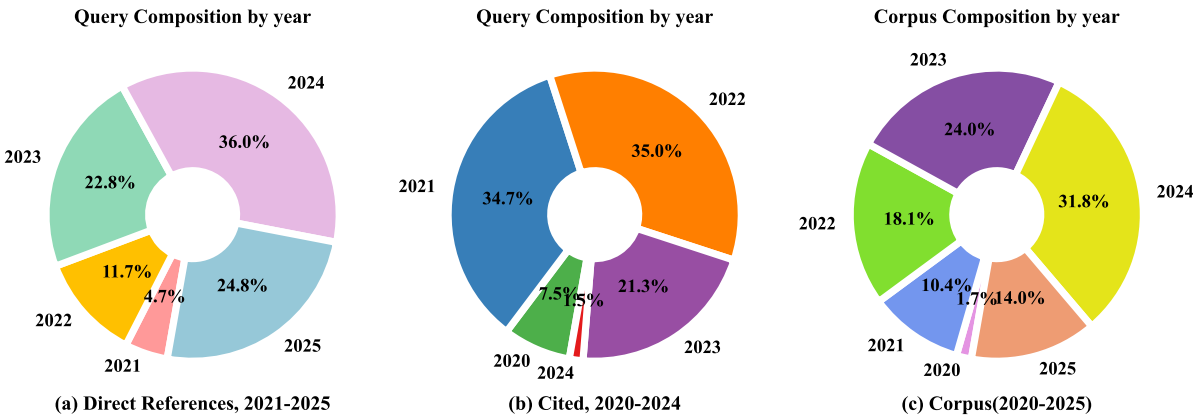

Figure 6: PatentFullBench composition by published year.

Table 9: Statistics of query documents in PatentFullBench.

	References	Citations
Nation	United States	
# of documents per split	400	400
# of tokens per abstract	119.50	120.31
# of tokens per document	190335.46	186504.29
average # of candidates per sample	37.9	53.93

Table 10: Corpus statistics in PatentFullBench.

	PATENTFULLBENCH
Total # of documents	5150
Avg. # of tokens per abstract	120.97
Avg. # of tokens per document	29141.85
Total # of segmented corpus	52584
Avg. # of tokens per segment	2865.06

Appendix C.1. For more details, please see Appendix J.

Inference Setting We evaluate our patent to patent retrieval setting with the same domain-agnostic embedding models that we utilized for our evaluation, additionally reporting abstract-to-abstract retrieval results with domain-specific retrievers, PAT-SPECTER and PaECTER (Ghosh et al., 2024). Moreover, we utilized proprietary LLM, GPT-4o-Mini-2024-0718 with temperature 0 for query optimizers and truncated the input document context to 120000 tokens using the tiktoken[20] library since patent context length is significantly longer than scientific papers, occasionally exceeding the context window of GPT-4o-Mini-2024-0718 model.

E Supplementary Experiments

Here, we provide additional experiments to further support the results and analysis in the main paper.

Consistent Generalization Across Embedding Models In our main results, we showed three variations of embedding models to demonstrate the effectiveness of our CoR framework. In this section, we further provide additional experiments on other existing long-context embedding models, namely **Granite-Embeddings-R2**, **QWEN3-0.6B-Embedding**, and **Dewey-en-Beta** in Table 11. The experimental results strongly support the robustness and general applicability of our research findings, as even without a trained model, by using proprietary **GPT-4o-2024-11-20** and **GPT-4.1-2025-04-14** with just a single round (depth) of retrieval, we were able to observe consistent performance improvements compared to naive F2F and A2A retrieval setups.

The improvements achieved using Granite-Embeddings-R2 indicate that our framework can perform robustly under a context window of 8192 as repeatedly demonstrated in our main results, and is compatible with even the most recently released model. Moreover, the results from the QWEN3-0.6B-Embedding model imply that our model can be applied to embedding models with longer context windows (32768) and also Inf-Retriever-v1-1.5B in Table 1, as well as to popular state-of-the-art long-context embedding models for general domain-agnostic multilingual retrieval tasks in the MTEB benchmark.

Furthermore, we provide evaluation results on Dewey-En-Beta, which is an embedding model with currently the longest context window among existing models, while demonstrating State-of-the-Art performance on LongEmbed benchmark. This suggests that our paper-retrieval framework is ap-

[20]https://github.com/openai/tiktoken

Table 11: Performance of CoR with different LLM backbones and neural retrievers under single-round retrieval.

Model + Retriever	ICLR-Citations Recall@300	NeurIPS-Citations Recall@100
Baseline		
Granite-Emb-Eng-R2 (A2A)	50.19	41.39
Granite-Emb-Eng-R2 (F2F)	54.00	44.37
QWEN3-0.6B-Emb (A2A)	51.47	41.76
QWEN3-0.6B-Emb (F2F)	54.03	44.81
Dewey-en-Beta (A2A)	47.76	40.54
Dewey-en-Beta (F2F)	53.15	42.06
Ours (CoR)		
GPT-4o + Granite-Emb-Eng-R2	**55.64**	**47.22**
GPT-4o + QWEN3-0.6B-Emb	**56.94**	**47.15**
GPT-4.1 + Dewey-en-Beta	**56.73**	**47.99**

plicable to models with the longest context windows up-to date, indicating the effectiveness and robustness of our suggested method to handle long context retrieval.

Robustness without Abstract-to-Abstract Retrieval Meanwhile, in Table 12, the results are reported when abstract-to-abstract retrieval (A2A) is not utilized within CoR. Despite A2A setup being omitted from its original multi-aspect queries, consistent improvement over baselines can be observed, although not to the extent when abstract-abstract retrieval is integrated into our pipeline, indicating the positive, balancing effect that A2A retrieval presents, along with robustness of our method. This in turn highlights the effectiveness of our pipeline in generating multiple aspect-aware queries in the absence of complementary abstract to abstract retrieval, strongly supporting our research contributions.

Comparison with Naive Chaining Table 13 shows the benefits of employing multi-round retrieval with CoR compared to that of naive chain-of-retrieval methods using Full Paper-to-Full Paper and Abstract-to-Abstract retrieval setup. The results indicate that our aspect-guided chain-of-retrieval along with post-order recursive aggregation is more effective, and trivial iterative retrieval is not sufficient. Note that the baseline experiments were conducted with under the same hyper parameter setting, with a cache to avoid redundant sampling likewise our framework, retrieving 300 candidates and aggregated via Reciprocal Rank Fusion. This shows that our approach in exploration and aggregation guided by aspect-centric queries is more effective compared to iterative top-k retrieval using raw content as queries, supporting the validity of our proposed retrieval system.

Starting Index Hyper Parameter Ablation In this section, we provide experiments on the hyper parameter selection starting index γ that we used

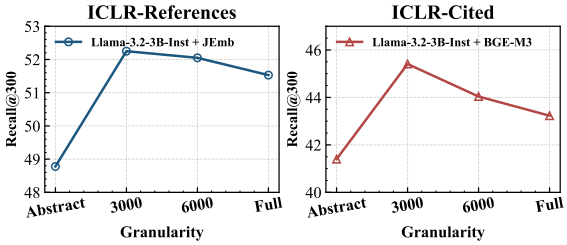

Figure 7: Results across candidate paper granularities.

Table 12: Performance analysis of ours when the original abstract to abstract retrieval is not utilized in our pipeline.

Retrieval Method	ACL-Citations Recall@200	ICLR-Citations Recall@200
Baseline		
JEmb-v2 (A2A)	35.33	39.19
JEmb-v2 (F2F)	35.88	42.44
BGE-M3 (A2A)	33.76	35.16
BGE-M3 (F2F)	32.64	36.06
Inf-Ret-v1 (A2A)	41.77	45.42
Inf-Ret-v1 (F2F)	38.86	43.00
Ours w/o A2A		
QWEN-2.5-3B-Inst + JEmb-v2	**41.13**	**44.41**
Llama-3.2-3B-Inst + BGE-M3	**35.83**	**37.49**
Llama-3.2-3B-Inst + Inf-Ret-v1	**47.05**	**50.65**

Table 13: Ablation study on the effect of our search framework compared to naive chain-of retrieval methods based on A2A and F2F retrieval setting.

Model + Retriever	ICLR-References Recall@100	ICLR-Citations Recall@200
Baseline		
JEmb-v2 (A2A)	33.04	39.10
JEmb-v2 (F2F)	33.19	41.54
BGE-M3 (A2A)	28.71	34.72
BGE-M3 (F2F)	31.00	36.12
Inf-Ret-v1 (A2A)	42.50	44.67
Inf-Ret-v1 (F2F)	36.67	42.38
Ours (w/o DPO)		
Llama-3.2-3B-Inst + JEmb-v2	**38.34**	**46.73**
Llama-3.2-3B-Inst + BGE-M3	**32.62**	**41.84**
Llama-3.2-3B-Inst + Inf-Ret-v1	**46.74**	**52.42**

for **Next Query Selection Algorithm**, to choose the starting point of choosing the representative paper with highest semantic similarity. In our case, value 1 is used to report our results in the main paper (always start from the second highest rank), to select the paper with highest similarity while avoiding the same papers as input query that may exist in our corpus (which inevitably appears at rank 0, since same papers display 100% semantic similarity). To further validate the benefits of selecting paper with highest-semantic similarity for subsequent retrieval, variations of CoR with different selection starting index γ is reported in Figure 8. The results demonstrate steady decrease in performance as selection starting index is increased, implying that CoR using the nearest-possible paper for next query selection is more desirable due to the semantic drift that occurs when relatively unrelated papers are selected as query for subsequent round of retrieval.

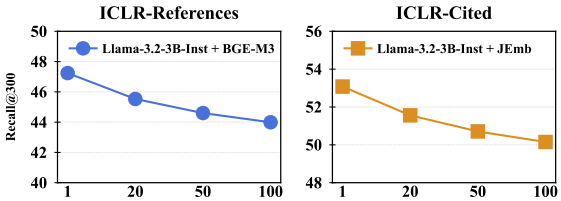

Figure 8: Hyperparameter Ablation Study on the effect of selection starting index γ on the overall performance. We use DPO-trained query optimizers on 3 rounds of retrieval.

Corpus Granularity Ablation Study Furthermore, in order to see how the granularity of candidate representations affects the performance on our multi-vector retrieval system, we conduct further analysis. The results in Figure 7 demonstrate the effectiveness of our multi-vector approach: segmenting each candidate document into 3K-token chunks yields the best performance, which outperforms not only abstract-only and full-document single-vector representations but also coarser chunking strategies (such as 6K-token segments), suggesting that finer-grained representations are effective in capturing diverse and localized signals within full-length papers, which enables our query to capture a more multi-faceted dimension of target papers, resulting in an enhanced performance.

Effect of Reciprocal Rank Fusion We conduct an additional ablation study on the effect of RRF (Reciprocal Rank Fusion) in our hybrid retrieval system (single round of CoR), compared to embedding-level merging approach in prior multi-vector retrieval approaches. We primarily compare with two traditional approaches in embedding-level merging strategy, the naive aggregation strategy that forcefully computes similarities (in our case L2 distance) between all the sub-vectors of respective query and candidates and naively sums it up to acquire similarity score of original query and documents. In our experiments, to mitigate the unfair penalization due to document length, we normalize the aggregated results with the total number of subvectors used for similarity calculation.

Meanwhile, we also present comparison with late interaction strategy (Khattab and Zaharia, 2020; Santhanam et al., 2021), where only the maximum similarity for subquery and its corresponding sub-documents is aggregated to form original query, document similarity (in our case, minimum l2 distance). Table 15 illustrates a performance drop when queries optimized in various aspects are used as subvector representations of original documents and are used to compute a single similarity value when matched with the segmented corpus of target candidates. This supports the validity of our design choice: our ranking-merging system is more robust to noise while still retrieving candidates that are strongly aligned in at least one aspect.

Table 14: Performance comparison against baselines augmented with metadata (i.e., author information).

Retrieval Method	ACL-References Recall@200	ACL-Citations Recall@200
Baseline w/ Author Information		
JEmb-v2 (A2A)	34.25	34.84
JEmb-v2 (F2F)	36.54	36.04
BGE-M3 (A2A)	32.16	33.73
BGE-M3 (F2F)	31.62	32.66
Inf-Ret-v1 (A2A)	46.66	42.36
Inf-Ret-v1 (F2F)	44.31	38.92
Ours (CoR)		
Llama-3.2-3B-Inst + JEmb-v2	**39.81**	**42.93**
Llama-3.2-3B-Inst + BGE-M3	**35.75**	**39.48**
Llama-3.2-3B-Inst + Inf-Ret-v1	**49.30**	**48.23**

Table 15: Analysis on the impact of incorporating rank fusion to attain unified results of our hybrid retrieval framework on the EMNLP-Citations split.

Retrieval Method	Recall@100	MRR@50
Ours w/o RRF & w/o A2A		
Naive Aggregation		
GPT-4o-Mini-2024-0718 + BGE-M3	21.61	32.61
Late Interaction (MaxSim)		
GPT-4o-Mini-2024-0718 + BGE-M3	26.00	31.53
Ours w/o A2A		
GPT-4o-Mini-2024-0718 + BGE-M3	**26.88**	**34.54**

Additional Diversity Analysis In this section, we provide further analysis on the improved diversity of semantics that our retrieval system presents. Figure 9 presents a linearly increasing trend where the Top@300 absolute Convex Hull Volume of 400 samples per benchmark split proportionally increases with respect to the retrieval depth, indicating that our CoR framework is able to diversify the semantics of search results as the rounds progress. Moreover, along with the quantitative metrics, we visualize the distribution of the embedding vectors of Top@300 retrieved documents for eight sampled queries on a two-dimensional plane using **t-distributed Stochastic Neighbor Embedding (t-SNE)** (Van der Maaten and Hinton, 2008) and its distribution boundaries via a two-dimensional convex hull area of the projected data points in Figure 10. We simultaneously visualize the distribution of t-SNE data points for both ground truth candidates, and baseline retrieved results (including A2A retrieved results and F2F results using the same embedding model), and retrieved candidates from CoR. The t-SNE Convex Hull visualization results also advocate our claim, where our method typically displays a more dispersed embedding distribution compared to dispersion of retrieved docu-

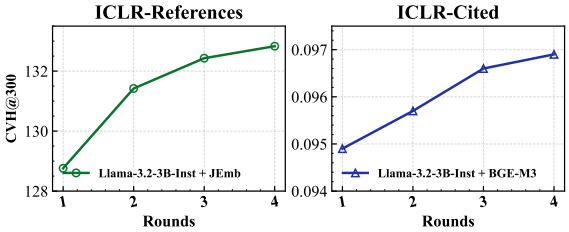

Figure 9: Diversity Analysis with retrieval depth progression. Note that we perform evaluation using DPO-trained Query Optimizers.

Table 16: Latency comparison across retrieval depths. The results are reported using jina-embeddings-v2-base-en, and untrained Llama-3.2-3B-Instruct query optimizer.

Depth	A2A	F2F	CoR
1	0.38 (s)	1.35 (s)	35.29 (s)
2	0.67 (s)	2.28 (s)	189.99 (s)
3	0.54 (s)	3.12 (s)	493.45 (s)

Table 17: Results of applying CoR to the A2A retrieval setting on SciFullBench with two retrieval rounds.

Retrieval Method	ICLR-References Recall@300	ICLR-Citations Recall@300
Baseline		
JEmb-v2 (A2A)	48.38	44.49
BGE-M3 (A2A)	41.95	39.61
Inf-Ret-v1 (A2A)	59.66	51.56
Ours (CoR)		
GPT-4o-Mini + JEmb-v2	**49.47**	**46.41**
GPT-4o-Mini + BGE-M3	**43.12**	**41.92**
GPT-4.1 + Inf-Ret-v1	**60.95**	**53.27**

ment embeddings from baselines.

Latency Analysis While the primary objective of our work is to improve retrieval effectiveness rather than efficiency, we also report the latency of CoR in Table 16, measured on a single NVIDIA RTX A6000 GPU. Although latency reduction lies outside the scope of our work, exploring techniques such as intermediate-branch or search-space pruning would be an exciting direction for future work.

Generalization to the A2A Retrieval Setting While our framework is primarily designed for the full-paper–to-full-paper retrieval setting, we additionally evaluate whether CoR is applicable to the A2A setting. To this end, we report its performance on the A2A configuration of SciFullBench in Table 17, where title–abstract pairs are used as inputs and retrieval candidates, while keeping the same query-optimization prompts. The results demonstrate that CoR is not limited to the full-paper–to-full-paper setting, but generalizes effectively to the A2A scenario, yielding meaningful performance gains.

Table 18: Performance analysis when a single comprehensive query covering the three aspects is used.

Retrieval Method	ACL-Citations Recall@200	ICLR-Citations Recall@200
Single Comprehensive Query w/o A2A		
QWEN-2.5-3B-Inst + JEmb-v2	38.66	41.68
Llama-3.2-3B-Inst + BGE-M3	32.45	34.03
Llama-3.2-3B-Inst + Inf-Ret-v1	45.36	48.55
Ours (CoR) w/o A2A		
QWEN-2.5-3B-Inst + JEmb-v2	**41.13**	**44.41**
Llama-3.2-3B-Inst + BGE-M3	**35.83**	**37.49**
Llama-3.2-3B-Inst + Inf-Ret-v1	**47.05**	**50.65**

Comparison against Metadata-Augmented Baselines We also consider a setting where additional metadata is provided to the baselines by concatenating author information to both queries and candidate documents. As shown in Table 14, CoR remains effective under this comparison and continues to outperform the baselines even when they are augmented with author metadata.

Comparison with a Single Aspect-Aware Comprehensive Query To examine the effectiveness of our multiple aspect-aware query design, we compare CoR against a baseline that uses a single comprehensive query covering all three aspects simultaneously. As shown in Table 18, CoR with multiple aspect-specific queries outperforms the single-query baseline, indicating the benefit of separating queries by aspect in paper-to-paper retrieval.

Comparison against Section-Level-Retrieval Baselines Moreover, we perform comparisons using retrieval setup using the subsection of a paper as query and candidates, particularly using Introduction section of papers as query and candidates in place of full paper or abstract content. Despite the difficulty in parsing the introduction section from papers, we have made our best effort to parse the introduction section through python regular expression while making sure to use full-paper content when introduction cannot be explicitly parsed. The results in Table 22 demonstrate that CoR is able to obtain substantial improvements over such baselines.

Case Study We provide the qualitative examples of the retrieved results from SciFullBench and PatentFullBench in Table 26 and Table 27, respectively, as well as the examples of the generated queries from them, in Appendix H.

F Robustness under Diverse Benchmarks and Evaluation Metrics

Effectiveness on the Existing Benchmark (A2A) While our main experiments are conducted on Sci-

Table 19: Performance of CoR on the SciDocs benchmark under the A2A setup; statistically meaningful gains are underlined based on paired t-tests.

Retrieval Method	Recall@10	mAP@10	nDCG@10
A2A Retrieval			
JEmb-v2 (A2A)	92.46 ± 0.00	85.53 ± 0.00	91.66 ± 0.00
BGE-M3 (A2A)	90.36 ± 0.00	80.88 ± 0.00	88.70 ± 0.00
Inf-Ret-v1 (A2A)	95.69 ± 0.00	90.02 ± 0.00	94.44 ± 0.00
Ours (CoR) w/ A2A			
GPT-4.1 + JEmb-v2	**93.20** ± 0.09	**86.20** ± 0.08	**92.18** ± 0.06
GPT-4.1 + BGE-M3	**90.52** ± 0.07	**81.15** ± 0.10	**88.88** ± 0.04
GPT-4.1 + Inf-Ret-v1	**96.16** ± 0.10	**90.40** ± 0.03	**94.80** ± 0.04

Table 20: Performance of CoR on the SciDocs benchmark under the F2F setup.

Retrieval Method	Recall@5	Recall@10	Average
A2A Retrieval			
JEmb-v2 (A2A)	63.61 ± 0.00	69.86 ± 0.00	66.74
BGE-M3 (A2A)	55.56 ± 0.00	65.97 ± 0.00	60.77
Inf-Ret-v1 (A2A)	65.28 ± 0.00	76.11 ± 0.00	70.70
F2F Retrieval			
JEmb-v2 (F2F)	65.97 ± 0.00	71.94 ± 0.00	68.96
BGE-M3 (F2F)	53.75 ± 0.00	62.92 ± 0.00	58.34
Inf-Ret-v1 (F2F)	65.42 ± 0.00	73.89 ± 0.00	69.66
Ours (CoR)			
GPT-4.1 + JEmb-v2	**66.25** ± 1.58	71.53 ± 1.10	68.89
GPT-4.1 + BGE-M3	**58.33** ± 0.64	**66.81** ± 0.96	**62.57**
GPT-4.1 + Inf-Ret-v1	**66.25** ± 1.46	**77.50** ± 0.96	**71.88**

FullBench, which provides a full-paper–to-full-paper retrieval setting with more recent publications, we also examine whether our findings hold under previously established benchmarks. To this end, we additionally evaluate CoR on the SciDocs-Cocite split (Cohan et al., 2020), which is originally defined under an A2A retrieval setting, using 993 queries whose abstract content is available[21]. As shown in Table 19, CoR remains effective under the existing benchmark with A2A setting, yielding statistically meaningful improvements over baseline methods. In addition, in Table 21, we report evaluation results on the CSFCube benchmark (Mysore et al., 2021), which consists of a total of 50 queries and with human-annotated ground truth signals based on three aspects (e.g., background, method, result) and demonstrate that CoR is able to gain robust performance improvements using diverse embedding models.

Effectiveness on the Existing Benchmark (F2F) In addition to evaluating CoR on the existing A2A benchmark setting of SciDocs, we further extend this benchmark to a F2F configuration. For this purpose, we construct a SciDocs-Full variant from the Cocite, CoRead, and CoView splits of SciDocs (Cohan et al., 2020), where both queries and candidate documents contain full-paper content, following the same preprocessing pipeline as

[21] https://huggingface.co/datasets/allenai/scirepeval

Table 21: Aggregated Performance of CoR on the CSFCube benchmark under the A2A setup using GPT-4.1 as query optimizers. We report the mean and standard deviation across three independent trials.

Retrieval Method	nDCG@20	mAP@20	Recall@20
JEmb-v2-Base-EN			
A2A	60.26 ± 0.00	30.67 ± 0.00	39.48 ± 0.00
CoR w/ A2A	**60.73** ± 0.14	**31.89** ± 0.22	**40.31** ± 0.21
BGE-M3			
A2A	57.83 ± 0.00	28.78 ± 0.00	37.45 ± 0.00
CoR w/ A2A	**59.07** ± 0.07	**30.37** ± 0.23	**38.96** ± 0.23
Inf-Ret-v1-1.5B			
A2A	63.80 ± 0.00	33.77 ± 0.00	41.86 ± 0.00
CoR w/ A2A	**65.26** ± 0.41	**35.26** ± 0.29	**43.44** ± 0.35
Granite-Emb-Eng-R2			
A2A	64.37 ± 0.00	33.52 ± 0.00	41.98 ± 0.00
CoR w/ A2A	**66.53** ± 0.36	**35.55** ± 0.15	**43.57** ± 0.27
QWEN3-0.6B-Emb			
A2A	64.13 ± 0.00	33.54 ± 0.00	42.35 ± 0.00
CoR w/ A2A	**65.28** ± 0.49	**35.46** ± 0.03	**43.59** ± 0.20
Dewey-EN-Beta			
A2A	61.61 ± 0.00	31.95 ± 0.00	40.05 ± 0.00
CoR w/ A2A	**63.77** ± 0.55	**33.46** ± 0.23	**41.76** ± 0.09

SciFullBench introduced in Appendix A. As shown in Table 20, CoR yields consistent improvements over baselines on this extended version of the benchmark, indicating that our method remains also effective across different retrieval configurations within the established benchmarks.

Effectiveness on Other Metrics In Table 24, we show the effectiveness of CoR under lower-k cutoff values on SciFullBench, where results are averaged between respective splits of two venues (ACL, EMNLP). Meanwhile, Tables 6, 20, 19, 21, and 25 report results at relatively lower k metrics as well. Overall, the results indicate that CoR achieves consistent performance gains even at lower k values, demonstrating its robustness under user-relevant evaluation settings.

Table 22: Comparison of retrieval performance under the Introduction-to-Introduction retrieval setup.

Retrieval Method	ACL-References		ACL-Citations	
	nDCG@300	**Recall@300**	**nDCG@300**	**Recall@300**
JEmb-v2-Base-EN				
Intro-to-Intro	23.06	34.97	24.32	34.41
CoR w/ Llama-3.2-3B-Instruct	**28.52**	**44.92**	**32.76**	**48.05**
CoR w/ QWEN-2.5-3B-Instruct	**28.62**	**45.67**	**33.40**	**49.00**
BGE-M3				
Intro-to-Intro	19.64	30.20	21.48	30.36
CoR w/ Llama-3.2-3B-Instruct	**25.12**	**40.19**	**30.18**	**44.46**
CoR w/ QWEN-2.5-3B-Instruct	**25.77**	**40.71**	**30.44**	**45.29**
Inf-Retriever-v1-1.5B				
Intro-to-Intro	34.03	51.94	31.72	45.89
CoR w/ Llama-3.2-3B-Instruct	**35.09**	**55.14**	**36.98**	**54.13**
CoR w/ QWEN-2.5-3B-Instruct	**35.16**	**54.68**	**37.74**	**54.78**

Table 23: Main results with additional metrics.

		ICLR				NeurIPS			
		References		**Citations**		**References**		**Citations**	
	IR Method	nDCG@300	Recall@300	nDCG@300	Recall@300	nDCG@300	Recall@300	nDCG@300	Recall@300
	Lexical-Based Retrievers								
	BM-25 (A2A)	24.78	34.69	23.35	32.59	31.79	41.73	31.68	42.22
	BM-25 (F2F)	31.09	44.54	34.57	46.28	21.59	29.49	38.38	48.65
	Domain-Specific Retriever								
	SciNCL-**A2A**	34.04	50.51	29.57	44.13	38.44	53.08	36.57	51.67
	SPECTER2-Base-**A2A**	32.71	48.99	30.97	45.23	37.77	51.82	37.99	52.95
	SPECTER2-Adapter-MTL CTRL-**A2A**	33.91	49.98	29.49	43.29	38.54	52.16	36.75	51.46
	SciMult-MHAExpert-**A2A**	28.26	42.56	24.61	36.88	33.64	47.52	32.03	45.25
Domain-Agnostic Retriever	**Jina-Embeddings-v2-BASE-EN**								
	A2A (Abstract-to-Abstract)	34.43	48.38	30.47	44.49	39.12	51.45	37.22	51.28
	F2F (Full-to-Full)	35.27	49.58	33.71	47.85	38.82	51.62	37.89	51.41
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**37.19**	**54.60**	**36.23**	**53.08**	**40.56**	**56.40**	**40.93**	**59.33**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**36.82**	**54.37**	**36.55**	**53.59**	**41.00**	**57.19**	**42.17**	**60.35**
	BGE-M3								
	A2A (Abstract-to-Abstract)	29.51	41.95	27.15	39.61	34.65	46.22	33.48	46.08
	F2F (Full-to-Full)	31.78	44.10	29.06	41.13	35.63	47.57	35.38	46.89
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	31.21	**47.24**	**31.96**	**46.53**	**36.40**	**51.53**	**37.12**	**52.01**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**32.10**	**48.16**	**32.94**	**47.97**	**37.58**	**52.69**	**38.50**	**53.57**
	Inf-Retriever-v1-1.5B								
	A2A (Abstract-to-Abstract)	43.03	59.66	35.40	51.56	48.64	63.77	41.81	57.70
	F2F (Full-to-Full)	39.01	53.85	34.06	47.89	23.17	30.69	37.09	49.69
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**44.65**	**63.19**	**40.66**	**58.41**	**49.96**	**67.43**	**46.66**	**64.66**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**44.65**	**62.93**	**41.36**	**59.24**	**49.77**	**66.62**	**47.06**	**64.88**

		ACL				EMNLP			
		References		**Citations**		**References**		**Citations**	
	IR Method	nDCG@300	Recall@300	nDCG@300	Recall@300	nDCG@300	Recall@300	nDCG@300	Recall@300
	Lexical-Based Retrievers								
	BM-25 (A2A)	21.47	31.80	21.90	30.74	20.57	29.78	19.82	28.00
	BM-25 (F2F)	22.77	34.64	27.47	37.95	24.00	35.55	28.69	39.21
	Domain-Specific Retriever								
	SciNCL-**A2A**	26.72	41.63	26.29	39.79	25.52	39.78	24.86	38.02
	SPECTER2-Base-**A2A**	25.82	40.48	27.61	41.19	24.32	37.55	26.09	39.22
	SPECTER2-Adapter-MTL CTRL-**A2A**	25.76	40.13	26.31	39.61	24.87	37.59	24.65	37.47
	SciMult-MHAExpert-**A2A**	24.19	37.49	23.76	35.31	22.02	34.75	21.38	32.26
Domain-Agnostic Retriever	**Jina-Embeddings-v2-BASE-EN**								
	A2A (Abstract-to-Abstract)	25.50	38.80	27.66	40.52	24.41	37.00	25.23	37.32
	F2F (Full-to-Full)	27.27	41.66	28.80	40.95	27.83	41.73	27.87	40.34
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**28.52**	**44.92**	**32.76**	**48.05**	**28.15**	**43.96**	**30.84**	**46.00**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**28.62**	**45.67**	**33.40**	**49.00**	**27.80**	**43.99**	**31.68**	**47.60**
	BGE-M3								
	A2A (Abstract-to-Abstract)	23.65	35.81	25.95	38.40	22.25	33.29	22.87	34.15
	F2F (Full-to-Full)	24.39	36.67	25.78	37.01	23.88	35.08	23.92	34.73
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**25.12**	**40.19**	**30.18**	**44.46**	**24.12**	**38.24**	**27.55**	**40.87**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**25.77**	**40.71**	**30.44**	**45.29**	**24.66**	**38.97**	**28.22**	**42.14**
	Inf-Retriever-v1-1.5B								
	A2A (Abstract-to-Abstract)	34.17	51.51	32.20	47.33	32.68	48.68	29.68	43.72
	F2F (Full-to-Full)	33.35	49.59	30.61	43.72	32.37	46.40	30.02	43.57
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**35.09**	**55.14**	**36.98**	**54.13**	**33.86**	**52.11**	**35.37**	**51.77**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**35.16**	**54.68**	**37.74**	**54.78**	**33.65**	**51.91**	**35.99**	**52.40**

Table 24: Retrieval Performance Analysis on additional metrics with lower cutoff k values on SCIFULLBENCH.

Retrieval Method	nDCG@20	Recall@20	nDCG@30	Recall@30
JEmb-v2-Base-EN				
A2A	16.82	13.05	17.38	15.85
F2F	18.52	14.20	19.27	17.41
CoR w/ Llama-3.2-3B-Instruct	**18.86**	**14.84**	**19.91**	**18.64**
CoR w/ QWEN-2.5-3B-Instruct	**18.77**	**14.82**	**19.87**	**18.66**
BGE-M3				
A2A	15.36	11.80	16.00	14.59
F2F	16.57	12.61	17.13	15.36
CoR w/ Llama-3.2-3B-Instruct	**16.87**	**13.05**	**17.73**	**16.39**
CoR w/ QWEN-2.5-3B-Instruct	**17.13**	**13.21**	**17.98**	**16.60**
Inf-Retriever-v1-1.5B				
A2A	20.85	16.43	21.86	20.31
F2F	21.16	16.75	22.29	20.80
CoR w/ Llama-3.2-3B-Instruct	**22.47**	**17.88**	**23.75**	**22.41**
CoR w/ QWEN-2.5-3B-Instruct	**22.83**	**18.14**	**24.15**	**22.82**

Table 25: Main results with additional metrics.

		ICLR				NeurIPS			
		References		Citations		References		Citations	
	IR Method	Recall@100	Recall@200	nDCG@200	Recall@100	Recall@100	Recall@200	nDCG@200	Recall@100
	Lexical-Based Retrievers								
	BM-25 (A2A)	24.51	30.77	21.83	22.57	30.86	37.60	30.08	31.17
	BM-25 (F2F)	31.66	39.59	32.72	33.01	20.19	25.07	36.84	37.57
	Domain-Specific Retriever								
	SciNCL-**A2A**	34.92	44.55	27.33	29.49	38.10	47.37	34.40	36.90
	SPECTER2-Base-**A2A**	33.47	43.43	28.66	30.32	36.65	46.15	36.05	37.74
	SPECTER2-Adapter-MTL CTRL-**A2A**	34.51	43.72	27.43	29.13	37.30	46.60	34.70	37.24
	SciMult-MHAExpert-**A2A**	28.44	37.13	22.57	23.51	33.01	41.98	30.14	31.66
Domain-Agnostic Retriever	**Jina-Embeddings-v2-BASE-EN**								
	A2A (Abstract-to-Abstract)	34.71	43.35	28.38	30.09	37.91	46.30	35.16	36.85
	F2F (Full-to-Full)	35.17	44.29	31.60	33.25	37.25	46.15	35.88	37.43
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**38.36**	**48.71**	**33.94**	**36.50**	**41.16**	**50.57**	**38.57**	**42.55**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**38.52**	**48.33**	**34.16**	**36.52**	**41.36**	**51.07**	**39.91**	**43.81**
	BGE-M3								
	A2A (Abstract-to-Abstract)	29.69	37.23	25.44	26.66	33.28	41.65	31.69	33.40
	F2F (Full-to-Full)	32.46	39.41	27.12	28.50	34.81	42.98	33.57	34.31
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**32.88**	**42.47**	**29.99**	**32.13**	**36.47**	**46.17**	**35.29**	**37.58**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**33.76**	**43.28**	**30.95**	**33.07**	**37.34**	**47.40**	**36.50**	**39.19**
	Inf-Retriever-v1-1.5B								
	A2A (Abstract-to-Abstract)	44.49	53.91	33.03	35.03	48.20	57.89	39.88	42.97
	F2F (Full-to-Full)	39.21	48.37	32.15	34.65	22.95	28.06	35.35	37.46
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**47.15**	**57.14**	**38.26**	**41.44**	**50.87**	**61.38**	**44.45**	**48.48**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**46.81**	**57.22**	**38.83**	**41.99**	**50.78**	**60.62**	**44.91**	**49.11**

		ACL				EMNLP			
		References		Citations		References		Citations	
	IR Method	Recall@100	Recall@200	mAP@30	Recall@200	Recall@100	Recall@200	mAP@30	Recall@200
	Lexical-Based Retrievers								
	BM-25 (A2A)	23.24	28.45	4.82	27.61	21.50	26.44	4.36	24.64
	BM-25 (F2F)	23.68	30.25	7.24	33.81	24.92	31.12	7.44	34.90
	Domain-Specific Retriever								
	SciNCL-**A2A**	29.18	36.62	5.67	34.69	27.23	34.55	4.87	33.05
	SPECTER2-Base-**A2A**	28.28	35.81	6.21	36.38	25.47	33.12	5.33	34.14
	SPECTER2-Adapter-MTL CTRL-**A2A**	28.42	35.73	5.81	34.77	26.28	33.14	4.88	32.36
	SciMult-MHAExpert-**A2A**	25.96	33.07	5.11	30.78	23.37	30.32	4.13	27.35
Domain-Agnostic Retriever	**Jina-Embeddings-v2-BASE-EN**								
	A2A (Abstract-to-Abstract)	27.75	34.49	6.20	35.33	26.08	32.64	5.24	32.43
	F2F (Full-to-Full)	28.78	36.46	7.03	35.88	29.51	36.98	6.32	35.51
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**31.96**	**39.81**	**7.67**	**42.93**	**30.76**	**39.21**	**6.57**	**40.31**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**32.02**	**40.41**	**7.92**	**43.75**	**30.83**	**38.96**	**6.66**	**41.86**
	BGE-M3								
	A2A (Abstract-to-Abstract)	25.88	32.01	5.56	33.76	23.50	29.36	4.57	29.63
	F2F (Full-to-Full)	25.15	32.21	5.98	32.64	24.81	30.84	5.24	30.08
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**28.43**	**35.75**	**6.84**	**39.48**	**26.19**	**33.31**	**5.78**	**35.63**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**28.65**	**36.33**	**6.75**	**40.22**	**26.77**	**34.28**	**6.01**	**37.01**
	Inf-Retriever-v1-1.5B								
	A2A (Abstract-to-Abstract)	37.55	46.24	7.65	41.77	34.73	43.40	6.35	39.00
	F2F (Full-to-Full)	36.35	44.37	7.81	38.86	34.65	41.83	6.98	38.92
	CoR w/ Llama-3.2-3B-Instruct (w/ DPO)	**39.44**	**49.3**	**9.13**	**48.23**	**37.26**	**46.81**	**8.05**	**46.29**
	CoR w/ QWEN-2.5-3B-Instruct (w/ DPO)	**39.73**	**49.31**	**9.60**	**48.82**	**36.80**	**46.37**	**8.38**	**46.51**

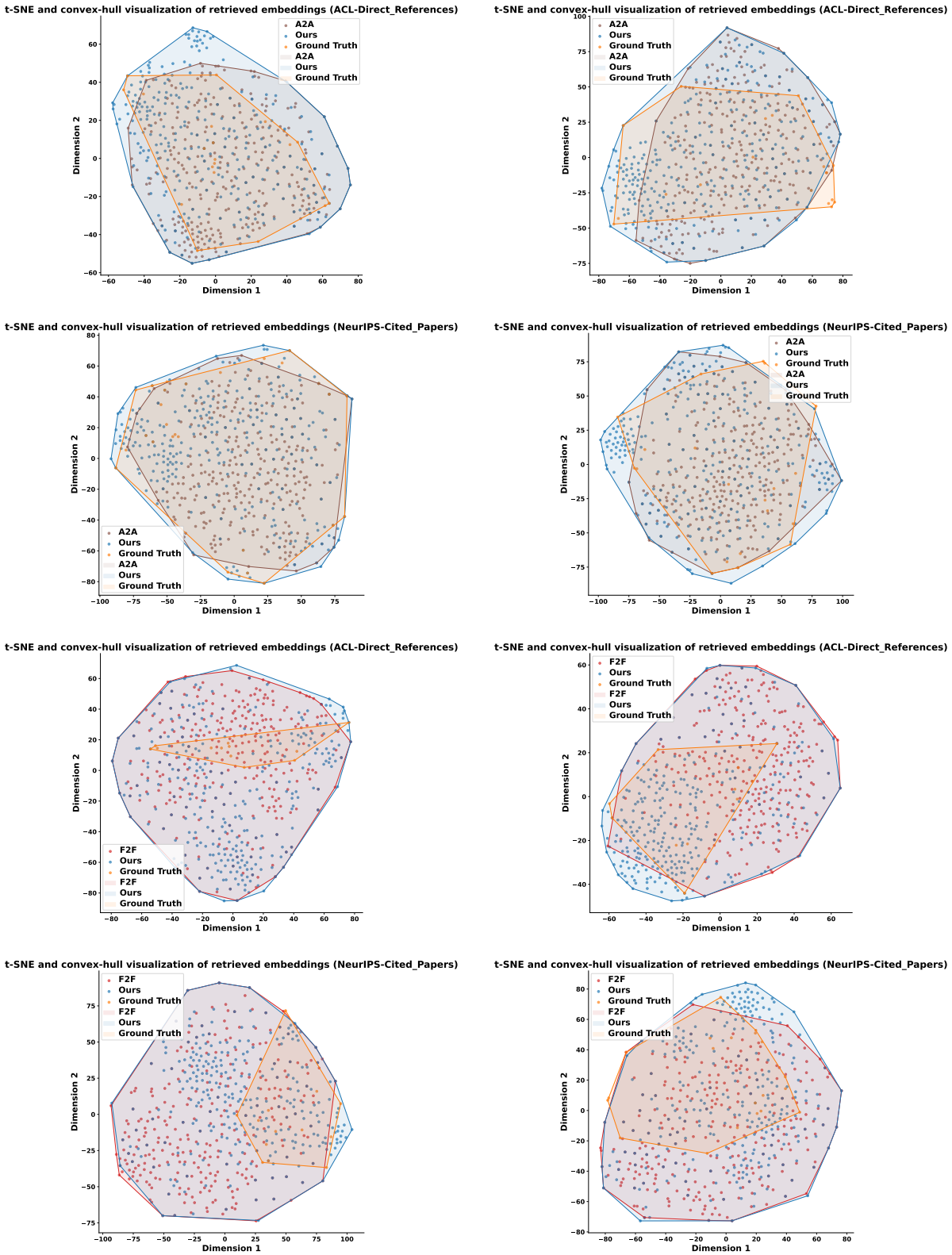

Figure 10: t-SNE and convex hull boundary visualization of Top-300 retrieved document embeddings across eight queries from respective benchmark splits. We report the results with Jina-Embeddings-v2-Base-EN as the embedding model.

G Algorithm Details

In this section, we provide further details on the functional roles of our CoR framework with formal algorithms. Specifically, Algorithm 2 defines an aspect-aware retrieval process from Section 3.2, where the query optimizers decompose the input document into aspect-aware queries for exploration and then the papers are retrieved from multi-vector corpora. Algorithm 3 describes the aspect-aware next query selection process in Section 3.3. Algorithm 4 shows the overall exploration process of CoR, using Algorithm 2 and Algorithm 3. Finally, Algorithm 5 explains the Post-Order Aggregation algorithm of CoR explained in Section 3.3.

Algorithm 2 Aspect-Aware Multi-Vector Retrieval

Require: Input paper D; Input paper abstract D_{abstract}; Top-k per query K;
Name of the Parent Retrieved Results PARENT; Corpora $\mathcal{C} \leftarrow \{\mathcal{C}_{\text{abstract}}, \mathcal{C}_{\text{chunked}}\}$;

Initialize: Functional aspect-aware LLM query optimizer agents $\mathcal{F} = \{f_{\text{M}}, f_{\text{E}}, f_{\text{R}}\}$

function ONEHOPRETRIEVAL(D, PARENT, $\mathcal{C}$, K)
  $R \leftarrow [\,]$ ▷ Memory to save aspect-aware retrieved results per document
  for all $f_i \in \mathcal{F}$ **do**
    NAME $\leftarrow$ PARENT$\|f_i$ ▷ Update name for aspect-aware retrieval for document D
    $q \leftarrow f_i(D)$
    $T \leftarrow \texttt{Retrieve}(q, C_{\text{chunked}}, K)$ ▷ Top-K Retrieved Results from chunked corpus
    $\mathcal{R}_q \leftarrow [\, h(x) \mid x \in T \,]$ ▷ Retrieved Results mapped back to paper
    R.APPEND($(\mathcal{R}_q,$ PARENT, NAME, $f_i)$)
  NAME $\leftarrow$ PARENT$\|$abstract
  $q \leftarrow D_{\text{abstract}}$
  $T \leftarrow \texttt{Retrieve}(q, \mathcal{C}_{\text{abstract}}, K)$ ▷ Top-K Retrieved Results from abstract corpus
  $\mathcal{R}_q \leftarrow [\, h(x) \mid x \in T \,]$ ▷ Retrieved Results mapped to paper
  R.APPEND($(\mathcal{R}_q,$ PARENT, NAME, abstract$)$)
  return R

Algorithm 3 Aspect-Aware Next Query Selection

Require: Root paper D; Max depth R; selection starting index γ; Top-k per query K
Aspect Aware Cache CACHE; memory to save query paper for next round retrieval Q_{next}

function INITIALASPECT(NAME) ▷ Obtain Initial Branching Aspect of the root document
  Parse NAME as (ROOT $\|\, a_1 \,\|\, a_2 \,\|\, \cdots$)
  return a_1

function SELECTNEXTQUERY(P, CACHE, γ)
  for all $r \in P$ **do**
    TOPK, PARENT, NAME, PREVIOUS ASPECT $\leftarrow$ r
    if PREVIOUS ASPECT $=$ abstract **then**
      CONTINUE
      ▷ Skip selection if previous retrieved results were from abstract query

    INITIAL ASPECT $\leftarrow$ INITIALASPECT(NAME)
    for $j = \gamma$ **to** K **do**
      DOC $\leftarrow$ TOPK$[j]$
      if DOC $\notin$ CACHE[INITIAL ASPECT] **then**
        CACHE[INITIAL ASPECT].append(DOC)
        Q_{next}.APPEND((DOC, NAME))
        break

      ▷ Select most similar document not in aspect-aware cache starting from the γ index.

Algorithm 4 Chain-of-Retrieval

Require: Root paper D; Max depth R; selection starting index γ; Top-k per query K; Corpora $\mathcal{C} \leftarrow \{\mathcal{C}_{\text{abstract}}, \mathcal{C}_{\text{chunked}}\}$

Initialize:
query queue $Q \leftarrow [(D, \textsc{Root})]$
memory to save retrieved results $\mathcal{M} \leftarrow [\,]$
$\textsc{Cache}[a] \leftarrow \emptyset \quad \forall a \in \{\text{M}, \text{E}, \text{R}\}$ ▷ Initialize Aspect-Aware Cache for Method, Experiment, and Research Question Aspect

for $r = 0$ **to** $R - 1$ **do**
  $\mathcal{M}[r] \leftarrow [\,]$, $Q_{\text{next}} \leftarrow [\,]$
  for all $(d, \textsc{parent})$ in Q **do**
    $P \leftarrow \textsc{OneHopRetrieval}(d, \textsc{parent}, \mathcal{C}, \text{K})$
    $\mathcal{M}[r].\textsc{Extend}(P)$
    $NQ \leftarrow \textsc{SelectNextQuery}(NQ, P, \textsc{Cache}, \gamma)$
  $Q \leftarrow NQ$

Algorithm 5 Recursive Post-Order Merging

Require: Root paper D; Max depth R; All Retrieved Results after depth R of retrieval $\mathcal{M}$;

Ensure: Final top-k candidates for D

function $\textsc{Group}(L)$ ▷ Group list of per round retrieved results with same parent nodes
  return $\{\, p \mapsto [\, x \in L : \text{Parent}(x) = p \,] \,\}$

function $\textsc{MergeSiblings}(L)$ ▷ Merge retrieved results with same parent nodes
  $S \leftarrow \{\}$
  for $(\textsc{Parent}, G) \in \textsc{Group}(L)$ **do**
    $S[\textsc{Parent}] \leftarrow \text{RRF}(G)$
  return S

function $\textsc{MergeEdges}(P, S)$ ▷ Merge retrieved results from parent nodes and merged child nodes
  $O \leftarrow [\,]$
  for $(\text{TopK}, \textsc{Parent}, \textsc{Name}) \in P$ **do**
    if $\textsc{Parent} \in Keys(S)$ **then**
      O.append(RRF(TopK, $S[\textsc{Parent}]$))
  return O

$r \leftarrow R - 1$

while $r > 0$ **do** ▷ Post-Order Merge, starting from the bottom level nodes
  $S \leftarrow \textsc{MergeSiblings}(\mathcal{M}[r])$
  $\mathcal{M}[r-1] \leftarrow \textsc{MergeEdges}(\mathcal{M}[r-1], S)$
  $r \leftarrow r - 1$
return $\text{RRF}(\mathcal{M}[0])$

H Case Study

Table 26: Example of the retrieved documents for the input document with CoR from a EMNLP-Citations split in SciFullBench. Due to the length of input documents, we provide only the title and abstract for the input, as well as the titles for the retrieved documents. Retrieved documents that belong to ground truth are highlighted in blue.

Input Document Meta Data	**[Title]** Stance Detection on Social Media with Background Knowledge **[Abstract]** Identifying users2019 stances regarding specific targets/topics is a significant route to learning public opinion from social media platforms. Most existing studies of stance detection strive to learn stance information about specific targets from the context, in order to determine the user2019s stance on the target. However, in real-world scenarios, we usually have a certain understanding of a target when we express our stance on it. In this paper, we investigate stance detection from a novel perspective, where the background knowledge of the targets is taken into account for better stance detection. To be specific, we categorize background knowledge into two categories: episodic knowledge and discourse knowledge, and propose a novel Knowledge-Augmented Stance Detection (KASD) framework. For episodic knowledge, we devise a heuristic retrieval algorithm based on the topic to retrieve the Wikipedia documents relevant to the sample. Further, we construct a prompt for ChatGPT to filter the Wikipedia documents to derive episodic knowledge. For discourse knowledge, we construct a prompt for ChatGPT to paraphrase the hashtags, references, etc., in the sample, thereby injecting discourse knowledge into the sample. Experimental results on four benchmark datasets demonstrate that our KASD achieves state-of-the-art performance in in-target and zero-shot stance detection.
Ground-Truth Document Meta Data	**[Title]** A More Advanced Group Polarization Measurement Approach Based on LLM-Based Agents and Graphs **[Title]** A Survey of Stance Detection on Social Media: New Directions and Perspectives **[Title]** Chain of Stance: Stance Detection with Large Language Models **[Title]** A Challenge Dataset and Effective Models for Conversational Stance Detection **[Title]** Mitigating Biases of Large Language Models in Stance Detection with Counterfactual Augmented Calibration **[Title]** Multi-modal Stance Detection: New Datasets and Model **[Title]** A Logically Consistent Chain-of-Thought Approach for Stance Detection **[Title]** Stance Detection with Collaborative Role-Infused LLM-Based Agents **[Title]** Prompting and Fine-Tuning Open-Sourced Large Language Models for Stance Classification **[Title]** Ladder-of-Thought: Using Knowledge as Steps to Elevate Stance Detection **[Title]** Investigating Chain-of-thought with ChatGPT for Stance Detection on Social Media
Top@30 Retrieved Document Meta Data	**[Title]** A Survey of Stance Detection on Social Media: New Directions and Perspectives **[Title]** Stance Detection with Collaborative Role-Infused LLM-Based Agents **[Title]** A Survey on Stance Detection for Mis- and Disinformation Identification **[Title]** Prompting and Fine-Tuning Open-Sourced Large Language Models for Stance Classification **[Title]** Chain of Stance: Stance Detection with Large Language Models **[Title]** A Challenge Dataset and Effective Models for Conversational Stance Detection **[Title]** Enabling Contextual Soft Moderation on Social Media through Contrastive Textual Deviation **[Title]** Stance Detection on Social Media with Fine-Tuned Large Language Models **[Title]** Stance Detection in Web and Social Media: A Comparative Study **[Title]** Multi-modal Stance Detection: New Datasets and Model **[Title]** TATA: Stance Detection via Topic-Agnostic and Topic-Aware Embeddings **[Title]** A Benchmark for Cross-Domain Argumentative Stance Classification on Social Media **[Title]** Mitigating Biases of Large Language Models in Stance Detection with Counterfactual Augmented Calibration **[Title]** Advancing Annotation of Stance in Social Media Posts: A Comparative Analysis of Large Language Models and Crowd Sourcing **[Title]** DEEM: Dynamic Experienced Expert Modeling for Stance Detection **[Title]** FarExStance: Explainable Stance Detection for Farsi **[Title]** Investigating Chain-of-thought with ChatGPT for Stance Detection on Social Media **[Title]** Relative Counterfactual Contrastive Learning for Mitigating Pretrained Stance Bias in Stance Detection **[Title]** A Logically Consistent Chain-of-Thought Approach for Stance Detection **[Title]** Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey **[Title]** Reinforcement Tuning for Detecting Stances and Debunking Rumors Jointly with Large Language Models **[Title]** Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification **[Title]** KCD: Knowledge Walks and Textual Cues Enhanced Political Perspective Detection in News Media **[Title]** Deciphering Political Entity Sentiment in News with Large Language Models: Zero-Shot and Few-Shot Strategies **[Title]** Argumentative Stance Prediction: An Exploratory Study on Multimodality and Few-Shot Learning **[Title]** A More Advanced Group Polarization Measurement Approach Based on LLM-Based Agents and Graphs **[Title]** Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research **[Title]** WIBA: What Is Being Argued? A Comprehensive Approach to Argument Mining **[Title]** KGAP: Knowledge Graph Augmented Political Perspective Detection in News Media **[Title]** "We Demand Justice!": Towards Social Context Grounding of Political Texts

Table 27: Example of the retrieved documents for the input document with CoR from a Citations split in PATENT-FULLBENCH. Due to the length of input documents, we provide only the title and abstract for the input, as well as the titles for the retrieved documents. Retrieved documents that belong to ground truth are highlighted in blue.

Input Document Meta Data	**[Abstract]** The present disclosure describes methods and systems directed towards providing scaled engagement and views of an e-sports event. Instead of providing the same distribution of live e-sport event data to all remote viewers of a live e-sports event, features associated with e-sports gaming network could be used to customize the distribution of live e-sport event data to promote immersive viewer experience. The enhanced immersion can also be carried out in a virtual reality or augmented reality setting. The features would be capable of providing additional information, different views, and a variety of different commentators for the e-sports event so that the viewer can be more engaged when viewing the particular e-sports event. With the increased engagement from remote viewers, the distribution of live e-sports event data can also be further modified for monetization by incorporating advertisements as well.
Ground-Truth Document Meta Data	**[Title]** Statistical driven tournaments **[Title]** Scaled VR engagement and views in an e-sports event **[Title]** Player to spectator handoff and other spectator controls **[Title]** Creation of winner tournaments with fandom influence **[Title]** User-driven spectator channel for live game play in multi-player games **[Title]** Methods and systems to increase interest in and viewership of content before, during and after a live event **[Title]** Discovery and detection of events in interactive content **[Title]** Integrating commentary content and gameplay content over a multi-user platform **[Title]** Statistically defined game channels **[Title]** Online tournament integration **[Title]** De-interleaving gameplay data
Top@30 Retrieved Document Meta Data	**[Title]** Methods and systems to increase interest in and viewership of content before, during and after a live event **[Title]** Scaled VR engagement and views in an e-sports event **[Title]** Creation of winner tournaments with fandom influence **[Title]** Player to spectator handoff and other spectator controls **[Title]** Discovery and detection of events in interactive content **[Title]** Real-time modifications in augmented reality experiences **[Title]** User-driven spectator channel for live game play in multi-player games **[Title]** Integrating commentary content and gameplay content over a multi-user platform **[Title]** Online tournament integration **[Title]** Integrating augmented reality experiences with other components **[Title]** AR-based connected portal shopping **[Title]** External screen streaming for an eyewear device **[Title]** Systems and methods for generating and facilitating access to a personalized augmented rendering of a user **[Title]** Systems, methods and apparatuses of digital assistants in an augmented reality environment and local determination of virtual object placement and apparatuses of single or multi-directional lens as portals between a physical world and a digital world component of the augmented reality environment **[Title]** Statistically defined game channels **[Title]** AR/VR enabled contact lens **[Title]** Shared augmented reality unboxing experience **[Title]** De-interleaving gameplay data **[Title]** Statistical driven tournaments **[Title]** Automated augmented reality experience creation based on sample source and target images **[Title]** Augmented reality unboxing experience **[Title]** Real-time video dimensional transformations of video for presentation in mixed reality-based virtual spaces **[Title]** Systems and methods for pinning content items to locations in an augmented reality display based on user preferences **[Title]** Avatar customization system **[Title]** Controlling interactive fashion based on body gestures **[Title]** Systems, methods, and apparatus for enhanced headsets **[Title]** User interfaces for wide angle video conference **[Title]** User interface for multi-user communication session **[Title]** Real-time upper-body garment exchange **[Title]** Applying animated 3D avatar in AR experiences

Query Case Analysis for Paper-to-Paper Retrieval

Document Meta-Data

Abstract: We introduce GAIA, a benchmark for General AI Assistants that, if solved, would represent a milestone in AI research. GAIA proposes real-world questions that require a set of fundamental abilities such as reasoning, multi-modality handling, web browsing, and generally tool-use proficiency. GAIA questions are conceptually simple for humans yet challenging for most advanced AIs: we show that human respondents obtain 92% vs. 15% for GPT-4 equipped with plugins. This notable performance disparity contrasts with the recent trend of LLMs outperforming humans on tasks requiring professional skills in e.g. law or chemistry. GAIA's philosophy departs from the current trend in AI benchmarks suggesting to target tasks that are ever more difficult for humans. We posit that the advent of Artificial General Intelligence (AGI) hinges on a system's capability to exhibit similar robustness as the average human does on

such questions. Using GAIA's methodology, we devise 466 questions and their answer. We release our questions while retaining answers to 300 of them to power a leader-board hereby accessible.

Method-Focused Query:

GAIA, a benchmark for General AI Assistants, presents a novel approach to evaluating the capabilities of large language models (LLMs) in performing real-world tasks that require a range of fundamental abilities, including reasoning, multi-modality handling, web browsing, and tool-use proficiency. The methodology employed in developing GAIA involves creating a set of 466 carefully crafted questions and their answers, covering various assistant use cases such as daily personal tasks, science, and general knowledge. Each question is designed to admit a short, single correct answer, enabling simple and robust automatic evaluation.

To address the limitations of traditional LLM benchmarks, GAIA focuses on targeting tasks that are conceptually simple for humans yet challenging for most advanced AIs. The questions are intended to be easy to create, yet vary, root in the real world, and pose a unique combination of factual and inferential demands. Furthermore, GAIA's design principles prioritize interpretability, robustness against memorization, and easiness of use, aiming to provide a more nuanced evaluation framework for assessing the capabilities of LLMs.

The development of GAIA involved several stages, starting with the creation of a seed set of questions by the authors themselves, followed by extensive annotation and validation phases involving human curators and evaluators. Two independent annotators assessed each question for ambivalence, and discrepancies were resolved through iterative refinements. This rigorous process ensured that the questions remained unambiguous and suitable for evaluation.

Upon releasing the benchmark, GAIA evaluates LLMs' performance across various capabilities, including web browsing, multi-modality processing, coding, and diverse filetype reading. The evaluation procedure leverages a zero-shot prompt strategy, utilizing a prefix prompt to guide the model toward producing a factual answer. Results indicate significant disparities between human and LLM performances, highlighting the need for more sophisticated evaluation frameworks that account for the complexities inherent in real-world scenarios.

Notably, GAIA's methodology diverges from previous benchmarks by focusing on the overall performance of AI assistants rather than individual tasks or narrow skill sets. This shift underscores the importance of considering the broader cognitive landscape encompassed by general intelligence, particularly in light of the growing emphasis on achieving true Artificial General Intelligence (AGI). Future directions for GAIA entail exploring extensions to accommodate emerging topics, incorporating human-in-the-loop feedback, and investigating novel methodologies for tackling issues such as linguistic and cultural diversity.

Overall, GAIA represents a pioneering endeavor in establishing a standardized benchmark for evaluating the capabilities of General AI Assistants, offering valuable insights into the strengths and weaknesses of contemporary LLM architectures. By fostering a deeper understanding of the intricacies underlying effective AI assistance, GAIA paves the way for continued innovation in the pursuit of AGI.

Experiment-Focused Query:

Here's a detailed summary of the experiment setup, datasets, evaluation metrics, baselines, and key findings of the paper "GAIA: A Benchmark for General AI Assistants":

Experiment Setup:

The authors propose a novel benchmark called GAIA (General AI Assistant) consisting of 466 carefully crafted questions and their answers. Each question requires a series of steps to arrive at a factual answer, mimicking real-world scenarios where AI assistants interact with humans.

The questions cover various domains, including science, literature, and everyday life, and are designed to be conceptually simple yet challenging for current AI systems.

Datasets:
The authors develop a custom dataset of 466 questions, divided into three levels of difficulty: Level 1 (simplest), Level 2 (moderate), and Level 3 (most challenging). The questions are paired with their answers, reasoning traces, and metadata describing the required capabilities and tools.

Evaluation Metrics:
The authors employ a combination of human evaluation and automated scoring to assess the performance of AI systems on the GAIA benchmark. Specifically:

* **Human Score:**
The authors annotate a subset of questions with their expected answers and measure the similarity between the predicted answer and the true answer.

* **Automated Scoring:**
The authors implement a scoring mechanism that evaluates the correctness of the predicted answer based on the presence of certain keywords and phrases in the answer.

Baselines:
The authors compare the performance of several popular large language models (LLMs) on the GAIA benchmark, including GPT-4, AutoGPT, and human evaluators. The baseline performances serve as a reference point for assessing the effectiveness of the proposed GAIA benchmark.

Key Findings:

* Despite achieving impressive results on various NLP tasks, current LLMs struggle significantly on the GAIA benchmark, particularly on Level 3 questions.
* The authors observe that equipping LLMs with external tools and APIs can significantly enhance their performance on the GAIA benchmark, highlighting the importance of integrating external resources into AI systems.
* The GAIA benchmark demonstrates the limitations of traditional evaluation metrics, such as perplexity and BLEU score, in capturing the nuances of human-AI interaction.
The authors conclude that solving the GAIA benchmark represents a significant milestone toward developing truly general-purpose AI assistants that can effectively assist humans in a wide range of contexts.

Overall, the GAIA benchmark offers a comprehensive framework for evaluating the capabilities of AI systems in simulating real-world human-AI interactions, emphasizing the importance of integrating external resources, improving reasoning and decision-making, and enhancing overall human-AI alignment.

Research-Question Focused Query:
Here's a detailed explanation of the scientific paper, focusing on its motivation, research questions, and contributions:

Motivation:
The paper introduces a novel benchmark called GAIA (General AI Assistant) aimed at assessing the capabilities of General AI Assistants. The author argues that traditional benchmarks, such as GLUE and MMLU, are insufficient for evaluating AI systems' ability to generalize across various tasks and domains. Current benchmarks tend to focus on narrow tasks, whereas real-world applications require AI systems to tackle complex, multifaceted problems. GAIA

addresses this limitation by presenting a comprehensive framework for evaluating AI assistents' reasoning, multi-modality handling, web browsing, and tool-use proficiency.

Research Questions:

The primary research question driving this paper is: "Can we develop a benchmark that comprehensively evaluates the capabilities of General AI Assistants?" Specifically, the author seeks to investigate:

* Can AI systems generalize across various tasks and domains?
* How effective are current benchmarks in capturing the complexities of real-world applications?

To address these questions, the author designs and implements GAIA, a benchmark comprising 466 carefully crafted questions covering various aspects of AI assistance, including reasoning, multi-modality handling, web browsing, and tool-use proficiency.

Contributions: The paper contributes several significant ideas to the field of Natural Language Processing (NLP):

1. **Comprehensive Benchmark for General AI Assistants**: GAIA offers a standardized framework for evaluating AI assistents' capabilities, filling the gap left by existing benchmarks.
2. **Realistic Task Set**: The benchmark features questions that mimic real-world scenarios, encouraging AI systems to apply their knowledge and skills in context-dependent settings.
3. **Focus on Reasoning and Multi-Modality Handling**: GAIA emphasizes the importance of reasoning and multi-modality handling in AI assistents, highlighting areas where current models fall short.
4. **Evaluation Methodology**: The paper presents a systematic approach to evaluating AI assistents, incorporating human evaluation, model-based evaluation, and automatic scoring techniques.

Overall, the paper demonstrates the feasibility of developing a comprehensive benchmark for General AI Assistants, paving the way for improved evaluation methodologies and more sophisticated AI systems. . . .

Query Case Analysis for Patent-to-Patent Retrieval

Document Meta-Data

Abstract: The present disclosure describes methods and systems directed towards providing scaled engagement and views of an e-sports event. Instead of providing the same distribution of live e-sport event data to all remote viewers of a live e-sports event, features associated with e-sports gaming network could be used to customize the distribution of live e-sport event data to promote immersive viewer experience. The enhanced immersion can also be carried out in a virtual reality or augmented reality setting. The features would be capable of providing additional information, different views, and a variety of different commentators for the e-sports event so that the viewer can be more engaged when viewing the particular e-sports event. With the increased engagement from remote viewers, the distribution of live e-sports event data can also be further modified for monetization by incorporating advertisements as well.

Method-Focused Query: The disclosed method for providing scaled engagement and views of a live electronic sports (e-sport) event involves a systematic approach to enhance viewer immersion through customized data distribution. Initially, the method collects comprehensive information from the live e-sport event, including in-game elements, player statistics, and audience reactions. This data is processed to generate detailed statistics about the players, which are then used to create a dynamic display of the event from an in-game camera perspective. The generated display is further modified based on individual viewer preferences, which may include adjustments for augmented reality (AR)

environments, different camera angles, and the integration of advertisements. The modifications are tailored to enhance the viewer's experience by providing relevant statistics, commentary, and interactive features, such as the ability to follow specific players or engage with other viewers. The final modified display is distributed to the viewer's device, ensuring a personalized and immersive viewing experience that can also facilitate monetization through targeted advertisements. The system architecture includes an event server for data collection, a gaming server for processing and display generation, and a public network for distribution, all of which work together to deliver a rich, engaging e-sport viewing experience.

Claim-Focused Query: The claims of this patent outline a method and system for enhancing viewer engagement and experience during live electronic sports (e-sports) events through customized data distribution. The independent claim details a method that includes collecting event data, generating player statistics, and creating a display that presents the event from an in-game camera perspective. This display is then modified based on viewer preferences, which may include augmented reality integration and personalized camera angles, before being distributed to remote viewers. The dependent claims specify refinements such as the inclusion of game-related information from player devices, audience reactions, advertisement incorporation, and user profile-based modifications that tailor the viewing experience to individual preferences. Additional claims address the provision of multiple viewing perspectives, interactive functions for viewers, and audio from other viewers, enhancing the immersive experience. The system claims describe an architecture comprising an event server and a gaming server that collectively manage data collection, processing, and distribution, while also allowing for the integration of advertisements and user interactions. Overall, the claims provide broad protection for methods and systems that facilitate a customizable and engaging viewing experience for remote participants in e-sports events.

Background Focused Query: The present invention pertains to the field of electronic sports (e-sports), specifically focusing on enhancing viewer engagement and experience during live e-sports events through customized data distribution. Existing methods of broadcasting e-sports events typically provide a uniform viewing experience to all remote viewers, lacking personalization and immersive features that could significantly enhance engagement. Current streaming platforms do not adequately leverage the wealth of data generated during live competitions, such as player statistics, audience reactions, and in-game dynamics, which limits the depth of viewer interaction and understanding of the event. Furthermore, the integration of advanced technologies like virtual reality (VR) and augmented reality (AR) remains underutilized, preventing viewers from experiencing events in a more immersive manner akin to being physically present. There is a pressing need for systems that can dynamically modify the presentation of e-sports data based on individual viewer preferences, including customizable perspectives, additional commentary, and interactive features, while also addressing challenges related to monetization through targeted advertisements. Operating constraints include the need for real-time data processing, compatibility with various user devices, and the ability to handle high viewer throughput without compromising latency or quality. The application context spans competitive gaming leagues, online streaming services, and event organizers, all of whom seek to enhance viewer satisfaction and engagement while navigating the complexities of data management and user interaction in a rapidly evolving digital landscape.

I Human Evaluation Guideline

In this section, we provide the guidelines for human evaluation provided to annotators in Table 6.

Human Evaluation Guideline

1. Evaluation Guidelines
Each task contains Title, Abstract information of Input Query Document, and has 3 different randomly shuffled retrieval variants with 5 retrieved documents with their titles and abstracts respectively. You will annotate each Query–Document pair with **EXACTLY ONE** of the following labels:

Method
Experiment
Research Question/Motivation
Irrelevant

A total of 15 retrieved documents must be labeled per Input paper.

The evaluation unit is one retrieved document for one query.
For each retrieved document, Exactly One label can be assigned.

2. Evaluation Instruction

1. Evaluation Criteria

1.1 Research Question Definition:
Choose this label when the retrieved document addresses the same or a highly similar research problem, task, or scientific question as the query.

Criteria:
Same or very similar task/problem/research objectives Overlapping motivation or problem definition High-level goals align strongly Do not consider overly vague and broad similarities as Motivation-Wise Related.

Examples:

Correct: Both works aim to address the fundamental challenge of retrieving relevant information from long-context documents, focusing on the issues of efficiently utilizing the important content within the query and candidate documents, and mitigate the effect of irrelevant noise information.

Incorrect: Both papers focus on Large Language Models.

1.2 Method Definition:
Choose this label when the retrieved document is relevant mainly because it provides similar methodology, to solve its respective problem. Focus on the similarity between Algorithms, Model Architectures, Training Methods, Optimization Strategies, or Theoretical Frameworks.

Criteria:

1) Similar algorithm or model architecture
2) Methodological or optimization insights overlap
3) Technical design relevant to the query
4) Try to Avoid Overly Vague Methodological Similarities

Examples:

Correct Although the query document A focuses on research agents for solution generation and the retrieved document B addresses automated web agents, their core methodological frameworks are highly similar. Both frameworks implement a training-free, self-evolving memory architecture in which past experiences with the highest rewards are stored during an offline stage and later retrieved during inference. This mechanism enables the policy model to generalize to unseen tasks and effectively perform a form of curriculum learning through experience reuse.

Incorrect Both Paper A and Paper B leverage large language models as autonomous agents and employ multi-agent frameworks to address their respective problem settings.

1.3 Experiment Definition:

Choose this label when the retrieved document is primarily relevant due to its experimental framework: datasets, benchmarks, evaluation methodologies, or baseline setups.

Criteria:

1) Introduces a similar dataset or benchmark used in the query
2) Uses similar evaluation protocol
3) Provides relevant baselines or experiment structures
4) Try to Avoid Overly Vague Experimental Similarities

Examples:

Correct Both paper A and paper B attempts to solve the long-horizon forgetting problem of automated Agents with different methodologies and research objectives. Both paper A and paper B performs evaluation on OsWorld benchmark and VisualWebArena benchmark, while CoACT and UITars as their strong baselines.

Incorrect Paper A and Paper B both utilize benchmarks to evaluate agents.

1.4 Irrelevant Definition:

Choose this label when the retrieved document does not meaningfully relate to the query in terms of research question, method, or experimental setup.

Criteria:

1) Only superficial keyword overlap (e.g., both mention "LLM")
2) Different domain or unrelated task
3) No substantial conceptual or technical relevance

Examples:

Vision Language Model layer compression for GUI Agents for efficiency improvement <-> Preference Optimization for Vision Language Models

2. How to Resolve Ambiguous Cases When a retrieved document could belong to multiple categories, use the following priority:

Phase 1: Check the aspect that aligns strongly among three given aspects.

Research Question → if the research problem aligns strongly
Method → if the alignment is mainly methodological
Experiment → if the relevance is primarily datasets/benchmarks/evaluation

Phase 2 Irrelevant → if none of the above apply

3. Quality Checklist Before submitting, verify the following:

Exactly one label selected per document
Strong task alignment → **Research Question**
Methodological overlap → **Method**
Dataset/evaluation relevance → **Experiment**
Unrelated documents → **irrelevant**
There are duplicate candidate documents across variants. Please make sure that the labels are applied consistently across all query, document pairs.

J Prompts

Here, we provide prompts used for experiments.

Method-Focused Agent Prompt for Paper-to-Paper Retrieval

Instruction:
You are a specialized research assistant tasked with generating a structured, detailed explanation of a scientific paper based on its **Methodology**. Your goal is to provide a clear yet comprehensive summary that makes it easy to identify relevant papers by emphasizing the **methodology** of given paper.

IMPORTANT
- Your explanation is going to be used as a query to retrieve similar papers **METHOD** wise.
- Make sure that your explanation can retrieve highly relevant papers easily.
- There are also other agents who are tasked with generating explanation on given paper. Unlike you, they are focused on experiments, and research questions of given paper. You must try to avoid overlap with possible explanations that the other two agents might generate.

Input:
You will be given the full text of a scientific paper. Carefully analyze its content, with a particular focus on the **METHODOLOGY** section, to extract its main approaches.

Key Considerations:
Highlight specific method/approach details, avoiding vague or overly general descriptions. Use precise language to ensure clarity while maintaining depth.

Output Format:
Generate a well-structured, detailed and yet clear paragraph that effectively captures the paper's approach, and key concepts in a concise yet informative manner. Focus on high-level insights rather than excessive detail.
You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

Experiment-Focused Agent Prompt for Paper-to-Paper Retrieval

You are a specialized research assistant tasked with generating a structured, detailed explanation of a scientific paper's experimental setup.
Your goal is to clearly outline the **datasets, evaluation metrics, baselines, and key experimental findings**, making it easy to understand how the paper validates its approach.

IMPORTANT
- Your explanation is going to be used as a query to retrieve similar papers **EXPERIMENT** wise.
- Make sure that your explanation can retrieve highly relevant papers easily.
- There are also other agents who are tasked with generating explanation on given paper. Unlike you, they are focused on methods, and research questions of given paper. You must try to avoid overlap with possible explanations that the other two agents might generate.

Input:
You will be provided with the full text of a scientific paper. Carefully analyze its content, paying particular attention to the **Experiments, Results, and Evaluation** sections to extract the key experimental details.

Key Considerations:
Datasets & Benchmarks: Clearly specify the datasets and benchmarks used for evaluation.
Baselines & Comparisons: Identify what methods or models the paper compares against.
Key Results & Insights: Summarize the main experimental findings without excessive detail.

Output Format:
Generate a clear, well structured and detailed paragraph that highlights the experimental methodology, datasets, evaluation metrics, baselines, and key results. Focus on high-level insights rather than excessive detail.

You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

Research Question-Focused Agent Prompt for Paper-to-Paper Retrieval

Instruction:
You are a specialized research assistant tasked with generating a structured, detailed explanation of a scientific paper based on its **Motivation, Research Questions, and Contributions**. Your goal is to provide a clear yet comprehensive summary that makes it easy to identify relevant papers by emphasizing the **core problem, key contributions, and research objectives.**

IMPORTANT
- Your explanation is going to be used as a query to retrieve similar papers **RESEARCH QUESTION** wise.
- Make sure that your explanation can retrieve highly relevant papers easily.
- There are also other agents who are tasked with generating explanation on given paper. Unlike you, they are focused on experiments, and methods of given paper. You must try to avoid overlap with possible explanations that the other two agents might generate.

Input:
You will be given the full text of a scientific paper. Carefully analyze its content, with a particular focus on the Introduction and Conclusion, to extract its main contributions, research questions, and motivations.

Key Considerations:
Highlight specific motivations, research questions, and contributions, avoiding vague or overly general descriptions.
Use precise language to ensure clarity while maintaining depth.

Output Format: Generate a well-structured, detailed and yet clear paragraph that effectively captures the paper's motivation, problem statement, research questions, and key contributions in a concise yet informative manner. Focus on high-level insights rather than excessive detail

You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

Method-Focused Agent Prompt for Patent-to-Patent Retrieval

You are a specialized research assistant tasked with generating a structured, detailed explanation of a patent's **METHOD**.

Your goal is to summarize how the invention works in practice: its core technical principles, implementation procedures, system architecture, and functional mechanisms. This explanation will be used as a query to retrieve patents with similar methods or implementation techniques.

IMPORTANT

- Focus **ONLY** on the Detailed Description and Embodiments sections.
- Capture the technical processes, structures, system flows, or algorithms.
- Paraphrase at a high-level while keeping enough technical detail for retrieval.
- Other agents will generate explanations about the invention claims and the background/problems of the given patent. You must avoid overlap with those aspects.

Input
You will be provided with the full text of a patent. Carefully analyze its Detailed Description, Examples, and Figures to extract the methodological details.

Key Considerations
Here are some key aspects that you may focus on.
- **Core technical principle:** what mechanism enables the invention to function?
- **Implementation structure:** key components, modules, or subsystems.
- **Operational flow:** how the method proceeds step by step.
- **Variants or embodiments:** different configurations or modes of execution.
- **Integration context:** how it interacts with external systems or environments.

Output Format
Produce a single, clear, and well-structured paragraph that captures the **METHOD** of the invention. Write in concise, retrieval-friendly technical language that highlights implementation strategies and system functionality.

Claim-Focused Agent Prompt for Patent-to-Patent Retrieval

You are a specialized research assistant tasked with generating a structured, detailed explanation of a patent's **CLAIMS**.

Your goal is to clearly outline the legal protection scope: what is being claimed, how broad the claims are, and what technical features or components are covered.
This explanation will be used as a query to retrieve patents with similar **CLAIM** structures and protection scopes.

IMPORTANT
- Focus **ONLY** on the **CLAIMS** section.
- Emphasize the scope of protection, claimed components, processes, and relationships among them.
- Avoid background information, prior art, or detailed embodiments (those are handled by other agents).
- Do not directly copy claim sentences; paraphrase into concise, high-level, retrieval-friendly language.
- Include both the broad independent claims (core invention scope) and notable dependent claims (specific refinements).
- Other agents will generate explanations about the invention **details/method** and the **background/problems** of the given patent. You must avoid overlap with those aspects.

Input
You will be provided with the full text of a patent. Carefully analyze the **CLAIMS** section.

Output Format
Produce a clear, well-structured paragraph that captures the scope and focus of the **CLAIMS**. Use neutral, technical language suitable for retrieval so that patents with overlapping protection scopes

can be surfaced.

Background-Focused Agent Prompt for Patent-to-Patent Retrieval

You are a specialized research assistant tasked with generating a structured, detailed explanation of a patent's **Problem / Background** (i.e., the technical field and the shortcomings of prior art).

Your goal is to clearly state the problem space—technical domain, prior-art limitations, unresolved challenges, operating constraints, and desired (non-solution) performance objectives—so that similar patents can be retrieved by shared problem patterns rather than specific solutions.

IMPORTANT

- Your explanation is going to be used as a query to retrieve patents that are similar in the **PROBLEM SPACE.**
- Include concrete domain terms (components, materials, data types), relevant standards/regulations, operating conditions (e.g., throughput, latency, power, temperature), failure modes, and application contexts that characterize the problem.
- Other agents will generate explanations about the invention details/method and the claims made from the input patent. You must avoid overlap with those aspects.

Input

You will be provided with the full text of a patent.

Key Considerations

Here are some key aspects that you may focus on.

- **Technical Field:** the domain and subdomain (e.g., "wireless edge inference for medical imaging").
- **Prior Art & Limitations:** concrete bottlenecks, inefficiencies, failure cases, safety/privacy concerns, interoperability issues.
- **Unmet Technical Objectives:** what must be achieved (targets or constraints).
- **Operating Constraints & Edge Cases:** data distributions, environmental/thermal limits, network conditions, power/memory budgets, size/weight/cost constraints, lifecycle/maintenance issues.
- **Application Contexts & Stakeholders:** deployment scenarios, industries, users, and integration boundaries.

Output Format

Generate a single, clear, well-structured paragraph in neutral technical language that captures **ONLY** the **Problem / Background**. Paraphrase in your own words at a high level; do not copy text or include any solution, embodiment, or claim content. The paragraph should be information-dense and retrieval-friendly so that it can surface patents that confront the same technical challenges.

Base Agent Prompt for Paper-to-Paper Retrieval

Instruction: You are a specialized research assistant tasked with generating a structured,detailed explanation of a scientific paper based three different aspects.

1. Method-Specific Queries:

Generate a structured, detailed explanation of a scientific paper based on its Methodology Your goal is to provide a clear yet comprehensive summary that makes it easy to identify relevant papers by emphasizing the methodology of given paper.

IMPORTANT

- Your explanation is going to be used as a query to retrieve similar papers METHOD wise.
- Make sure that your explanation can retrieve highly relevant papers easily.

Input:

You will be given the full text of a scientific paper. Carefully analyze its content, with a particular focus on the **METHODOLOGY** section, to extract its main approaches.

Key Considerations:

Highlight specific method/approach details, avoiding vague or overly general descriptions. Use precise language to ensure clarity while maintaining depth.

Output Format:

Generate a well-structured, detailed and yet clear paragraph that effectively captures the paper's approach, and key concepts in a concise yet informative manner. Focus on high-level insights rather than excessive detail You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

2. Experiment-Specific Queries:

Generate a structured, detailed explanation of a scientific paper's experimental setup Your goal is to clearly outline the datasets, evaluation metrics, baselines, and key experimental findings, making it easy to understand how the paper validates its approach.

IMPORTANT

- Your explanation is going to be used as a query to retrieve similar papers **EXPERIMENT** wise.
- Make sure that your explanation can retrieve highly relevant papers easily.

You will be provided with the full text of a scientific paper. Carefully analyze its content, paying particular attention to the Experiments, Results, and Evaluation sections to extract the key experimental details.

Key Considerations:

Datasets & Benchmarks: Clearly specify the datasets and benchmarks used for evaluation.
Baselines & Comparisons: Identify what methods or models the paper compares against.
Key Results & Insights: Summarize the main experimental findings without excessive detail.

Output Format:

Generate a clear, well structured and detailed paragraph that highlights the experimental methodology, datasets, evaluation metrics, baselines, and key results. Focus on high-level insights rather than excessive detail. You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

3. Research Question-Specific Queries:

Generate a structured, detailed explanation of a scientific paper based on its Motivation, Research Questions, and Contributions. Your goal is to provide a clear yet comprehensive summary that makes it easy to identify relevant papers by emphasizing the core problem, key contributions, and research objectives.

IMPORTANT

- Your explanation is going to be used as a query to retrieve similar papers **RESEARCH QUESTION** wise.
- Make sure that your explanation can retrieve highly relevant papers easily.

Input:

You will be given the full text of a scientific paper. Carefully analyze its content, with a particular focus on the Introduction and Conclusion, to extract its main contributions, research questions, and motivations.

Key Considerations:
Highlight specific motivations, research questions, and contributions, avoiding vague or overly general descriptions. Use precise language to ensure clarity while maintaining depth.

Output Format:
Generate a well-structured, detailed and yet clear paragraph that effectively captures the paper's motivation, problem statement, research questions, and key contributions in a concise yet informative manner. Focus on high-level insights rather than excessive detail You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

Output:
Return a structured json file for respective Method-Specific Queries, Experiment-Specific Queries, and Research Question-Specific queries, with the respective keys as "method_query", "experiment_query", and "research_question_query". Each key should contain the generated explanation as a string.

Prompt for Single Comprehensive Query for Paper-to-Paper Retrieval

Instruction:
You are a specialized research assistant tasked with generating a structured, detailed explanation of a scientific paper based three different aspects. You should generate a single query covering method, experiment, and research question aspects. Carefully read the below instructions and generate a single comprehensive query that covers below three aspects.

1. **Method-Specific Queries:**
Generate a structured, detailed explanation of a scientific paper based on its Methodology. Your goal is to provide a clear yet comprehensive summary that makes it easy to identify relevant papers by emphasizing the methodology of given paper.

IMPORTANT
- Your explanation is going to be used as a query to retrieve similar papers **METHOD** wise.
- Make sure that your explanation can retrieve highly relevant papers easily.

Input:
You will be given the full text of a scientific paper. Carefully analyze its content, with a particular focus on the **METHODOLOGY** section, to extract its main approaches.

Key Considerations:
Highlight specific method/approach details, avoiding vague or overly general descriptions. Use precise language to ensure clarity while maintaining depth.

Output Format:
Generate a well-structured, detailed and yet clear paragraph that effectively captures the paper's approach, and key concepts in a concise yet informative manner. Focus on high-level insights rather than excessive detail You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

2. **Experiment-Specific Queries:**
Generate a structured, detailed explanation of a scientific paper's experimental setupYour goal is to clearly outline the datasets, evaluation metrics, baselines, and key experimental findings, making it easy to understand how the paper validates its approach.

IMPORTANT
- Your explanation is going to be used as a query to retrieve similar papers **EXPERIMENT** wise.
- Make sure that your explanation can retrieve highly relevant papers easily.

You will be provided with the full text of a scientific paper. Carefully analyze its content, paying particular attention to the Experiments, Results, and Evaluation sections to extract the key experimental details.

Key Considerations:
Datasets & Benchmarks: Clearly specify the datasets and benchmarks used for evaluation.
Baselines & Comparisons: Identify what methods or models the paper compares against.
Key Results & Insights: Summarize the main experimental findings without excessive detail.

Output Format:
Generate a clear, well structured and detailed paragraph that highlights the experimental methodology, datasets, evaluation metrics, baselines, and key results. Focus on high-level insights rather than excessive detail. You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

3. Research Question-Specific Queries:
Generate a structured, detailed explanation of a scientific paper based on its Motivation, Research Questions, and Contributions. Your goal is to provide a clear yet comprehensive summary that makes it easy to identify relevant papers by emphasizing the core problem, key contributions, and research objectives.

IMPORTANT
- Your explanation is going to be used as a query to retrieve similar papers **RESEARCH QUESTION wise.**
- Make sure that your explanation can retrieve highly relevant papers easily.

Input:
You will be given the full text of a scientific paper. Carefully analyze its content, with a particular focus on the Introduction and Conclusion, to extract its main contributions, research questions, and motivations.

Key Considerations:
Highlight specific motivations, research questions, and contributions, avoiding vague or overly general descriptions. Use precise language to ensure clarity while maintaining depth.

Output Format: Generate a well-structured, detailed and yet clear paragraph that effectively captures the paper's motivation, problem statement, research questions, and key contributions in a concise yet informative manner. Focus on high-level insights rather than excessive detail You must not include title and abstract of given paper in your answer, and try to put it into your own words with high level reasoning after reading the paper.

Output:
Return a comprehensive query that covers the Methodology, Research Questions, and Experimental Details within a single paper.